

Uniwersytet Warszawski
Wydział Filozofii

Wojciech Stempniak

Nr albumu: 433855

**Struktura zależnościowa koordynacji
– analiza korpusów
Universal Dependencies**

**Praca licencjacka
na kierunku KOGNITYWISTYKA**

Praca wykonana pod kierunkiem
prof. dr. hab. Adama Przepiórkowskiego
Uniwersytet Warszawski

Warszawa, czerwiec 2024

Streszczenie

Istnieje wiele poglądów na temat struktury zależnościowej koordynacji, czyli konstrukcji współrzędnie złożonej. W literaturze opisane są cztery główne podejścia – model praski, londyński, stanfordzki i moskiewski (Popel i in., 2013).

Poprzednie badania (Przepiórkowski i Woźniak, 2023) opisują metodę pozwalającą na testowanie poprawności tych modeli. Polega ona na analizie tendencji do umieszczania krótszego członu koordynacji na początku konstrukcji współrzędnie złożonej. Wykorzystuje ona zasadę Dependency Length Minimization (DLM, Temperley 2007), czyli tendencję do formułowania zdań tak, aby łączna długość relacji między słowami w zdaniu była jak najmniejsza. Przepiórkowski i Woźniak (2023) na podstawie analizy koordynacji w korpusie języka angielskiego argumentują za poprawnością modeli symetrycznych, czyli podejścia praskiego i londyńskiego.

Cztery główne modele struktury zależnościowej koordynacji zostały opracowane na podstawie analiz języków inicjalnych, czyli takich, w których głowy znajdują się zwykle na początku fraz. Jednak podejścia te mogą nie opisywać prawidłowo koordynacji w językach finalnych, tj. takich, w których głowa zwykle jest na końcu frazy. Kanayama i in. (2018) postulują wprowadzenie alternatywnych modeli struktury zależnościowej koordynacji dla języków finalnych. W niniejszej pracy przedstawiam przewidywania 12 modeli struktury zależnościowej koordynacji. Zestawiam je z wynikami analizy korpusów językowych opisanych w standardzie Universal Dependencies (UD, De Marneffe i in. 2021). W badaniu uwzględniono korpusy 13 języków, w tym 9 inicjalnych i 2 finalnych.

Wyniki badania potwierdzają występowanie w języku angielskim tendencji zaobserwowanych w pracy Przepiórkowski i Woźniak (2023). Podobne zależności występują także w języku czeskim oraz w łacinie. Niemniej jednak w pozostałych badanych językach przewidywane tendencje nie zostały zaobserwowane. Pokazuję, że może to wynikać z niewystarczającej ilości oraz złej jakości danych użytych w badaniu. Proponuję poprawę metodologii i dalsze badania dotyczące struktury zależnościowej koordynacji.

Słowa kluczowe

koordynacja, struktura koordynacji, zależności składniowe, drzewa zależnościowe, Universal Dependencies (UD), Dependency Length Minimization (DLM), języki inicjalne, języki finalne, badanie korpusowe

Tytuł pracy w języku angielskim

Dependency structure of coordination – an analysis of Universal Dependencies corpora

Podziękowania

Pragnę złożyć serdeczne podziękowania mojemu promotorowi prof. dr. hab. Adamowi Przepiórkowskiemu za inspirację i pomoc w prowadzeniu badań oraz cierpliwość, wyrozumiałość i zaangażowanie podczas kierowania moją pracą.

Ukończenie niniejszej pracy nie byłoby możliwe bez pomocy mgr. Berkego Şenşerkerci, któremu serdecznie dziękuję za pomoc w dostosowaniu stosowanych przeze mnie algorytmów do warunków języków finalnych i ewaluacji algorytmu użytego do analizy języka tureckiego.

Dziękuję również Magdalenie Borysiak, Katarzynie Zrobek i Oskarowi Pruszyńskiemu za koleżeńską pomoc w prowadzeniu badań i zgłębianiu wiedzy na temat badanych przeze mnie zjawisk.

Spis treści

1. Wprowadzenie	7
1.1. Cel pracy	7
1.2. Struktura pracy	8
2. Wstęp teoretyczny	9
2.1. Zależności składniowe	9
2.2. Universal Dependencies	10
2.3. Koordynacja	10
2.4. Dependency Length Minimization	12
2.5. Kolejność członów koordynacji	13
2.6. Języki inicjalne oraz finalne	15
2.7. Struktura zależnościowa koordynacji	16
3. Struktura zależnościowa koordynacji	18
3.1. Języki inicjalne (Przepiórkowski i Woźniak, 2023)	18
3.1.1. Metody	18
3.1.2. Podejścia	19
3.1.3. Wyniki i interpretacja	21
3.2. Języki finalne	21
3.2.1. Różnice względem języków inicjalnych	21
3.2.2. Podejścia	22
3.2.3. Metody	24
3.2.4. Predykcje	24
3.3. Języki mieszane	29
4. Przetwarzanie danych	30
4.1. Dane wejściowe	30
4.1.1. Korpusy zależnościowe	30
4.1.2. Format danych	33
4.2. Wyciąganie koordynacji	33
4.2.1. Relacja conj	33

4.2.2. Wyznaczanie głów członów, nadzędnika i spójnika koordynacji	34
4.2.3. Wyznaczanie granic członów	36
4.2.4. Określanie pozycji nadzędnika	39
4.2.5. Obliczanie długości członu	39
4.2.6. Koordynacje zagnieżdżone	41
4.2.7. Procedura znajdowania koordynacji z uwzględnieniem koordynacji zagnieżdżonych	42
4.3. Weryfikacja działania algorytmu	44
4.3.1. Ograniczenia	44
4.3.2. Dobór języków	44
4.3.3. Losowanie wyciągniętych koordynacji	45
4.3.4. Ocena poprawności	45
4.3.5. Wyniki	45
5. Metody statystyczne	47
5.1. Względna pozycja głównego członu	47
5.2. Pozycja nadzędnika	48
5.3. Długość członów koordynacji	51
5.4. Różnica długości członów a pozycja krótszego członu	52
6. Dyskusja wyników	58
6.1. Replikacja poprzednich badań	58
6.1.1. Język angielski	58
6.1.2. Języki słowiańskie	58
6.1.3. Języki romańskie	59
6.1.4. Języki mieszane	59
6.1.5. Języki finalne	59
6.2. Przewidywania modeli struktury zależnościowej koordynacji	60
6.2.1. Języki inicjalne	60
6.2.2. Języki finalne	61
6.3. Wyjaśnienia wyników	62
6.3.1. Gramatykalizacja krótszego prawego członu	62
6.3.2. Monotoniczność tendencji	62
6.3.3. Przyczyny tendencji wzrostowych	63
7. Zakończenie	66
7.1. Ograniczenia	66
7.1.1. Korpusy zależnościowe UD	66
7.1.2. Dependency Length Minimisation	67
7.1.3. Ograniczenia metodologiczne	68

7.2. Przyszłe badania	68
7.2.1. Surface Syntactic Universal Dependencies (SUD)	68
7.2.2. Analiza koordynacji różnych długości	69
A. Rozszerzone zależności składniowe	70
A.1. Enhanced Dependencies	70
A.2. Języki inicjalne	71
A.3. Języki finalne	71
B. Wyniki poprzednich badań	73
C. Długość członów koordynacji	75
C.1. Języki inicjalne	75
C.2. Języki mieszane	78
C.3. Języki finalne	78
D. Różnica długości członów a pozycja krótszego członu	79
D.1. Języki inicjalne	79
D.2. Języki mieszane	89
D.3. Języki finalne	92

Rozdział 1

Wprowadzenie

1.1. Cel pracy

Jedno z pytań, na które próbuje odpowiedzieć teoria składni, brzmi: „Dlaczego układamy słowa w zdaniach w ten sposób, a nie inaczej?”

Rozważmy dwa zdania:

- (1) Wczoraj czytałem [[gazetę] i [bardzo ciekawą książkę przygodową]].
- (2) Wczoraj czytałem [[bardzo ciekawą książkę przygodową] i [gazetę]].

W zdaniach (1) i (2) znajdują się koordynacje zawierające dwa człony: *gazetę* oraz *bardzo ciekawą książkę przygodową*. Jedyną różnicą między tymi zdaniami jest kolejność ustawienia tych członów. Człony mają podobne własności syntaktyczne i semantyczne. W związku z tym można by przypuszczać, że nie ma znaczenia, który człon zostanie umieszczony w zdaniu jako pierwszy. Niemniej jednak człony różnią się długością. Długość członu można liczyć na rozmaite sposoby – w słowach, tokenach, sylabach lub znakach. Na przykład, *bardzo ciekawą książkę przygodową* ma 4 tokeny i 33 znaki, zaś *gazetę* ma 1 token i 6 znaków.

Pierwszą rzeczą, jaką wykazuję, jest to, że długość członów koordynacji jest istotnym czynnikiem w wyborze ich kolejności. Wynika to ze zjawiska znanego jako Dependency Length Minimization (DLM). Jest to uniwersalna zależność występująca w języku naturalnym. Polega ona na tym, że słowa układane są w zdaniu w sposób minimalizujący sumę długości relacji składniowych między nimi (Temperley, 2007; Futrell i in., 2015). Ponieważ długość relacji wewnętrz czelonów koordynacji jest stała, jedynie ustawienie członów w zdaniu ma wpływ na sumę długości relacji. W tej pracy badam, w jaki sposób pozycja nadziednika wpływa na ustawienie kolejności członów. Pokazuję, że w językach inicjalnych, takich jak polski czy angielski, zdanie o strukturze takiej jak (1) ma większą szansę na pojawienie się w języku naturalnym niż zdanie takie, jak (2).

Drugim problemem, jaki podejmuję, jest kwestia opisu relacji składniowych. Wśród lingwistów nie ma zgody co do tego, jakie dokładnie relacje zależnościowe łączą poszczególne człony.

gólne elementy koordynacji. Istnieją cztery uznawane sposoby opisu – model praski, londyński, stanfordzki i moskiewski (Popel i in., 2013; Przepiórkowski i Woźniak, 2023). Niemniej jednak, zakładając prawdziwość DLM, można wykazać, że nie wszystkie podejścia są poprawne. W tej pracy sprawdzam, które modele tworzą drzewa zależnościowe zgodne z przewidywaniami DLM.

Przepiórkowski i Woźniak (2023) przeprowadzili analizę koordynacji w języku angielskim. Wykazali, że modele stanfordzki i moskiewski nie są poprawnymi metodami opisu koordynacji. Moim głównym celem jest replikacja tego badania oraz rozszerzenie go na trzynaście języków. W ramach badania analizuję koordynacje osobno w językach inicjalnych oraz finalnych.

W badaniu wykorzystuję korpusy zależnościowe opisane w wersji 2.14 standardu Universal Dependencies (De Marneffe i in., 2021). Korpusy te pochodzą ze strony <https://universaldependencies.org/>. Wszystkie skrypty użyte w analizie dostępne są w repozytorium pod adresem <https://github.com/wjstempniak/Dependency-Structure-of-Coordination>.

1.2. Struktura pracy

Niniejszy pierwszy rozdział pracy poświęcam wprowadzeniu do tematu i opisaniu struktury pracy. W rozdziale 2 przedstawiam podstawy teoretyczne mojej pracy. Przedstawiam dokładnie problem opisywania struktury koordynacji i omawiam standard Universal Dependencies (UD). W rozdziale 3 przedstawiam przewidywania różnych modeli dotyczących struktury koordynacji w językach inicjalnych i finalnych. W rozdziale 4 przedstawiam kolejne etapy przeprowadzonego przeze mnie badania. Szczególną uwagę poświęcam problemowi automatycznego wyznaczania granic członów koordynacji. W rozdziale 5 opisuję analizę statystyczną uzyskanych przeze mnie wyników. W rozdziale 6 przedstawiam wnioski wyciągnięte z badania. Pracę kończę rozdziałem 7, w którym opisuję ograniczenia mojej pracy oraz proponuję dalsze możliwości badań w tej dziedzinie.

Rozdział 2

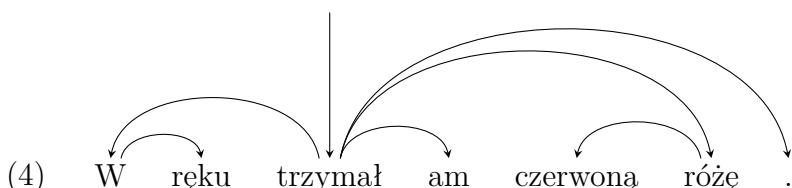
Wstęp teoretyczny

2.1. Zależności składniowe

Zależności składniowe są centralnym zagadnieniem gramatyki zależnościowej, czyli dziedziny teorii składni zajmującej się połączeniami pomiędzy poszczególnymi słowami wchodzącyymi w skład zdania. Są to relacje łączące dwa elementy: nadzrębniok oraz zależny od niego podrzędniok. Elementami łączonymi przez zależności są tokeny – przez to pojęcie należy rozumieć słowa oraz znaki interpunkcyjne, a także, według niektórych podejść, części niektórych słów.

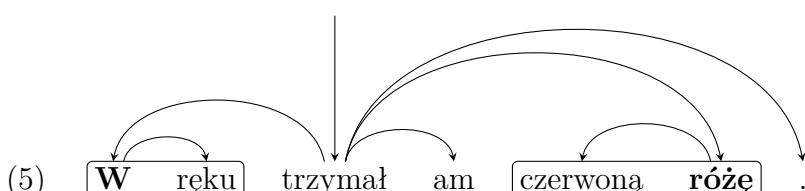
Relacje występujące w obrębie zdania można opisać za pomocą grafu zwanego drzewem zależnościowym. (4) jest przykładowym drzewem zależnościowym zdania (3):

- (3) W ręku trzymałam czerwoną różę.



Warto zauważyć, że słowo *trzymałam* składa się z dwóch tokenów – *trzymał* oraz *am*.

Należy odróżnić zależności składniowe od relacji składnikowych. Relacje składnikowe łączą poszczególne słowa we frazy oraz frazy w bardziej złożone frazy w zdaniu (Chomsky, 1956). Schemat (5) przedstawia drzewo zależnościowe zdania (3) z wyróżnionymi niektórymi frazami.



Słowo określające kategorię syntaktyczną danej frazy nazywane jest głową tej frazy (Hoeksema, 1992). Przykładowo, głową frazy przyimkowej *W ręku* jest przyimek *W*, zaś głową frazy rzeczownikowej *czerwoną różę* – rzeczownik *różę*.

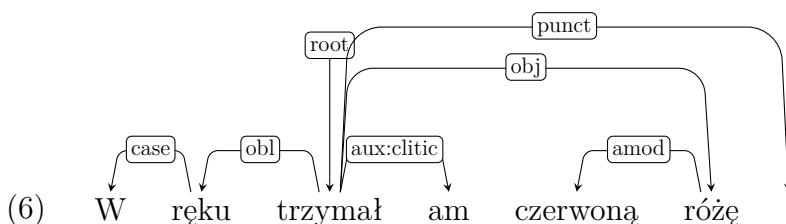
W drzewie zależnościowym głowę danej frazy można rozpoznać po tym, że nie posiada nadrzędników w jej obrębie.

2.2. Universal Dependencies

Obecnie nie ma w lingwistyce jednej, szeroko akceptowanej formuły opisu relacji zależnościowych. Standardem, z którego korzystam, jest Universal Dependencies. Jest to formalizm tworzący „ramy dla spójnego opisu gramatyki” (De Marneffe i in., 2021) języków, a więc takich cech jak części mowy, cechy morfologiczne czy właśnie relacje składniowe.

Universal Dependencies opisuje zależności składniowe za pomocą zestawu uniwersalnych znaczników, charakterystycznych dla rodzaju relacji. Na przykład, relację łączącą podmiot nominalny z orzeczeniem opisuje znacznik *nsubj*, a dopełnienie bliższe – *obj*. Znaczniki UD posiadają różne warianty. Podmiot zdania w trybie biernym opisuje znacznik *nsubj:pass*. W tej pracy traktuję każdy wariant jako osobny znacznik, chyba że zaznaczam inaczej.

Poniżej znajduje się drzewo zależnościowe zdania (3) opisane według standardu UD.¹



Należy zwrócić uwagę na krawędzie łączące tokeny *W*, *ręku* oraz *trzymał*. Według Universal Dependencies głową frazy przyimkowej *W ręku* jest rzeczownik *ręku*. Jest to sprzeczne z teorią lingwistyczną i wskazuje na niedoskonałość tego standardu. Niemniej jednak w niniejszej pracy będę zakładał poprawność UD.

2.3. Koordynacja

Koordynacja lub konstrukcja współrzędnie złożona to zestawienie w zdaniu dwóch lub więcej elementów pełniących tę samą funkcję syntaktyczną. Elementy te nazywa się członami koordynacji.

¹Drzewa zależnościowe w formacie UD wszystkich zdań niepochodzących z korpusów zostały stworzone przez automatyczne parsowanie parserem Trankit (Van Nguyen i in., 2021).

Problem wyznaczania granic członów jest jednym z ważniejszych aspektów analizy koordynacji. Nie jest to zadanie trywialne i często wymaga odwołania się do semantyki. Rozważmy następujące przykłady:

- (7) Niesforne dzieci i nauczyciele pojechali na wycieczkę.
- (8) Niesforni uczniowie i nauczyciele pojechali na wycieczkę.

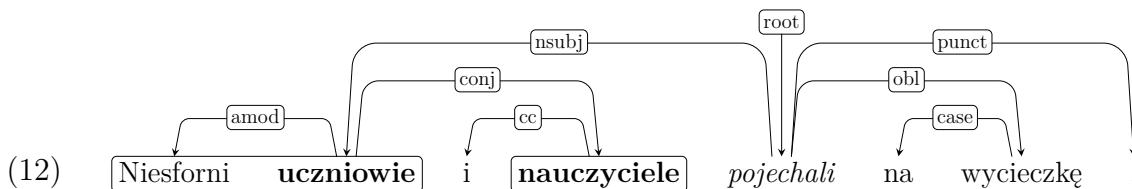
W przypadku zdania (7) słowo *Niesforne* jest niekontrowersyjnym podrzędnikiem słowa *dzieci*, ponieważ wynika to ze związku zgody. Z tego wynika, że granice członów koordynacji wyglądają następująco:

- (9) [[Niesforne dzieci] i [nauczyciele]] *pojechali* na wycieczkę.

W zdaniu (8) *Niesforni* może opisywać zarówno *uczniowie*, jak i *nauczyciele*. W takim wypadku syntaktyka dopuszcza dwie interpretacje zdania:

- (10) [[Niesforni uczniowie] i [nauczyciele]] *pojechali* na wycieczkę.
- (11) [Niesforni [[uczniowie] i [nauczyciele]]] *pojechali* na wycieczkę.

W tej sytuacji prawdopodobne granice członów można wyznaczyć jedynie odwołując się do znaczeń poszczególnych słów. Zdroworozsądkowa semantyka nakazuje nam przyjąć interpretację (10). Poniżej znajduje się drzewo zależnościowe zdania (8) w interpretacji (10), opisane w standardzie UD, z zaznaczonymi istotnymi elementami koordynacji.



Człony koordynacji często połączone są spójnikami. Przez spójnik koordynacji rozumiem właśnie ten spójnik, który łączy jej człony. Zakładam, że każda konstrukcja współrzędnie złożona ma co najwyżej jeden spójnik. W przypadku koordynacji (10) jest to *i*.

Przez lewy i prawy człon koordynacji rozumiem odpowiednio pierwszy i ostatni człon występujący w zdaniu (niezależnie od tego, ile jest tych członów) – w (10) są to *Niesforne dzieci* oraz *nauczyciele*.

Główą członu jest token, który nie ma nadziednika w obrębie tego członu – dla członu *Niesforni uczniowie* będzie nim słowo *uczniowie*.

Nadziednik koordynacji to token, który jest najbliższym wspólnym przodkiem wszystkich członów koordynacji i jej spójnika. W koordynacji (10) będzie to słowo *pojechali*.

Ostatnią ważną charakterystyką konstrukcji współrzędnie złożonej jest pozycja nadziednika – parametr określający umiejscowienie nadziednika koordynacji względem

ozn.	pozycja	przykład zdania
(L)	po lewej stronie	Drzewo <i>sadzą</i> [[Pat] i [Mat]].
(0)	brak nadziednika	[[Posadzili] i [podlali]] drzewo.
(R)	po prawej stronie	[[Pat] i [Mat]] <i>posadzili</i> drzewo.
(M)	po środku	[[Pat] – <i>powtórzyłem</i> – oraz [Mat]].

Tabela 2.1: Pozycja nadziednika

jej członów. Tabela 2.1 przedstawia możliwe pozycje nadziednika wraz z przykładowymi zdaniami.

Konstrukcje współrzędnie złożone z nadziednikiem po środku są bardzo rzadkie i nie występują w wielu językach (m.in. w języku angielskim). Oznacza to, że uzyskanie istotnych statystycznie wyników dla koordynacji (M) jest bardzo trudne. Ponadto, wstępna analiza koordynacji typu (M) w języku polskim wykazała, że mniej niż 20% z nich zostało opisane w sposób prawidłowy. W związku z tym w niniejszej pracy biorę pod uwagę wyłącznie koordynacje typu (L), (0) i (R).

2.4. Dependency Length Minimization

Szyk zdania ma wiele ograniczeń. Najlepiej opisane są te dotyczące sposobu ustawienia najważniejszych części zdania, takich jak podmiot, orzeczenie i dopełnienia. Nie mają one większego wpływu na ustawienie innych elementów zdania, takich jak między innymi umiejscowienie członów koordynacji.

W tej pracy skupiam się na zjawisku, które zdaniem wielu lingwistów istotnie wpływa na sposób układania słów w zdaniu – Dependency Length Minimization (DLM). W języku występuje naturalna tendencja do jak największego skracania sumy długości relacji zależnościowych. Innymi słowy, słowa układane są w takiej kolejności, żeby każde dwa słowa połączone bezpośrednio relacją składniową stały możliwie blisko siebie (Temperley, 2007).

Wyjaśnienie przyczyn tego zjawiska jest dość intuicyjne. King i Just (1991) pokazują, że podczas składania i odkodowywania zdań w umysłach użytkowników powstają reprezentacje relacji składniowych. Utrzymywanie tych reprezentacji w czasie wykorzystuje pamięć roboczą (King i Just, 1991, s. 596). Im dłuższa jest relacja, tym dłużej jej reprezentacja jest utrzymywana w pamięci. Zbyt długo utrzymywane reprezentacje powodują błędy w konstrukcji i rozumieniu zdań. W związku z tym, w celu oszczędzania pamięci roboczej i zmniejszania liczby błędów, należy przechowywać aktywne reprezentacje zależności składniowych jak najkrócej, a co za tym idzie, minimalizować długość zależności.

W badaniach nad DLM używane są rozmaite sposoby mierzenia długości zależności. Do najpopularniejszych jednostek miary należą morfemy, sylaby i fonemy (Lohmann,

2014). Przepiórkowski i Woźniak (2023) w swoim badaniu mierzą długość członów w słowach, sylabach i znakach. W celu replikacji badań w niniejszej analizie stosuję identyczną metodologię.

Lohmann (2014) wskazuje, że najlepszą metodą pomiaru długości członu jest prawdopodobnie złożoność syntaktyczna. Należy przez to rozumieć liczbę węzłów w relacjach składniowych łączących słowa wchodzące w skład członu. Większa złożoność syntaktyczna przekłada się bezpośrednio na większe zaangażowanie pamięci roboczej w przetwarzanie języka.

Spośród używanych przeze mnie miar najlepszym estymatorem złożoności syntaktycznej frazy jest liczba słów. W związku z tym, na potrzeby dyskusji wyników mojej pracy uznaję liczbę słów w członie za najpewniejszą miarę długości członu konstrukcji współrzędnie złożonej.

Uniwersalne występowanie DLM w językach naturalnych jest potwierdzone w badaniach (Futrell i in., 2015). W tej pracy zakładam, że jest to istotny czynnik mający także wpływ na ustawnienie członów koordynacji.

2.5. Kolejność członów koordynacji

DLM nie jest jedynym czynnikiem wpływającym na kolejność ustawniacia członów koordynacji. Lohmann (2014) opisuje wiele innych zjawisk, które to robią. W niniejszym punkcie omawiam najważniejsze z nich i opisuję ich możliwe interakcje z DLM.

Konwencja lub następstwo czasowe W przypadku zdania (13) kolejność doboru członów wynika z konwencji, natomiast zdania (14) i (15) mają zupełnie inne znaczenie. W przykładach takich jak powyższe nie można więc mówić o determinowaniu ustawnienia kolejności członów przez DLM.

- (13) [[Panie] i [Panowie]]!
- (14) [[Pobrali się] i [mieli dziecko]].
- (15) [[Mieli dziecko] i [pobrali się]].

Te tendencje działają jednak w dwie strony (w tym samym oraz w przeciwnym kierunku, co DLM). W związku z tym wychodzę z założenia, że podczas analizy obszernych korpusów językowych ich łączny wpływ na kolejność członów koordynacji jest pomijalny.

Czynniki pragmatyczne Lohmann (2014) wymienia rozmaite czynniki pragmatyczne wpływające na ustawnienie członów koordynacji. Wskazuje różne hipotezy, zgodnie z którymi pierwszeństwo mają człony mówiące o obiektach bliższych lub lepiej znanych nadawcy wypowiedzi.

Warto zwrócić uwagę na sytuacje, w których koordynacja wprowadza do dyskursu nowe obiekty.

- (16) [[Paweł] i [jego sąsiad Gaweł]]
- (17) ?? [Jego sąsiad Gaweł] i [Paweł]

W przykładzie (16) treść członu *[Paweł]* stanowi punkt odniesienia do wprowadzenia kolejnego obiektu, czyli treści członu *[jego sąsiad Gaweł]*. W takiej sytuacji ustawienie członów w odwrotnej kolejności jest niepoprawne pragmatycznie (zob. (17)). Punkty odniesienia, takie jak *Paweł*, są co do zasady opisywane w krótszy sposób niż nowe obiekty, takie jak *Gaweł*. Oznacza to, że w takich sytuacjach krótszy człon występuje po lewej stronie koordynacji niezależnie od DLM. Niemniej jednak, jest to sytuacja występująca relatywnie rzadko.

Częstość używania słowa Fenk-Oczlon (1989) stawia hipotezę, że człony składające się z częściej występujących w języku słów częściej pojawiają się jako pierwsze w konstrukcjach współrzędnie złożonych. Ponieważ w języku naturalnym krótsze słowa występują co do zasady częściej niż długie (Zipf, 1946), to zjawisko może powodować ustawianie krótszego członu na początku koordynacji (Lohmann, 2014, s. 54).

Zgodnie z tą hipotezą, koordynacja (18) ma większą szansę na pojawienie się w języku niż (19), ponieważ w języku polskim słowo *pies* występuje częściej niż *hipopotam* (Przepiórkowski, 2012).

- (18) [[Pies] i [hipopotam]]
- (19) [[Hipopotam] i [pies]]

Prozodia i akcent Kolejnym czynnikiem mającym duży wpływ na ustawienie członów koordynacji jest prozodia. Lohmann (2014) wskazuje, że koordynacje o krótkich członach są konstruowane tak, żeby akcentowane sylaby tworzyły rytm.

- (20) [[Jan] i | Maria]
- (21) [[Maria] i Jan]]

W przykładzie (20) sylaby nieakcentowane i akcentowane występują naprzemianie. Dzięki temu zdanie zawierające taką koordynację zawiera rytm oparty o stopy metryczne². Natomiast w przykładzie (21) sylaby nie tworzą sekwencji rytmicznej. Powoduje to, że konstrukcja (20) ma większe szanse pojawić się w języku, niż (21) (McDonald i in. 1993, Wright i in. 2005).

²Na schemacie \cup oznacza sylabę krótką (nieakcentowaną), zaś — sylabę długą (akcentowaną). Jeśli w zdaniu można wydzielić stopy, są one rozdzielone znakiem | . Sekwencje | \cup — | oraz | — \cup | tworzą odpowiednio jamb i trochej – dwie najprostsze stopy metryczne.

Na ostatnie dwa czynniki (częstość występowania słowa oraz rozkład sylab akcentowanych) znaczący wpływ ma długość członów. Dotyczą one głównie koordynacji krótkich, o członach jedno- lub dwuwyrzazowych i o niewielkich różnicach długości członów. Oznacza to, że mogą one w tych przypadkach wzmacniać lub osłabiać efekt DLM. Z tego powodu w niniejszym badaniu analizuję również koordynacje dłuższe i takie, w których różnica długości członów jest znaczna.

Skutki istnienia innych czynników mających wpływ na ustawienie członów koordynacji opisuję dokładniej w rozdziale 7.

2.6. Języki inicjalne oraz finalne

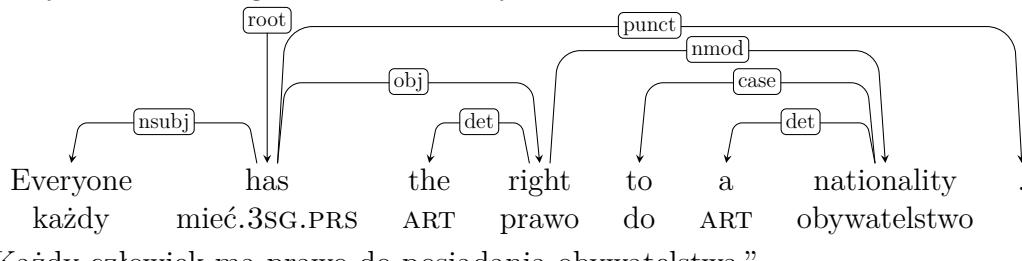
Jedną z cech, według których można sklasyfikować języki naturalne, jest generalna pozycja nadrzednika względem podrzednika. Ze względu na ten parametr w jazykoznawstwie wyróżnia się dwa rodzaje języków: inicjalne (ang. *head-initial*) oraz finalne (ang. *head-final*). Klasyfikacja ta oparta jest na cechach struktur gramatycznych występujących w danym języku, w tym przede wszystkim na szyku zdaniowym (Polinsky i Magyar, 2020).

W językach inicjalnych występuje generalna tendencja do umieszczania nadrzednika przed podrzednikiem. Do tej kategorii należą wiele języków indoeuropejskich, między innymi angielski, francuski, hiszpański, grecki oraz polski. W językach finalnych natomiast nadrzednik zwykle umieszczany jest za podrzednikiem. Przykładami takich języków są japoński, koreański oraz turecki (Polinsky, 2012).

Podział ten nie jest zero-jedynkowy. Istnieje wiele przykładów granicznych oraz takich, w których przypadek tendencja jest słaba. Na przykład język niemiecki jest zasadniczo finalny, jednak występuje w nim wiele struktur gramatycznych nadających mu cechy języka inicjalnego (Polinsky, 2012). W niniejszej pracy języki nienależące ściśle do żadnej z tych dwóch kategorii nazywam językami mieszanymi.

Poniżej znajdują się drzewa zależnościowe przykładowego zdania w języku typowo inicjalnym (angielskim) oraz finalnym (tureckim). W (22) widoczna jest tendencja do umieszczania głów członów po lewej, zaś w (23) po prawej stronie zdania.

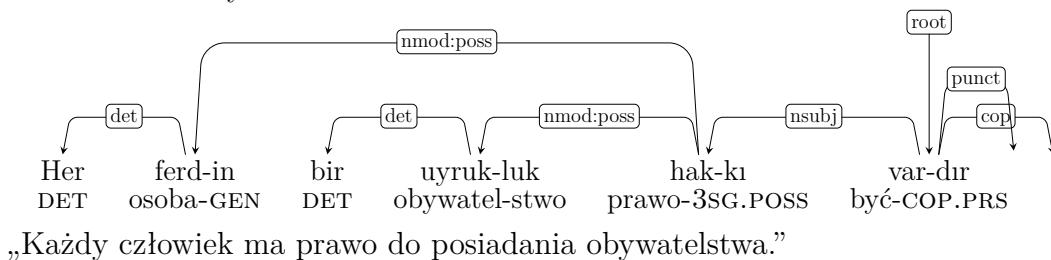
- (22) Everyone has the right to a nationality.³



„Każdy człowiek ma prawo do posiadania obywatelstwa.”

³Powszechna Deklaracja Praw Człowieka, art. 15, ust. 1. <https://www.un.org/en/about-us/universal-declaration-of-human-rights>

- (23) Her ferdin bir uyrulkuk hakkı vardır.⁴

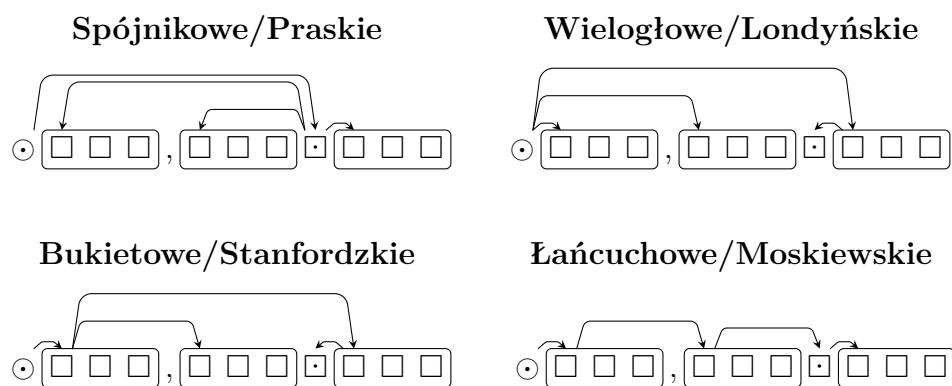


„Każdy człowiek ma prawo do posiadania obywatelstwa.”

2.7. Struktura zależnościowa koordynacji

W językoznawstwie nie ma pełnej zgody co do tego, jak opisywać relacje zależnościowe. Kontrowersje dotyczą nie tylko nazw i rodzajów zależności (czyli etykiet na krawędziach), lecz także struktury zależnościowej struktur gramatycznych (czyli tego, które tokeny są połączone z którymi w ramach danej struktury składniowej).

Istnieją cztery główne podejścia do opisu struktury konstrukcji współrzędnie złożonych. Przedstawiają je poniższe schematy, w których \odot oznacza nadzrębniaka koordynacji, \square spójnik, \Box pozostałe tokeny, zaś duże prostokąty symbolizują granice członów koordynacji. Relacje wewnętrzczne członów koordynacji nie są zaznaczone⁵.



Przepiórkowski i Woźniak (2023) pokazują, że jedynie podejścia praskie oraz londyńskie mogą poprawnie opisywać strukturę zależnościową koordynacji przy założeniu o poprawności DLM. Wynika to z dynamiki proporcji koordynacji z krótszym lewym członem w zależności od obecności i pozycji nadzrębniaka oraz z założenia, że umieszczenie krótszego członu na początku koordynacji uległo gramatykalizacji. Rozumowanie to jest szczegółowo przedstawione w rozdziale 3.

Przepiórkowski i Woźniak (2023) opierają swoją analizę na dwóch istotnych założeniach. Po pierwsze, autorzy zakładają, że gramatykalizacja umieszczania krótszego

⁴Tłumaczenie pochodzi ze strony <https://www.ohchr.org/en/human-rights/universal-declaration/translations/turkish-turkce>. Słowa w języku tureckim zostały zanalizowane przy użyciu TRmorph (Cöltekin, 2010) oraz oznakowane na podstawie reguł określonych w pracach Haspelmath (2014) i Bedir i in. (2021). Dziękuję Berkemu Şenşekerci za sprawdzenie glos.

⁵Poniższe schematy oparte są na schematach z pracy Przepiórkowski i Woźniak (2023).

członie po lewej stronie w języku angielskim wynika z faktu, że lewe człony są istotnie częściej krótsze niż prawe. Po drugie, głowa członu co do zasady znajduje się na początku członu.

W niniejszej pracy pokazuję, że te założenia prawdziwe są jedynie w przypadku języków inicjalnych oraz przedstawiam analogiczną analizę dla języków finalnych.

Rozdział 3

Struktura zależnościowa koordynacji

3.1. Języki inicjalne (Przepiórkowski i Woźniak, 2023)

3.1.1. Metody

Przepiórkowski i Woźniak (2023) pokazują, że przyjęcie konkretnego modelu struktury zależnościowej koordynacji pozwala na predykcję częstości występowania konkretnych rodzajów zdań w języku naturalnym. Rozpatrują oni sześć typów konstrukcji współrzędnie złożonych ze względu na pozycję nadrzędnika koordynacji oraz pozycję krótszego członu⁶:

(L-L)	○ [□ □ □] · [□ □ □ □ □ □]	Nadrzędnik po lewej, krótszy człon po lewej.
(L-R)	○ [□ □ □ □ □] · [□ □ □]	Nadrzędnik po lewej, krótszy człon po prawej.
(0-L)	[□ □ □] · [□ □ □ □ □ □]	Brak nadrzędnika, krótszy człon po lewej.
(0-R)	[□ □ □ □ □ □] · [□ □ □]	Brak nadrzędnika, krótszy człon po prawej.
(R-L)	[□ □ □] · [□ □ □ □ □ □] ○	Nadrzędnik po prawej, krótszy człon po lewej.
(R-R)	[□ □ □ □ □ □] · [□ □ □] ○	Nadrzędnik po prawej, krótszy człon po prawej.

Analiza dotyczy wyłącznie koordynacji binarnych, tj. posiadających dwa człony.

Przy założonej pozycji nadrzędnika, w języku naturalnym wraz ze wzrostem różnicy długości członów koordynacji coraz częściej pojawiają się zdania tego typu, dla którego suma długości relacji zależnościowych jest mniejsza. W związku z tym, każdy z modeli przewiduje, czy wraz ze wzrostem różnicy długości członów koordynacji:

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L)⁷ rośnie, czy spada;

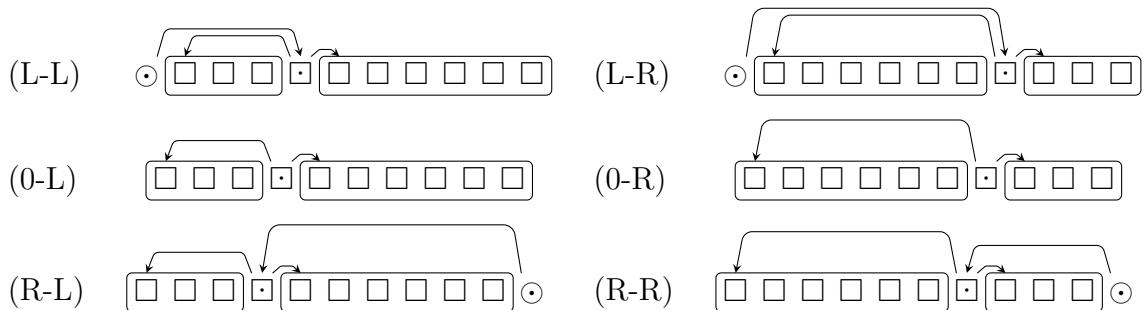
⁶W niniejszej pracy przy nazywaniu typów koordynacji przyjęto następującą konwencję: Pierwsza litera oznacza pozycję nadrzędnika (L, 0 lub R), zaś druga – pozycję krótszego członu (L lub R).

⁷Tj. procent koordynacji o krótszym pierwszym (lewym) członie wśród koordynacji z nadrzędniakiem po lewej stronie.

- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) rośnie, czy spada;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) rośnie, czy spada.

3.1.2. Podejścia

Podejście praskie



W celu predykcji tendencji do umieszczania krótszego członu koordynacji należy policzyć sumę długości relacji zależnościowych. Ze schematu wynika, że:

- w zdaniach (L-L) suma długości relacji jest **mniejsza**, niż w zdaniach (L-R);
- w zdaniach (0-L) suma długości relacji jest **mniejsza**, niż w zdaniach (0-R);
- w zdaniach (R-L) suma długości relacji jest **taka sama** jak w zdaniach (R-R).

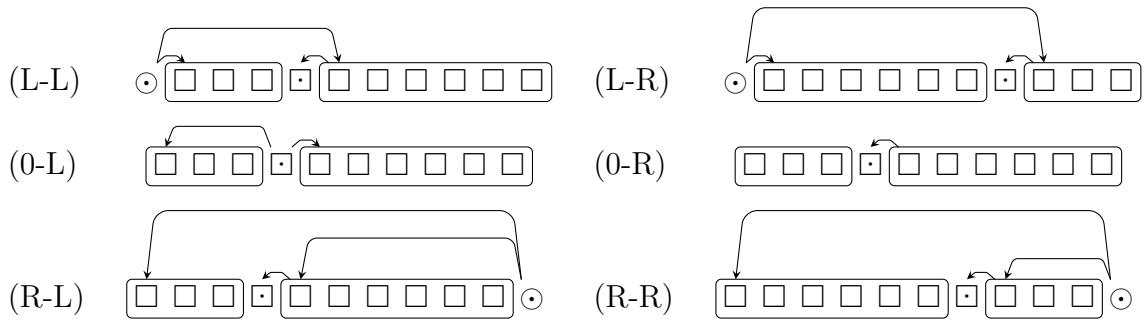
Powyzsze różnice rosną wraz ze wzrostem różnicy długości członów. Ze względu na efekt DLM, użytkownicy języka są skłonni tworzyć zdania o krótszej łącznej długości członów tym częściej, im bardziej mają możliwość skrócić łączną długość relacji. Oznacza to, że wielkość różnicy długości członów koordynacji przekłada się bezpośrednio naczęstość występowania zdań z krótszym lewym członem.

Na tej podstawie Przepiórkowski i Woźniak (2023) wyciągają wniosek, że model praski przewiduje, że wraz ze wzrostem różnicy długości członów:

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **rośnie**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **nie zmienia się**.

W następnych punktach analogiczne rozumowanie dla pozostałych podejść przedstawione są w sposób skrócony.

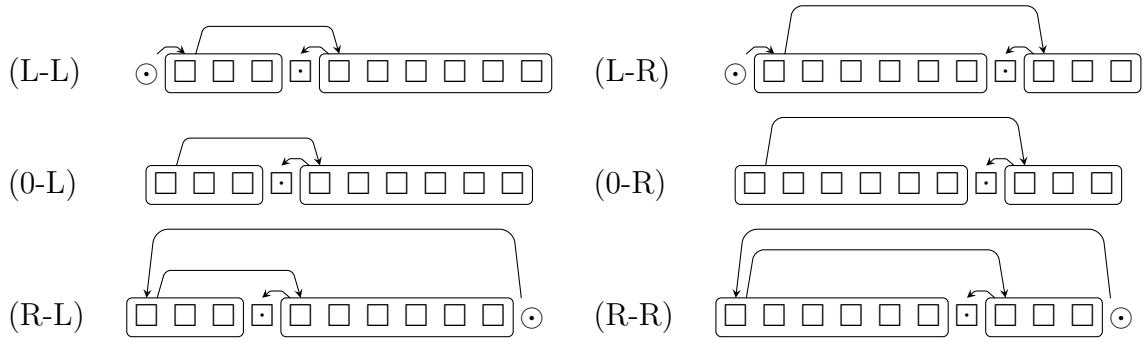
Podejście londyńskie



Model londyński przewiduje, że wraz ze wzrostem różnicy długości członów:

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **nie zmienia się**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **spada**.

Podejścia stanfordzkie i moskiewskie



Model stanfordzki przewiduje, że wraz ze wzrostem różnicy długości członów:

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **rośnie**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **rośnie**.

W przypadku koordynacji dwuczlonowych podejście moskiewskie przewiduje bardzo podobną strukturę koordynacji, co podejście stanfordzkie. Modele te różnią się jedynie dwiema krawędziami – jedną, łączącą głowę lewego członu z głową prawego członu oraz drugą, łączącą głowę prawego członu ze spójnikiem koordynacji. Różnice te nie mają istotnego wpływu na sumę długości relacji. Z tego powodu przewidywania modelu moskiewskiego co do występowania częstości zmian są identyczne, jak w przypadku modelu stanfordzkiego.

3.1.3. Wyniki i interpretacja

Przepiórkowski i Woźniak (2023) przeprowadzili analizę 21 825 koordynacji binarnych występujących w korpusie Penn Treebank (PTB) języka angielskiego.

Analiza wykazała, że wraz ze wzrostem różnicy długości członów:

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **rośnie**;
- **nie ma istotnej statystycznie tendencji** dotyczącej zmiany odsetka koordynacji (R) względem wszystkich koordynacji (R).

Pierwsze dwie z opisywanych tendencji uzyskały wysoką istotność statystyczną ($p < 0,001$). Nieznacznie malejąca prawidłowość opisująca koordynacje (R) nie była istotna statystycznie ($p = 0,921$). Niemniej jednak Przepiórkowski i Woźniak (2023) wskazują, że tendencja dotycząca koordynacji z nadzrębnikiem po prawej stronie jest istotnie inna od pozostałych dwóch zależności.

Wykres przedstawiający modele regresji logistycznych opisujących wyżej omówione zależności stanowi dodatek B do niniejszej pracy.

Jak zauważają Przepiórkowski i Woźniak (2023), taki stan rzeczy pokrywa się z przewidywaniami modelu praskiego. Niemniej jednak zauważają również możliwy wpływ innej tendencji na te zjawiska. Stawiają tezę, że ponieważ w języku angielskim zdania z krótszym lewym członem występują częściej, to umieszczanie krótszego członu po lewej stronie mogło ulec gramatykalizacji. Siła tej tendencji jest nieznana, jednak można przypuszczać, że jest ona na tyle duża, żeby zrównoważyć tendencję do tworzenia zdań typu (R-R) częściej niż (R-L) oraz wpływać naczęstość stawiania krótszego członu po lewej stronie, gdy nie ma nadzrębniaka. W takiej sytuacji należałoby również dopuścić interpretację londyńską.

Wykorzystując powyższe rozumowanie, Przepiórkowski i Woźniak (2023) argumentują, że podejścia stanfordzkie i moskiewskie nie mogą poprawnie opisywać struktury zależnościowej koordynacji w języku angielskim.

3.2. Języki finalne

3.2.1. Różnice względem języków inicjalnych

Analizując strukturę zależnościową koordynacji w językach finalnych należy wziąć pod uwagę dwa dodatkowe fakty.

Po pierwsze, głowy członów zwykle znajdują się bliżej ich końców (wynika to w trywialny sposób z natury tych języków).

Po drugie, szeroko uznawane podejścia dotyczące struktury koordynacji zostały utworzone głównie na podstawie analizy języków inicjalnych. Kanayama i in. (2018) zauważają, że wyżej omawiane modele, w szczególności podejścia asymetryczne, nie nadają się do opisu języków finalnych. Podają przykłady struktur występujących w języku japońskim oraz koreańskim, które interpretowane zgodnie z narzucanym przez UD podejściem stanfordzkim tworzą drzewa niezgodne z teorią lingwistyczną. Podkreślają, że z tego powodu z japońskich korpusów Universal Dependencies zostały usunięte wszystkie koordynacje, a w korpusach koreańskich występują równolegle dwa różne standardy opisu struktur. Postulują, żeby w analizie języków finalnych dopuścić odwrócone podejście, według którego główna prawego członu jest nadrzędnikiem pozostałych (Kanayama i in., 2018).

Nie jest to jednak jedynie podejście do tematu struktury zależnościowej koordynacji w językach finalnych. Choi i Palmer (2011) proponują, żeby „każdy człon był pododziennikiem następnego”, tworząc podejście odwrócone względem podejścia moskiewskiego.

Niemniej jednak analizowanie języków finalnych według modeli zaproponowanych w pracach Kanayama i in. (2018) oraz Choi i Palmer (2011) nie jest idealnym rozwiąza niem. Podejścia te nie określają miejsca spójnika w strukturze zależnościowej. Ponadto, Kanayama i in. (2018) postulują inne modele struktury zależnościowej dla języków inicjalnych i finalnych. Może być to dobre rozwiąza nanie ad hoc w celu inkorporacji koordynacji do japońskich korpusów UD, jednak stoi ono w sprzeczności z podstawowym celem Universal Dependencies. Podział języków na inicjalne i finalne nie jest podziałem sztywnym, lecz raczej wskazaniem tendencji występujących w danym języku, więc nie ma podstaw, żeby sądzić, że struktury gramatyczne w językach inicjalnych i finalnych są różne. Struktury gramatyczne powinny być opisywane przez uniwersalne standardy.

W związku z powyższymi argumentami, w niniejszej pracy analizuję nie tylko predykaty „klasycznych” podejść opisanych w punkcie 3.1.2 oraz postulowanych w badaniach Kanayama i in. (2018) oraz Choi i Palmer (2011), lecz także przewidywania wszystkich możliwych modeli struktury zależnościowej koordynacji.

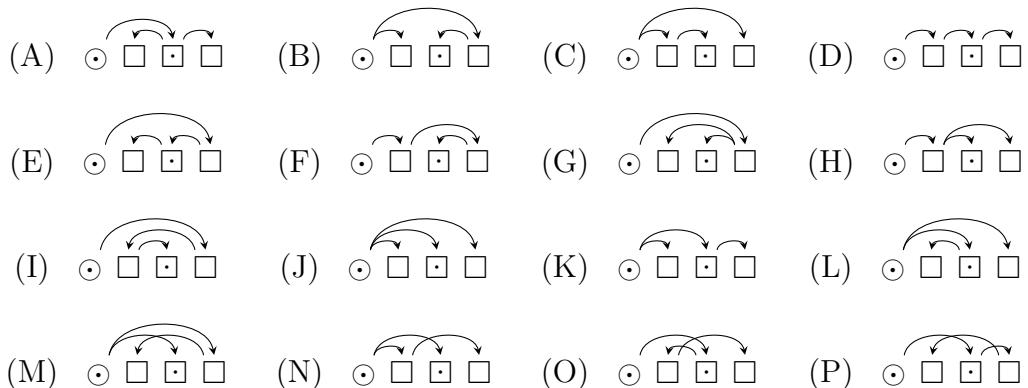
3.2.2. Podejścia

Dla analizy struktury zależnościowej koordynacji binarnej (zarówno w językach finalnych, jak w inicjalnych) istotne są jej cztery elementy:

- nadrzędnik,
- głowa pierwszego członu,
- spójnik,
- głowa ostatniego członu.

W celu uproszczenia na poniższych schematach posługuję się tymi symbolami. W analizie uwzględniam koordynacje wieloczłonowe, jednak uwzględniam tylko ich pierwszy (lewy) i ostatni (prawy) człon.

Istnieje 16 sposobów narysowania drzewa dla linearne uporządkowanego ciągu czterech wierzchołków z wyróżnionym korzeniem:



Powyższe modele należy interpretować w następujący sposób:

- (A) podejście praskie;
- (B) podejście londyńskie;
- (C) odwrócone podejście londyńskie – spójnik jest podpięty pod lewy człon;
- (D) podejście moskiewskie;
- (E) odwrócone podejście moskiewskie (postulowane w pracy Choi i Palmer, 2011);
- (F) podejście stanfordzkie (stosowane w UD);
- (G) odwrócone podejście stanfordzkie (postulowane w pracy Kanayama i in., 2018);
- (H)–(I) pozostałe warianty podejścia stanfordzkiego;
- (J) podejście „nadzędnikowe” (według którego głowy wszystkich członów oraz spójnik są podrzędnikami nadzędnika koordynacji);
- (K)–(L) podejścia zakładające, że jedna z głów jest bezpośrednim podrzędniakiem nadzędnika koordynacji, zaś druga podrzędniakiem spójnika;
- (M)–(P) Podejścia zakładające występowanie krawędzi nieprojektywnych w strukturze koordynacji.

Podejścia (M)–(P) zakładają, że w strukturze zależnościowej koordynacji znajdują się krawędzie nieprojektywne, tj. przecinające się z innymi. W języku naturalnym projektowność relacji zależnościowych uznawana jest za normę, zaś relacje nieprojektywne

występują tylko w niektórych strukturach gramatycznych. Koordynacja nie należy do tego typu struktur (Nivre, 2006). W związku z tym nie należy zakładać, że istnienie krawędzi nieprojektywnych w strukturze koordynacji jest regułą. Wobec tego w niniejszej pracy odrzucam podejścia (M)–(P) i zajmuję się predykcjami modeli (A)–(L).

3.2.3. Metody

Dla każdego z 12 podejść rozważam sześć sytuacji na wzór analizy z pracy Przepiórkowski i Woźniak (2023). Ponieważ analizuję korpusy języków finalnych, zakładam, że głowa członu (oznaczona tutaj \blacksquare) znajduje się na jego końcu⁸:

(L-L)	$\odot \square \square \blacksquare \square \cdot \square \square \square \square \square \blacksquare$	(L-R)	$\odot \square \square \square \square \square \blacksquare \square \cdot \square \square \blacksquare$
(0-L)	$\square \square \blacksquare \square \cdot \square \square \square \square \square \blacksquare$	(0-R)	$\square \square \square \square \square \blacksquare \square \cdot \square \square \blacksquare$
(R-L)	$\square \square \blacksquare \square \cdot \square \square \square \square \square \blacksquare \odot$	(R-R)	$\square \square \square \square \square \blacksquare \square \cdot \square \square \blacksquare \odot$

Wszystkie poniższe obliczenia dotyczą liczby tokenów w koordynacji. Przez a rozumiem długość krótszego członu, przez b – różnicę długości członów, zaś przez S – sumę długości relacji w obrębie koordynacji⁹. Na potrzeby obliczeń w schematach tokeny wchodzące w skład członu zamieniam na długość członu:

(L-L)	$\odot a \square \cdot a+b \square$	(L-R)	$\odot a+b \square \cdot a \square$
(0-L)	$a \square \cdot a+b \square$	(0-R)	$a+b \square \cdot a \square$
(R-L)	$a \square \cdot a+b \square \odot$	(R-R)	$a+b \square \cdot a \square \odot$

3.2.4. Predykcje

Podejście praskie (A) Model przewiduje następujące długości relacji:

(L-L)	$\odot \overbrace{a \square}^{\curvearrowright} \cdot \overbrace{a+b \square}^{\curvearrowright}$	$S = 2a + b$	(L-R)	$\odot \overbrace{a+b \square}^{\curvearrowright} \cdot \overbrace{a \square}^{\curvearrowright}$	$S = 2a + b$
(0-L)	$a \square \cdot \overbrace{a+b \square}^{\curvearrowright}$	$S = a + b$	(0-R)	$a+b \square \cdot \overbrace{a \square}^{\curvearrowright}$	$S = a$
(R-L)	$a \square \cdot \overbrace{a+b \square}^{\curvearrowright} \odot$	$S = 2a + 2b$	(R-R)	$a+b \square \cdot \overbrace{a \square}^{\curvearrowright} \odot$	$S = 2a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

⁸Jest to pewnego rodzaju uproszczenie, ponieważ w językach finalnych głowa nie zawsze znajduje się na samym końcu frazy. Niemniej jednak dla opisywanych przeze mnie tendencji istotne jest to, z której strony głowy członu częściej pojawiają się dodatkowe podrzędniki głowy. W przypadku języków finalnych zakładam, że człon częściej „rośnie” w lewą stronę od głowy.

⁹Dokładniej jest to suma długości relacji, na które ma wpływ długość członów koordynacji.

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **nie zmienia się**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

Podejście londyńskie (B) Model przewiduje następujące długości relacji:

(L-L)		$S = 4a + 2b$	(L-R)		$S = 4a + 2b$
(0-L)		$S = a + b$	(0-R)		$S = a$
(R-L)		$S = 2a + 2b$	(R-R)		$S = 2a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **nie zmienia się**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

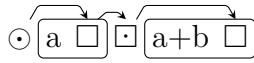
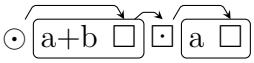
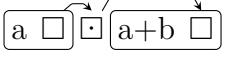
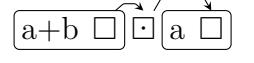
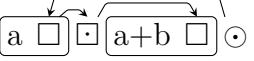
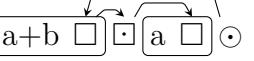
Odwrócone podejście londyńskie (C) Model przewiduje następujące długości relacji:

(L-L)		$S = 3a + b$	(L-R)		$S = 3a + 2b$
(0-L)		$S = 0$	(0-R)		$S = 0$
(R-L)		$S = a + b$	(R-R)		$S = a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **nie zmienia się**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **spada**.

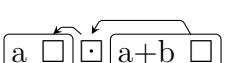
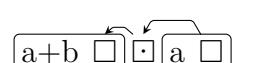
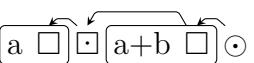
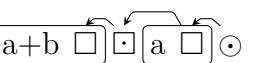
Podejście moskiewskie (D) Model przewiduje następujące długości relacji:

(L-L)		$S = 2a + b$	(L-R)		$S = 2a + b$
(0-L)		$S = a + b$	(0-R)		$S = a$
(R-L)		$S = 2a + 2b$	(R-R)		$S = 2a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **nie zmienia się**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

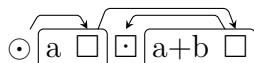
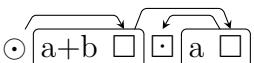
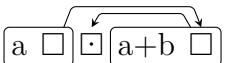
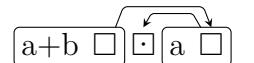
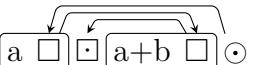
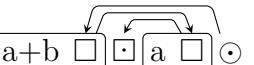
Odwrócone podejście moskiewskie (E) Model przewiduje następujące długości relacji:

(L-L)		$S = 3a + 2b$	(L-R)		$S = 3a + b$
(0-L)		$S = a + b$	(0-R)		$S = a$
(R-L)		$S = a + b$	(R-R)		$S = a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **spada**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **spada**.

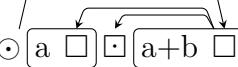
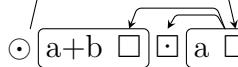
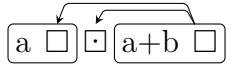
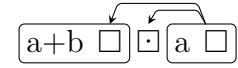
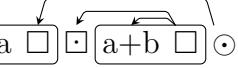
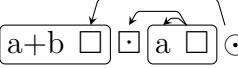
Podejście stanfordzkie (F) Model przewiduje następujące długości relacji:

(L-L)		$S = 3a + 2b$	(L-R)		$S = 3a + b$
(0-L)		$S = 2a + 2b$	(0-R)		$S = 2a$
(R-L)		$S = 3a + 3b$	(R-R)		$S = 3a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **spada**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **znacznie spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **bardzo znacznie spada**.

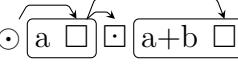
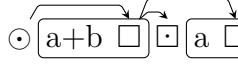
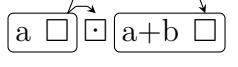
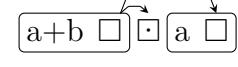
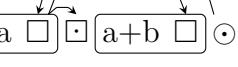
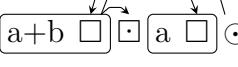
Odwrócone podejście stanfordzkie (G) Model przewiduje następujące długości relacji:

(L-L)		$S = 4a + 3b$	(L-R)		$S = 4a + b$
(0-L)		$S = 2a + 2b$	(0-R)		$S = 2a$
(R-L)		$S = 2a + 2b$	(R-R)		$S = 2a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **znacznie spada**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **znacznie spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

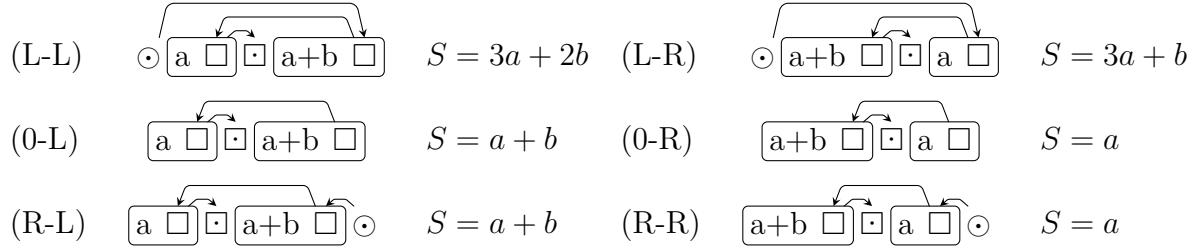
Podejście (H) Model przewiduje następujące długości relacji:

(L-L)		$S = 2a + b$	(L-R)		$S = 2a + b$
(0-L)		$S = a + b$	(0-R)		$S = a$
(R-L)		$S = 2a + 2b$	(R-R)		$S = 2a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **nie zmienia się**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

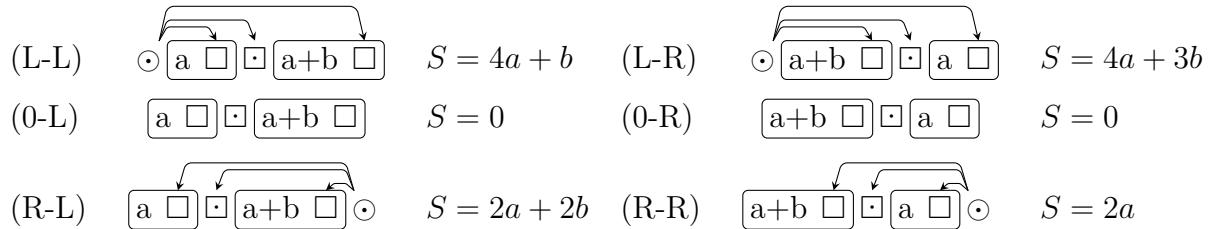
Podejście (I) Model przewiduje następujące długości relacji:



To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **spada**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **spada**.

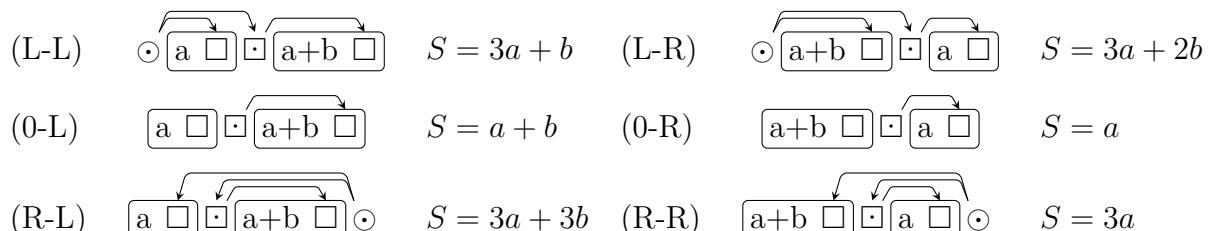
Podejście nadzędnikowe (J) Model przewiduje następujące długości relacji:



To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **znacznie wzrasta**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **nie zmienia się**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **znacznie spada**.

Podejście (K) Model przewiduje następujące długości relacji:



To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **bardzo znacznie spada**.

Podejście (L) Model przewiduje następujące długości relacji:

$$\begin{array}{llll}
 (\text{L-L}) & \textcircled{\text{S}} \xrightarrow{\quad} \boxed{a} \square \cdot \boxed{a+b} \square & S = 3a + b & (\text{L-R}) \quad \textcircled{\text{S}} \xrightarrow{\quad} \boxed{a+b} \square \cdot \boxed{a} \square \quad S = 3a + 2b \\
 (\text{0-L}) & \boxed{a} \square \xrightarrow{\quad} \boxed{a+b} \square & S = 0 & (\text{0-R}) \quad \boxed{a+b} \square \xrightarrow{\quad} \boxed{a} \square \quad S = 0 \\
 (\text{R-L}) & \boxed{a} \square \xrightarrow{\quad} \boxed{a+b} \square \cdot \textcircled{\text{S}} & S = a + b & (\text{R-R}) \quad \boxed{a+b} \square \cdot \boxed{a} \square \xrightarrow{\quad} \textcircled{\text{S}} \quad S = a
 \end{array}$$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **nie zmienia się**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **spada**.

W dalszej części pracy przedstawiam analizę danych w poszukiwaniu wyżej opisanych tendencji. Pokazuję, że w językach finalnych występuje tendencja wzrostowa do zmiany odsetka występowania koordynacji (R-L) względem wszystkich koordynacji (R). Zależność ta nie jest przewidywana przez żaden z modeli struktury zależnościowej koordynacji.

3.3. Języki mieszane

Językami mieszonymi nazywam te, w przypadku których nie ma wyraźnej tendencji do umieszczania głowy na początku lub na końcu fraz. Należą do nich m.in. niemiecczyzna i łacina (Polinsky i Magyar, 2020).

W niniejszej pracy obliczam opisywane powyżej tendencje w języku niemieckim i łacińskim. Ponieważ jednak w językach mieszanych nie sposób ustalić, z której strony członu znajduje się zwykle główna i z której strony głowy zwykle pojawiają się jej podrzędni, nie przedstawiam przewidywań modeli struktury zależnościowej koordynacji co do tych tendencji.

Rozdział 4

Przetwarzanie danych

4.1. Dane wejściowe

4.1.1. Korpusy zależnościowe

Analizie zostały poddane 72 korpusy zależnościowe 13 języków, spośród korpusów dostępnych na stronie internetowej Universal Dependencies (<https://universaldependencies.org/>).

Za główne kryterium doboru języka uznano istnienie korpusu zależnościowego opisaneego w standardzie UD o objętości co najmniej 700 tysięcy tokenów.

Język arabski został wykluczony z analizy, ponieważ znaczna część jego korpusów zawierała wyłącznie relacje zależnościowe między tokenami. W takich korpusach treść zdań została zamieniona podkreśnikami, w związku z czym policzenie długości członów w słowach, sylbach i znakach nie była możliwa. Korpusy języka arabskiego zawierające słowa nie przekroczyły łącznej objętości 700 tys. tokenów.

Ze względu na fakt, że w korpusach UD dla języka japońskiego nie występują koordynacje (Kanayama i in., 2018, s. 79), język ten również został wykluczony z analizy.

Ponadto do analizy przyjęto korpusy dwóch języków finalnych z największymi korpusami, nie licząc wykluczonych japońskiego i arabskiego, tj. koreańskiego i tureckiego.

Dodatkowo uwzględniono korpusy języka polskiego jako języka ojczystego autora pracy.

Tabela 4.1 przedstawia informacje na temat korpusów użytych w badaniu.

Język	Korpus	Rozmiar	Sposób anotacji relacji zależnościowych
Języki inicjalne			
angielski	GUM	184 478	ręcznie w formacie UD
	EWT	251 534	ręcznie w formacie UD
	Atis	61 879	ręcznie w formacie UD
	ParTUT	49 602	ręcznie w innym formacie, konwersja z poprawkami
	GENTLE	17 617	ręcznie w formacie UD
	PUD	21 058	ręcznie w formacie UD
	LinES	93 200	ręcznie w innym formacie, konwersja z poprawkami
	Pronouns	1 640	ręcznie w formacie UD
	ESLSpok	21 312	ręcznie w formacie UD
	GUMReddit	15 960	ręcznie w formacie UD
Razem		718 280	
czeski	CAC	494 420	ręcznie w innym formacie, automatyczna konwersja
	PDT	1 527 257	ręcznie w innym formacie, automatyczna konwersja
	FicTree	166 747	ręcznie w innym formacie, automatyczna konwersja
	CLTT	36 011	ręcznie w innym formacie, automatyczna konwersja
	PUD	18 578	ręcznie w formacie UD
	Poetry	6 273	ręcznie w formacie UD
	Razem	2 249 286	
hiszpański	AnCora	555 670	ręcznie w innym formacie, automatyczna konwersja
	GSD	423 345	ręcznie w innym formacie, automatyczna konwersja
	PUD	22 822	ręcznie w innym formacie, automatyczna konwersja
	Razem	1 001 837	
islandzki	Modern	80 392	ręcznie w innym formacie, automatyczna konwersja
	IcePaHC	983 671	ręcznie w innym formacie, automatyczna konwersja
	PUD	18 831	automatycznie z poprawkami
	GC	99 611	ręcznie w innym formacie, automatyczna konwersja
	Razem	1 182 505	
polski	PDB	347 319	ręcznie w innym formacie, automatyczna konwersja
	LFG	130 967	ręcznie w innym formacie, automatyczna konwersja
	PUD	18 333	ręcznie w innym formacie, automatyczna konwersja
	Razem	496 619	
portugalski	PetroGold	232 333	ręcznie w formacie UD
	Porttinari	157 490	ręcznie w formacie UD
	Bosque	210 958	ręcznie w innym formacie, konwersja z poprawkami
	CINTIL	441 991	ręcznie w innym formacie, automatyczna konwersja
	GSD	296 169	ręcznie w innym formacie, automatyczna konwersja
	PUD	21 917	ręcznie w innym formacie, automatyczna konwersja
	Razem	1 360 858	
rosyjski	Taiga	197 001	ręcznie w formacie UD
	Poetry	64 112	ręcznie w formacie UD
	SynTagRus	1 517 881	ręcznie w innym formacie, automatyczna konwersja
	GSD	97 994	ręcznie w formacie UD
	PUD	19 355	ręcznie w formacie UD
	Razem	1 896 343	

Język	Korpus	Rozmiar	Sposób anotacji relacji zależnościowych
Języki inicjalne			
rumuński	RRT	218 522	ręcznie w formacie UD
	SiMoNERo	146 020	ręcznie w innym formacie, automatyczna konwersja
	ArT	573	ręcznie w formacie UD
	Nonstandard	572 436	ręcznie w innym formacie, automatyczna konwersja
	Razem	937 551	
włoski	ISDT	278 460	ręcznie w innym formacie, automatyczna konwersja
	VIT	259 625	ręcznie w innym formacie, automatyczna konwersja
	Old	40 386	ręcznie w formacie UD
	ParTUT	51 614	ręcznie w innym formacie, konwersja z poprawkami
	ParlaMint	19 141	ręcznie w formacie UD
	TWITTIRO	28 384	ręcznie w innym formacie, automatyczna konwersja
	Valico	6 508	ręcznie w formacie UD
	PoSTWITA	119 334	automatycznie z poprawkami
	MarkIT	38 237	ręcznie w formacie UD
	PUD	22 182	ręcznie w innym formacie, automatyczna konwersja
	Razem	863 871	
Języki mieszane			
łacina	ITTB	450 480	ręcznie w innym formacie, konwersja z poprawkami
	LLCT	242 391	ręcznie w innym formacie, konwersja z poprawkami
	UDante	55 286	ręcznie w formacie UD
	Perseus	28 868	ręcznie w innym formacie, automatyczna konwersja
	PROIEL	205 566	ręcznie w innym formacie, automatyczna konwersja
	Razem	982 591	
niemiecki	GSD	287721	ręcznie w innym formacie, automatyczna konwersja
	PUD	21 001	ręcznie w innym formacie, automatyczna konwersja
	LIT	40 340	ręcznie w formacie UD
	HDT	3 399 390	ręcznie w innym formacie, konwersja z poprawkami
	Razem	3 748 452	
Języki finalne			
koreański	Kaist	350 090	ręcznie w innym formacie, automatyczna konwersja
	GSD	80 322	ręcznie w innym formacie, automatyczna konwersja
	PUD	16 584	ręcznie w innym formacie, automatyczna konwersja
	Razem	446 996	
turecki	Kenet	178 658	ręcznie w innym formacie, automatyczna konwersja
	Penn	183 555	ręcznie w innym formacie, automatyczna konwersja
	Tourism	91 152	ręcznie w innym formacie, automatyczna konwersja
	Atis	45 907	ręcznie w formacie UD
	GB	16 803	ręcznie w formacie UD
	FrameNet	19 223	ręcznie w innym formacie, automatyczna konwersja
	BOUN	121 835	ręcznie w formacie UD
	IMST	56 422	ręcznie w innym formacie, automatyczna konwersja
	PUD	16 535	ręcznie w formacie UD
	Razem	730 090	

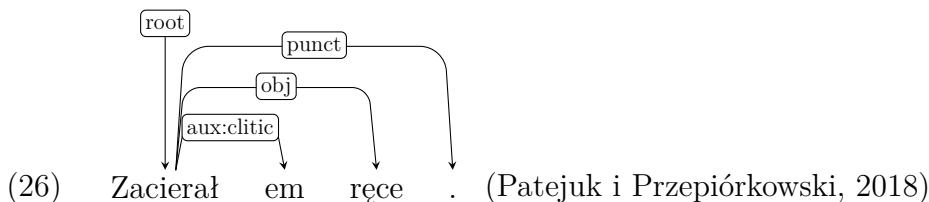
Tabela 4.1: Języki i korpusy analizowane w badaniu. Rozmiar korpusów podany jest w liczbie tokenów.

4.1.2. Format danych

Korpusy składają się ze zdań opisanych w formacie CONLL-U¹⁰, zawierającym wszystkie informacje potrzebne do utworzenia drzewa zależnościowego. Poniższe schematy przedstawiają przykładowe zdanie (24), jego opis w formacie CONLL-U (25) oraz drzewo zależnościowe (26).

(24) Zacierałem ręce.

(25) # sent_id = dev-1646
text = Zacierałem ręce.
converted_from_file = NKJP1M_1102000008_morph_6-p_morph_6.61-s-dis01.xml
genre = news
1 Zacierał zacierać VERB praet:sg:m1:imperf Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|
SubGender=Masc1|Tense=Past|VerbForm=Fin|Voice=Act 0 root 0:root SpaceAfter=No
2 em być AUX glt:sg:pri:imperf:wok Aspect=Imp|Number=Sing|Person=1| Variant=Long 1aux:clitic
1:aux:clitic _
3 ręce ręka NOUN subst:pl:acc:f Case=Acc|Gender=Fem|Number=Plur 1obj 1:obj SpaceAfter=No
4 . . PUNCT interp PunctType=Peri 1 punct 1:punct _



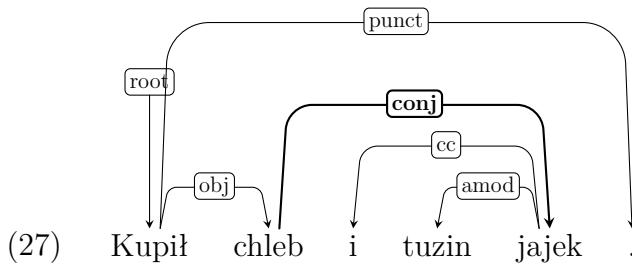
4.2. Wyciąganie koordynacji

W niniejszym punkcie omawiam proces wyciągania koordynacji, czyli zautomatyzowanego znajdowania i opisu konstrukcji współrzędnie złożonych w korpusach zależnościowych. Omawiana procedura zakłada, że zdania są anotowane w formacie UD, który przyjmuje stanfordzki model struktury koordynacji.

4.2.1. Relacja conj

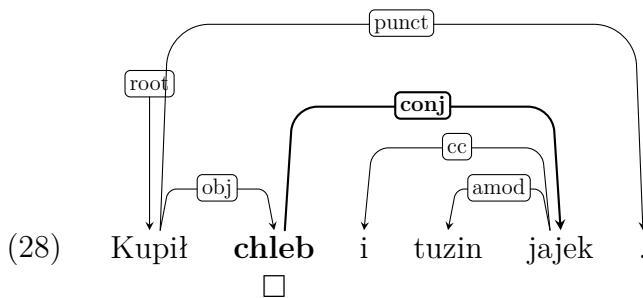
W standardzie Universal Dependencies zależność łącząca dwa człony koordynacji opisana jest zawsze etykietą `conj`. Taka etykieta sygnalizuje obecność konstrukcji współrzędnie złożonej, co pokazuje przykład (27). Ponieważ UD opisuje koordynacje według podejścia stanfordzkiego, to wiadomo, że każda krawędź drzewa podpisana etykietą `conj` łączy lewy człon koordynacji z jednym z pozostałych jej członów.

¹⁰Dokładny opis formatu znajduje się na stronie <https://universaldependencies.org/format.html>.

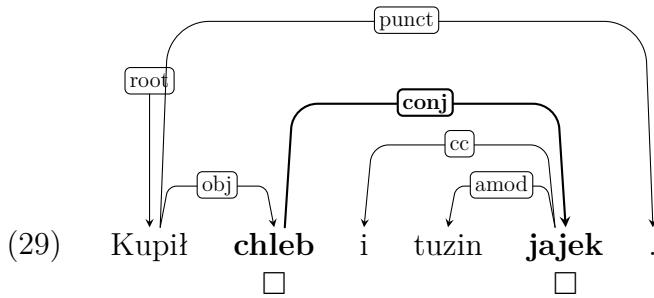


4.2.2. Wyznaczanie głów członów, nadrzędnika i spójnika koordynacji

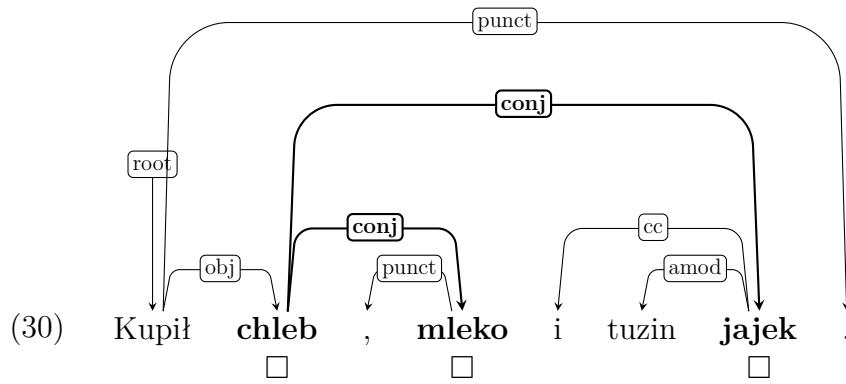
Głowa lewego członu Jeśli token jest nadrzędnikiem w relacji `conj`, to jest on głową lewego członu. W przykładzie (28) głową lewego członu jest token *chleb*.



Głowa prawego członu Jeśli z głowy lewego członu wychodzi tylko jedna relacja `conj`, to jej podrzędnik jest głową prawego członu. W (29) głową prawego członu jest token *jajek*.

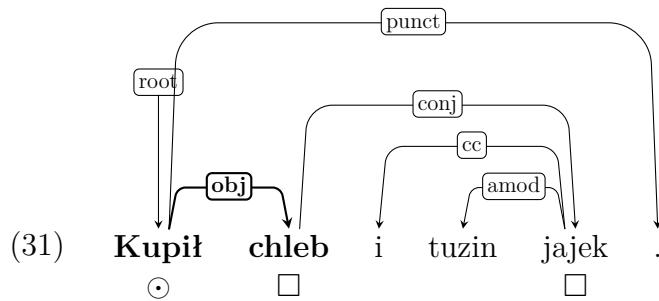


Jeśli natomiast głowa lewego członu ma kilka podrzędników z relacjami `conj`, jest to koordynacja wieloczłonowa. Podrzędni niki tych relacji to głowy pozostałych członów. Głowa, która występuje w zdaniu jako ostatnia, jest głową prawego członu. W zdaniu (30) tokeny *chleb*, *mleko* i *jajek* są głowami członów jednej konstrukcji współrzędnie złożonej. *Chleb* jest głową lewego, zaś *jajek* prawego członu koordynacji.

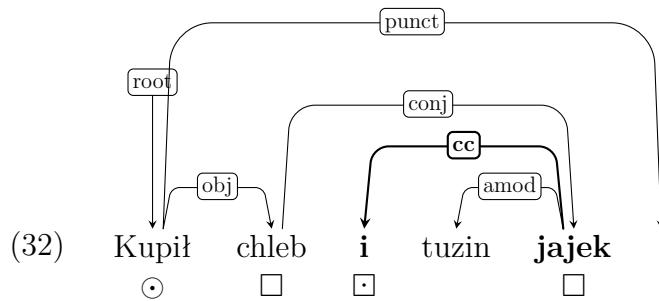


Ponieważ w analizie każda koordynacja traktowana jest jako binarna, środkowe człony (takie jak *mleko* w przykładzie (30)) są ignorowane.

Nadrzędnik Nadrzędnikiem koordynacji jest zawsze nadrzędnik głowy lewego członu. W przykładzie (31) jest to *Kupił*.



Spójnik Jeśli z głowy prawego członu wychodzi relacja *cc*, podrzędnik tej relacji jest spójnikiem koordynacji. W zdaniu (32) spójnikiem koordynacji *i*.



Podsumowując, procedura wyznaczania kluczowych elementów koordynacji wygląda następująco:

(33) Dla każdego tokenu *L*:

Jeśli *L* ma $n > 0$ podrzędników z relacją *conj* oraz:

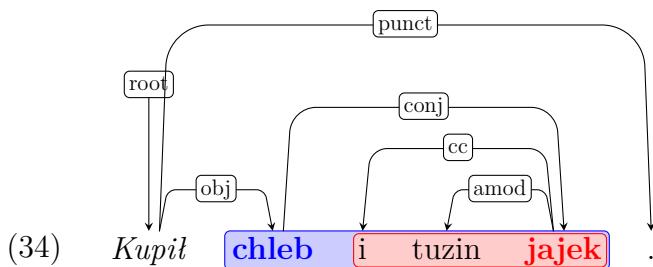
- nadrzędnik *L* to *G*,
- podrzędniki *L* z relacją *conj* to H_1, \dots, H_n ,
- opcjonalne podrzędniki H_1, \dots, H_n z relacją *cc* to spójniki – oznaczane odpowiednio C_1, \dots, C_n ,

to należy rozpatrzyć taką koordynację, w której nadziednikiem jest G , spójnikiem C_n , a głowami członów są L oraz H_1, \dots, H_n (przy czym głową lewego członu jest L , a głową prawego członu H_n).

H_i i C_i dla $i < n$ to odpowiednio pozostałe głowy i spójniki. Są one co do zasady pomijane w analizie¹¹.

4.2.3. Wyznaczanie granic członów

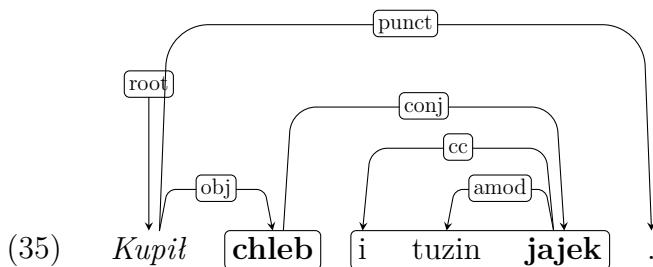
W celu automatycznego określenia granicy członu należy wziąć pod uwagę wszystkich potomków jego głowy:



W zdaniu (34) potomkami głowy lewego członu *chleb* są *i*, *tuzin* i *jajek*, zaś potomkami głowy prawego członu *i* oraz *tuzin*.

Spośród podrzędników głowy lewego członu należy wykluczyć głowy pozostałych członów (H_1, \dots, H_n) oraz wraz z ich podrzędnikami.

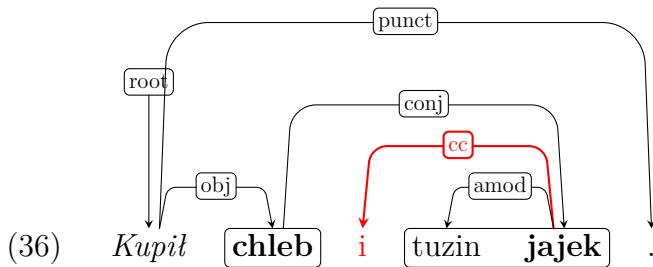
W ten sposób otrzymujemy wstępnie określone granice członów, co pokazuje przykład (35):



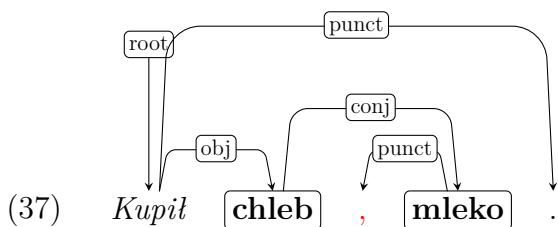
Następnie należy zastosować zestaw reguł w celu wykluczenia tokenów, które nie są elementami członów. W niniejszej pracy posługuję się następującymi heurystykami:

- (H1) Człon nie może zaczynać się od spójnika (słowa, które jest połączone z głową członu relacją cc).

¹¹Niektóre z heurystyk opisywanych w punktach 4.2.3 oraz 4.2.7 mogą brać pod uwagę istnienie lub treść średzkowych członów i pozostałych spójników. Odnoszę się do tego faktu przy omawianiu tych heurystyk.

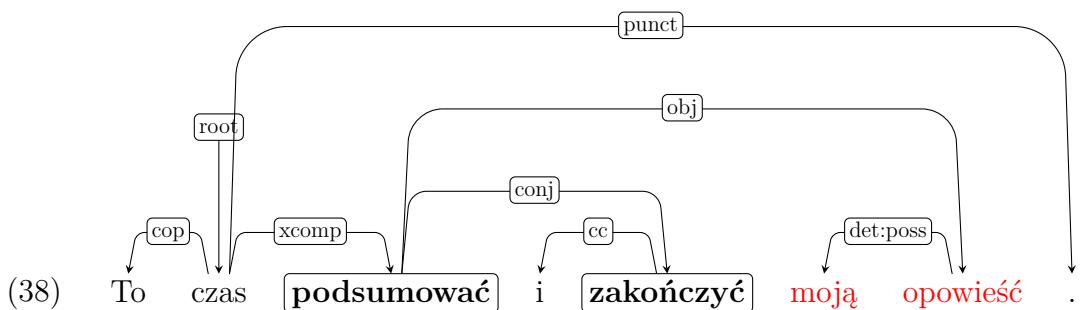


(H2) Człon nie może zaczynać się od znaku interpunkcyjnego (dokładniej rzecz ujmując, od przecinka, średnika, dwukropka ani myślnika).



(H3) Potomkowie głowy lewego członu znajdujący się na prawo od prawego członu nie wchodzą w skład lewego członu.

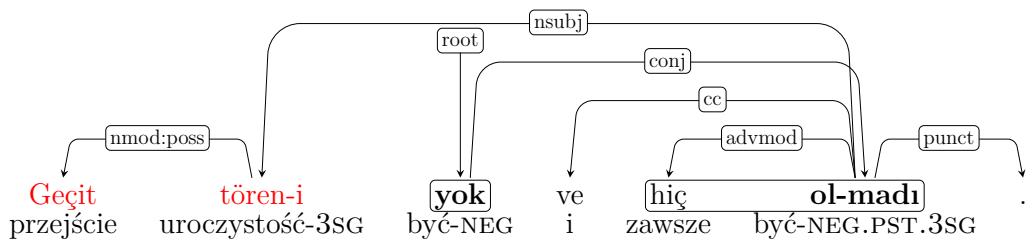
W rzeczywistości tokeny, o których mowa w (H3) są współdzielone przez oba człony, czyli należą do ich obu. Ponieważ interesuje mnie różnica długości członów, dla przejrzystości wykluczam ich wspólną część. W przykładzie (38) fraza *moją opowiesć* jest wykluczona z lewego członu na podstawie (H3).



(H4) Potomkowie głowy prawego członu znajdujący się na lewo od lewego członu nie wchodzą w skład prawego członu.

(H4) jest symetryczna względem (H3). Ma ona zastosowanie co do zasady wyłącznie w językach finalnych. W przykładzie (39) fraza *Geçit töreni* (parada) jest współdzielona przez oba człony koordynacji. W związku z tym nie jest traktowana jako element prawego członu.

(39) *Geçit töreni yok ve hiç olmadı.*

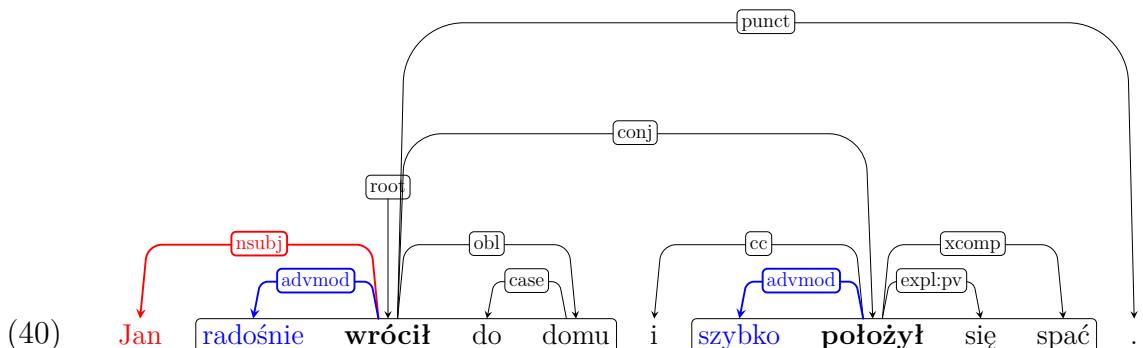


„Nie ma parady i nigdy nie było.” (Türk i in., 2019)

- (H5) Podrzednik głowy lewego członu po jego lewej stronie nie jest częścią lewego członu, jeśli z tą głową łączy go relacja o *unikalnej* etykiecie. Przez unikalną etykietę rozumiem taką, która nie występuje na żadnej relacji między głowami innych członów tej koordynacji i ich podrzędnikami.

Na potrzeby tej heurystyki etykiety **subj** i **subj:pass** oraz **nummod** i **nummod:gov** uznaję za identyczne.

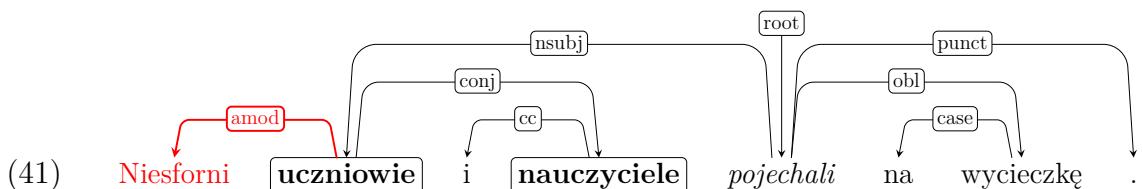
W przykładzie (40) tokeny *szymbko* i *radośnie* są połączone z głowami członów identycznymi etykietami **advmod**. Z tego powodu są potraktowane jako prywatne modyfikatory swoich nadrzędników. Token *Jan* jest podrzędnikiem relacji o unikalnej etykiecie **nsubj**, więc jest uznany za element obu członów.



Celem reguły (H5) jest rozróżnienie elementów „prywatnych” dla danego członu, takich jak okoliczniki *radośnie* i *szymbko* w (40), od elementów „wspólnych” dla obu członów, takich jak podmiot *Jan* w (40).

Niemniej jednak rozróżnienie to nie zawsze działa prawidłowo. Przykładowo, zastosowanie (H5) do powtórnego poniżej zdania (8) powoduje, że *Niesforni* zostaje uznane za element wspólny i prowadzi do interpretacji innej od tej, która wynika z semantyki.

- (8) Niesforni uczniowie i nauczyciele pojechali na wycieczkę.



Zasady (H1) i (H2) niemalże zawsze działają prawidłowo. Heurystyki (H3) i (H4) są bardziej zawodne. Najczęściej błędna jest reguła (H5).

4.2.4. Określanie pozycji nadziednika

Na potrzeby przetwarzania korpusów UD (korzystających z podejścia stanfordzkiego) pozycja nadziednika koordynacji określana jest w następujący sposób:

ozn.	pozycja nadziednika	definicja (zakladajaca opis w stylu UD)	przyklad koordynacji
(L)	po lewej stronie	przed poczatkem pierwszego czlonu	Drzewo <i>sadz</i> [[Pat] i [Mat]].
(0)	brak nadziednika	glowa lewego czlonu jest korzeniem zdania	[[Posadzili] i [podlali]] drzewo.
(R)	po prawej stronie	po koncu ostatniego czlonu	[[Pat] i [Mat]] <i>posadzili</i> drzewo.
(M)	po srodku	pozostałe przypadki	[[Pat] – powtórzyłem – oraz [Mat]].

Ze względu na małą liczbę koordynacji (M) oraz częste błędy w drzewach zdań zawierających koordynacje z nadziednikiem po środku w badaniu nie przeprowadzam osobnych analiz dla koordynacji (M).

4.2.5. Obliczanie długości członu

Drugą ważną z perspektywy analizy statystycznej cechą konstrukcji współrzędnie złożonej jest różnica długości jej członów.

Długość członu określana jest na cztery sposoby: jako liczba **słów**, **tokenów**, **sylab** i **znaków**.

Słowa Właściwym podejściem jest traktowanie słów jako podciągów tekstu rozdzielonych spacjami. Przyjęcie takiego rozwiązania skutkuje brakiem możliwości obliczenia długości członu, gdy jedynie część danego słowa należy do tego członu. Pokazuje to poniższy przykład:

- (42) Arma virumque canō.¹²

Arma vir-um que canō .
broń.ACC.PL mąż-ACC.SG i śpiewać.PRS.IND.ACT.1SG

„Śpiewam o broni i mężu.”

W zdaniu (42) znajduje się koordynacja binarna (43) ze spójnikiem *que*:

- (43) [[Arma] [virum]que]

Uznanie *virumque* za jedno słowo oznaczałoby, że prawy człon koordynacji (43) składa się z niecałkowitej liczby słów. Nie wiadomo, jak wielu koordynacji może dotyczyć ten problem, jednak nawet jeśli wyżej opisane zjawisko jest rzadkie, powoduje ono dojście do niedopuszczalnych wniosków. W celu uniknięcia opisanego powyżej problemu, przez liczbę słów rozumiem liczbę wszystkich tokenów oprócz tych będących znakami interpunkcyjnymi.

¹²Publius Vergilius Maro, *Eneida*, <https://www.thelatinlibrary.com/vergil/aen1.shtml>. W analizie morfologicznej użyto narzędzia Collatinus web (<https://utils.biblissima.fr/en/collatinus-web/>).

Tokeny Tokenami są wyrazy i znaki interpunkcyjne, a także klityki, części kontrakcji i złożień (Riedl i Biemann, 2018). W Universal Dependencies reguły tokenizacji różnią się nieznacznie między językami. Interpunkcja oraz części złożień są odrębnymi tokenami, chyba że stanowią „integralną część lematu” danego słowa (De Marneffe i in., 2021).

W zdaniu (44) *śmy* jest klityką dołączoną do tokenu *Wywiesili*, zaś przymiotnik złożony *biało-czerwoną* składa się z tokenów *biało* i *czerwoną* oraz z łącznika (który również jest osobnym tokenem).

W zdaniu (45) *ins* jest kontrakcją dwóch tokenów: *in* oraz *das*. Słowo *Krankenhaus* (szpital), pomimo że jest złożeniem słów *Kranken* (chorzy) oraz *Haus* (dom), jest potraktowane jako jeden token.¹³

(44) Wywiesiliśmy biało-czerwoną flagę.

Wywiesili śmy biało - czerwoną flagę .

(45) Ich gehe **ins** Krankenhaus.

Ich gehe **in das** Krankenhaus .
ja iść.1SG.PRS do DET.SG.N.ACC szpital

„Idę do szpitala.”

Sylaby Liczba sylab w obrębie danego członu jest sumą liczb sylab w poszczególnych słowach wchodzących w skład tego członu. Na potrzebę liczenia sylab przez „słowo” rozumiem część zdania oddzieloną od reszty spacją. Liczba sylab w tokenach została określona na podstawie opisanej niżej procedury.

- Algorytm sprawdza, czy słowo znajduje się na liście skrótów występujących w danym języku. Jeśli tak, rozpatruje rozwinięcie tego skrótu.
- Następnie algorytm sprawdza, czy słowo lub jego część jest liczbą. Jeśli tak, zamienia ją na formę słowną, wykorzystując do tego bibliotekę `numpy`¹⁴.
- Ostatecznie algorytm dzieli słowo na sylaby, wykorzystując następujące biblioteki języka Python:

`loguax` dla łaciny,

`turkishNLP` dla języka tureckiego,

`pyphen` dla pozostałych języków.

¹³Istnieją argumenty za rozdzieleniem niemieckich czasowników złożonych na tokeny (Riedl i Biemann, 2018), jednak autorzy standardu UD podjęli decyzję, żeby tego nie robić (De Marneffe i in., 2021).

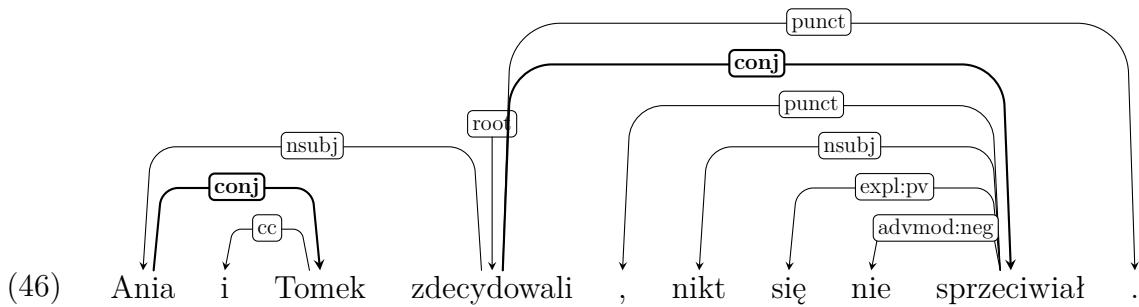
¹⁴Biblioteka ta nie obsługuje języka łacińskiego. Niemniej jednak prawie wszystkie (>99.5%) liczebniki w obrębie korpusów łaciny zapisane są słownie.

Znaki Długość członu wyrażona w znakach, tj. literach, spacjach i znakach interpunkcyjnych.

W przypadku języka koreańskiego każdy znak alfabetu (*jamo*) traktowany jest jak litera, zaś każdy blok znaków jako sylaba (Simpson i Kang, 2004). Przykładowo, słowo 꿀별 (pszczola) składa się z dwóch sylab (odpowiadających blokom 꿀 i 별) oraz z sześciu znaków (᥃, ᄊ, ᄂ, ᄅ i ᄃ) ¹⁵.

4.2.6. Koordynacje zagnieżdżone

Czasami zdarzają się sytuacje, w których jedna koordynacja jest częścią innej. Takie koordynacje nazywają się zagnieżdżonymi.

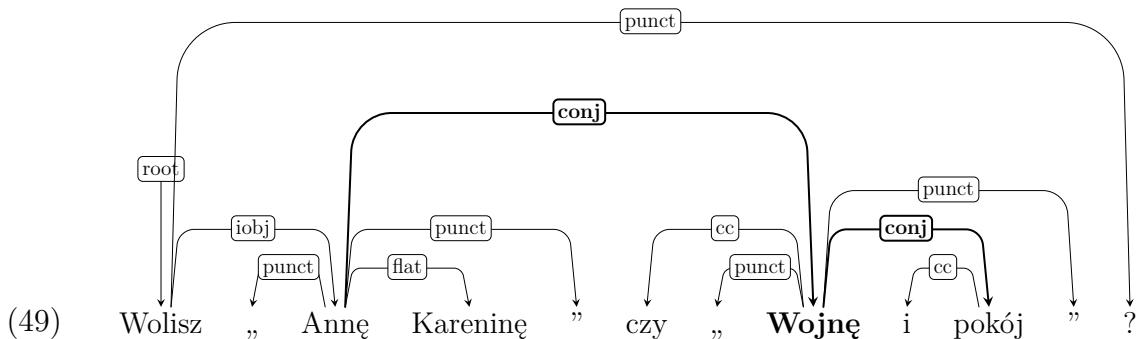


Należy traktować je jako dwie osobne konstrukcje współrzędnie złożone:

(47) [[Ania] i [Tomek]]

(48) [[Ania i Tomek zdecydowali], [nikt się nie sprzeciwiał]]

Jeden token może być głową lewego członu w jednej koordynacji i głową prawego w drugiej:



Nie stanowi to przeszkody dla wyciągania koordynacji według wyżej opisanych zasad. Algorytm (33) poprawnie wykrywa obie koordynacje:

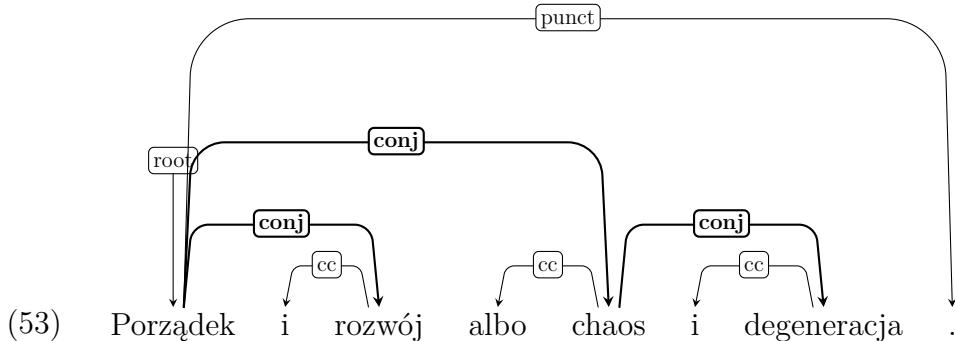
(50) [[Wojnę] i [pokój]]

(51) [[„Anne Kareninę”] czy [„Wojnę i pokój”]]

¹⁵Przykład pochodzi ze strony https://www.korean.go.kr/eng_hangeul/principle/001.html.

Podejście używane przez UD dopuszcza sytuacje, w których jeden token jest głową lewego członu dwóch różnych koordynacji. W takich sytuacjach rozdzielenie koordynacji zagnieżdżonych jest trudniejsze. Pokazuje to przykład zdania (52) i jego drzewa (53).

(52) Porządek i rozwój albo chaos i degeneracja.



Według wcześniejszej opisanej procedury wyciągania koordynacji, zdanie (52) powinno zawierać następujące koordynacje:

(54) *[[Porządek] i [rozwój] albo [chaos i degeneracja]]

(55) [[chaos] i [degeneracja]]

Opis ten jest niepoprawny. W rzeczywistości w zdaniu (52) występują następujące koordynacje:

(56) [[Porządek i rozwój] albo [chaos i degeneracja]]

(57) [[Porządek] i [rozwój]]

(58) [[chaos] i [degeneracja]]

Ponieważ wyżej opisane reguły znajdowania głów członów koordynacji nie zadziałały poprawnie, należy zmodyfikować algorytm.

4.2.7. Procedura znajdowania koordynacji z uwzględnieniem koordynacji zagnieżdżonych

W celu poprawnej analizy zdań zawierających koordynacje zagnieżdżone w procedurze znajdowania koordynacji przyjmuję następującą regułę:

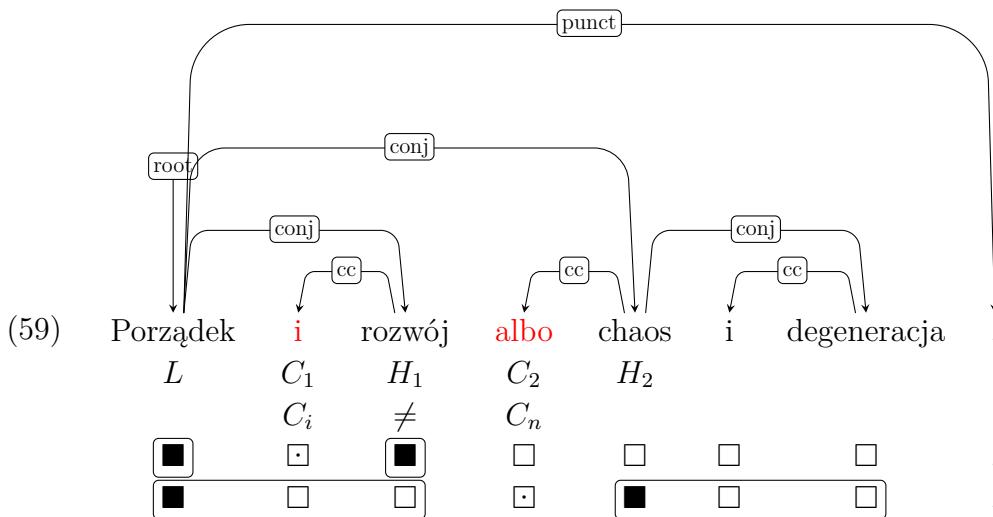
(H6) Jeśli występuje sytuacja opisana w (33) oraz:

- istnieją przynajmniej dwa spójniki i
- jeden ze spójników C_i różni się od ostatniego spójnika C_n ,

to należy rozpatrzyć dwie koordynacje:

- taką, w której nadziednikiem jest G , spójnikiem C_i , a głowami członów są L oraz H_1, \dots, H_i (przy czym głową lewego członu jest L , a głową prawego członu H_i);
- taką, w której nadziednikiem jest G , spójnikiem C_n , a głowami członów są L oraz H_{i+1}, \dots, H_n (przy czym głową lewego członu jest L , a głową prawego członu H_n).

Schemat (59) obrazuje zastosowanie (H6) do drzewa (53) zdania (52):



W pierwszej kolejności algorytm (33) rozpatruje token *Porządek*. Ponieważ wśród jego podrzędników znajdują się relacje `conj`, zostaje wykryta przynajmniej jedna koordynacja. Ponieważ spójniki *i* oraz *albo* są różne, zgodnie z (H6) algorytm rozpatruje dwie koordynacje.

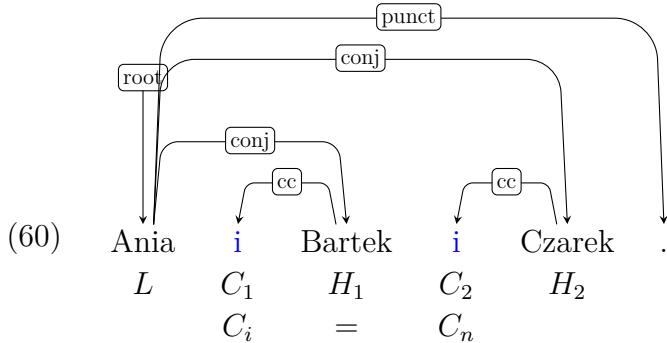
Pierwsza koordynacja nie ma nadziednika, głową jej lewego członu jest token *Porządek*, głową prawego członu *rozwój*, zaś spójnikiem token *i*. Jest to koordynacja (57), czyli `[[Porządek] i [rozwój]]`.

Druga koordynacja również nie posiada nadziednika. Głową jej lewego członu jest token *Porządek*, głową prawego członu *chaos*, zaś spójnikiem token *albo*. Zgodnie z (H6) token *rozwój* nie jest traktowany jako głowa członu w tej koordynacji. Dzięki temu po zastosowaniu reguł (H1)–(H5) do wykrytej koordynacji zostaje ona poprawnie opisana jako `[[Porządek i rozwój] albo [chaos i degeneracja]]`. W ten sposób procedura opisała koordynację (56).

Następnie algorytm rozpatruje pozostałe tokeny w poszukiwaniu krawędzi `conj`. Token *chaos* posiada podrzędnik *degeneracja* z taką relacją. Zgodnie z procedurą (33) znaleziona zostaje koordynacja, której nadziednikiem jest *Porządek*, głową lewego członu *chaos*, głową prawego członu *degeneracja*, a spójnikiem *i*. W ten sposób algorytm opisuje koordynację (58), czyli `[[chaos] i [degeneracja]]`. W tej koordynacji występuje tylko jeden spójnik, więc reguła (H6) nie jest stosowana.

W zdaniu nie występuje więcej tokenów posiadających podrzędniki `conj`, w związku z czym algorytm kończy rozpatrywanie zdania i zwraca prawidłowo trzy koordynacje: (56), (57) i (58).

W przykładzie (60) występuje podobna sytuacja, co w (53). Tym razem jednak wszystkie spójniki podrzędników tokenu *Ania* są identyczne:



W związku z tym warunki (H6) nie są spełnione. Algorytm (33) rozpatruje zdanie (60) jako zwykłą koordynację wieloczlonową:

(61) [[Ania] i [Bartek] i [Czarek]]

4.3. Weryfikacja działania algorytmu

4.3.1. Ograniczenia

Wyżej opisana procedura wyciągania koordynacji nie jest niezawodna. Jest wiele powodów, dla których koordynacje mogą być źle opisane:

- nieprawidłowe dane w obrębie korpusu (np. ciągi losowych znaków, które nie są częścią języka naturalnego),
- nieprawidłowe drzewa zależnościowe (błędy w automatycznym opisie lub konwersji znakowania, błędy w ręcznym opisywaniu, błędy wynikające z niedoskonałości standardu UD),
- nieprawidłowe działanie heurystyk.

Z tych przyczyn algorytm został poddany ewaluacji. Opisana niżej procedura została opracowana na podstawie ewaluacji używanej w badaniu Przepiórkowski i Woźniak (2023) oraz replikującej je analizie Przepiórkowski i in. (2024).

4.3.2. Dobór języków

Ewaluacji zostały poddane zdania w następujących językach:

- język polski – dwóch natywnych recenzentów,
- język angielski – dwóch recenzentów,
- język turecki – jeden natywny recenzent.

4.3.3. Losowanie wyciągniętych koordynacji

W przypadku języka polskiego i angielskiego wylosowano 300 koordynacji z uwzględnieniem pozycji nadziednika – po 100 z nadziednikiem po lewej, po prawej i bez nadziednika. W przypadku języka tureckiego wylosowano 60 koordynacji bez rozróżnienia na pozycję nadziednika.

4.3.4. Ocena poprawności

Dwóch recenzentów niezależnie ocenia poprawność wyciągania koordynacji na podstawie dwóch kryteriów:

- Lewy i prawy człon koordynacji są **dokładnie**¹⁶ takie, jakie powinny być.
- Pozycja nadziednika jest poprawnie określona.

Następnie recenzenci spotykają się i rozstrzygają konflikty. Miarą oceny algorytmu jest stosunek liczby poprawnie opisanych koordynacji do liczby wszystkich koordynacji danego typu.

4.3.5. Wyniki

Tabela 4.2 przedstawia odsetek koordynacji ocenionych jako poprawnie wyciągnięte w językach poddanych ewaluacji.

Język	Wszystkie koordynacje	Nadziednik po lewej	Nadziednik po prawej	Brak nadziednika
polski	0,79	0,83	0,89	0,66
angielski	0,72	0,72	0,61	0,84
turecki	0,58			

Tabela 4.2: Poprawność ewaluacji

Przed rozstrzygnięciem konfliktów współczynnik zgodności recenzentów κ wyniósł 56% dla języka polskiego i 54% dla języka angielskiego.

¹⁶Pomijano jedynie błędy dotyczące znaków interpunkcyjnych, ponieważ nie mają one wpływu na długość członów w słowach.

W korpusach języka tureckiego znajdowały się liczne zdania złożone pozbawione spójnika lub właściwej interpunkcji. Zostały one uznane za nieprawidłowe koordynacje. Jeśli przyjąć założenie, że takie zdania tworzą poprawne koordynacje zdaniowe, wynik ewaluacji dla języka tureckiego wynosi 73%.

Zarówno współczynnik zgodności recenzentów, jak i odsetek poprawnych koordynacji algorytmu mają niższe wartości niż analogiczne miary uzyskane w pracy Przepiórkowski i Woźniak (2023). Należy jednak zwrócić uwagę, że w ich badaniu ewaluacja algorytmu polegała na sprawdzeniu wyłącznie poprawności pozycji nadziednika.

Przepiórkowski i in. (2024) ewaluując swój algorytm sprawdzali zarówno pozycję nadziednika, jak i granice członów. Stosując tę metodę uzyskali dokładność równą 50,1%. Niemniej jednak nie można porównywać tego wyniku do uzyskanych w moim badaniu wskaźników z powodu dwóch istotnych różnic metodologicznych. Po pierwsze, Przepiórkowski i in. (2024) korzystali z danych gorszej jakości, ponieważ analizowali drzewa zależnościowe uzyskane za pomocą automatycznego parsowania zdań. Po drugie, stosowali oni bardziej rygorystyczną procedurę ewaluacji, losując więcej koordynacji z dłuższą różnicą długości członów, niż wynikało to z proporcji występujących w języku.

Podsumowując, wyniki ewaluacji wskazują na to, że opisana w niniejszym rozdziale procedura wyciągania koordynacji działa dobrze. Nie można jednak porównywać tych wyników do wyników ewaluacji w innych badaniach, ponieważ poprzednie analizy w tym zakresie stosowały inną metodologię.

Rozdział 5

Metody statystyczne

5.1. Względna pozycja głównego członu

Język	lewym człon				prawym człon			
	N	średnia	t	p	N	średnia	t	p
Języki inicjalne								
angielski	12326	0,36	-37	1.74e-288	15155	0,40	-31	1.99e-199
czeski	51416	0,34	-88		62872	0,39	-70	0
hiszpański	19685	0,30	-74		22137	0,31	-80	0
islandzki	31929	0,18	-196		36967	0,18	-225	0
polski	9976	0,23	-74		12049	0,33	-50	0
portugalski	19732	0,30	-71		22349	0,30	-84	0
rosyjski	36608	0,38	-56		46050	0,41	-47	0
rumuński	27224	0,37	-53		31024	0,37	-60	0
włoski	17728	0,43	-22	1,2e-107	20330	0,37	-49	0
Języki mieszane								
łaciński	19766	0,36	-47		25755	0,48	-5,8	7,98e-09
niemiecki	51068	0,53	17	9,27e-62	64616	0,52	14	2,61e-46
Języki finalne								
koreański	6801	0,78	59		12951	0,65	47	0
turecki	7994	0,64	28	7,97e-167	12763	0,69	55	0

Tabela 5.1: Względna pozycja głów członów koordynacji

Tabela 5.1 przedstawia rozkład względnej pozycji głównego członów w obrębie prawych i lewych członów w różnych językach. Pod uwagębrane są człony, których głowa nie jest

jedynym elementem (tj. takie, które mają przynajmniej dwa tokeny). Pozycja główny członu ustalana jest na podstawie następującego wzoru:

$$P = \frac{H - 1}{N - 1} \text{ dla } N \geq 2, \text{ gdzie:}$$

- P oznacza względną pozycję główny członu w członie;
- H oznacza pozycję główny członu w członie;
- N oznacza długość członu w tokenach.

W przykładzie (62) głowa lewego członu *Urzqd* znajduje się na samym początku członu, czyli w pozycji $P = \frac{1-1}{2-1} = 0$, zaś głowa prawego członu *Izba* na jego środku, czyli w pozycji $P = \frac{2-1}{3-1} = 0,5$.

(62)	Urzad Skarbowy	i	Krajowa Izba Rozliczeniowa
	1	2	1 2 3

W językach inicjalnych głowa członu występuje znacznie częściej na początku członu ($< 0,5$), zaś w językach inicjalnych częściej na jego końcu ($> 0,5$). Dotyczy to zarówno lewych, jak i prawych członów koordynacji.

W przypadku języków mieszanych nie można zaobserwować wyraźnej prawidłowości. W łacinie głowy lewych członów znajdują się częściej bliżej początku członu, zaś głowy prawych członów bliżej środka członu. W języku niemieckim głowy członów są w okolicach środka członu. Z tego powodu w dyskusji wyników nie interpretuję wyników dla języków mieszanych w kontekście przewidywań dotyczących struktury zależnościowej koordynacji.

Istotność statystyczna uzyskanych wyników jest potwierdzona testem t-Studenta sprawdzającym różnicę pozycji główny od 0,5, czyli środka członu¹⁷.

5.2. Pozycja nadziednika

Tabela 5.2 pokazuje rozkład występowania koordynacji o różnych pozycjach nadziednika w zależności od języka¹⁸.

Widoczne są trzy grupy językowe, pokrywające się z podziałem zaproponowanym w pracy Polinsky (2012).

W językach inicjalnych nadziednik występuje znacznie częściej po lewej stronie (ok. 60% koordynacji) niż po prawej (ok. 10% przypadków).

¹⁷Do analizy statystycznej użyto funkcji `t.test` języka R. Ilekroć w niniejszej pracy występuje $p = 0$, przez 0 należy rozumieć liczbę dodatnią mniejszą od $Mashine\$double.xmin \approx 5e - 324$ (R Core Team, 2023).

¹⁸Do wszystkich koordynacji należą również te z nadziednikiem po środku (M).

Język	Wszystkie		Nadzędnik po				Brak	
	koordynacje	N	lewej	P	prawej	N	P	nadrzędnika
Języki inicjalne								
angielski	21013	11171	0,53	2972	0,14	6829	0,32	
czeski	90566	49341	0,54	14279	0,16	26688	0,29	
hiszpański	28666	19557	0,68	2751	0,10	6300	0,22	
polski	16684	8407	0,50	2219	0,13	6023	0,36	
portugalski	29255	17661	0,60	3157	0,11	8364	0,29	
rosyjski	61004	31679	0,52	8485	0,14	20556	0,34	
rumuński	37247	21873	0,59	3088	0,08	11993	0,32	
włoski	25426	17014	0,67	2345	0,09	5992	0,24	
islandzki	43852	16986	0,39	2928	0,07	23877	0,54	
Języki mieszane								
łacina	39510	19635	0,50	9264	0,23	9112	0,23	
niemiecki	92115	43089	0,47	23637	0,26	24029	0,26	
Języki finalne								
koreański	21506	718	0,03	14289	0,66	6491	0,30	
turecki	19598	1760	0,09	11936	0,61	5758	0,29	

Tabela 5.2: Pozycja nadzędnika

W językach mieszanych (niemiecki i łacina) nadzędnik również występuje najczęściej po lewej stronie (ok. 50% koordynacji). Jednak koordynacje z nadzędziennikiem po prawej stronie są o wiele częstsze niż w przypadku języków inicjalnych (występują w ok. 25% przypadków).

Natomiast w przypadku języków finalnych (koreańskiego i tureckiego) występują tendencje przeciwnie do tych zaobserwowanych w językach inicjalnych. Nadzędnik występuje najczęściej po prawej stronie (ok. 60% koordynacji) i najrzadziej po lewej stronie (poniżej 10% przypadków).

W prawie wszystkich językach konstrukcje współrzędnie złożone pozabawione nadzędnika stanowią 22–39% koordynacji. Wyjątkiem jest język islandzki, w przypadku którego koordynacje bez nadzędnika są najczęstsze (54%). Jednak gdy w koordynacji występuje nadzędnik, pojawia się on zdecydowanie częściej po lewej stronie (39%), niż po prawej (7%).

Istotność statystyczna różnic częstości występowania koordynacji o różnych pozycjach nadzędnika została potwierdzona przy użyciu testu Wilcoxona. Wielkość statystyk testowych i ich istotność statystyczną pokazuje Tabela 5.2¹⁹.

¹⁹Do analizy statystycznej użyto funkcji `wilcox.test` języka R.

Język	Nadrzędnik po lewej – po prawej		Nadrzędnik po lewej – brak		Nadrzędnik po prawej – brak	
	V	p	V	p	V	p
Języki inicjalne						
angielski	3,07e+08	0	2,66e+08	0	2,61e+08	0
czeski	5,69e+09	0	5,13e+09	0	4,66e+09	0
hiszpański	6,52e+08	0	6,01e+08	0	4,62e+08	0
polski	1,91e+08	0	1,59e+08	6,18e-153	1,71e+08	0
portugalski	6,40e+08	0	5,64e+08	0	5,04e+08	0
rosyjski	2,57e+09	0	2,20e+09	0	2,23e+09	0
rumuński	1,04e+09	0	8,78e+08	0	8,60e+08	0
włoski	5,10e+08	0	4,63e+08	0	3,70e+08	0
islandzki	1,27e+09	0	8,10e+08	0	1,42e+09	0
Języki mieszane						
niemiecki	5,14e+09	0	5,12e+09	0	4,26e+09	0,037
łacina	9,85e+08	0	9,88e+08	0	7,78e+08	0,201
Języki finalne						
koreański	8,53e+07	0	1,69e+08	0	1,47e+08	0
turecki	9,23e+07	0	1,53e+08	0	1,32e+08	0

Tabela 5.3: Różnice między częstościami występowania różnych pozycji nadrzędnika

	medianą	średnią		medianą	średnią							
	lewy	prawy	lewy	prawy	V	p	lewy	prawy	lewy	prawy	V	p
Język angielski												
Wszystkie koordynacje (N = 21 013)												
znaki	13	20	22,79	32,69	5,3e+07	0	znaki	13	21	24,44	34,64	1e+09
sylaby	4	5	6,00	8,51	4e+07	0	sylaby	5	8	8,93	12,43	8,7e+08
słowa	2	4	4,19	5,96	2,3e+07	0	słowa	2	3	3,70	5,31	3,8e+08
Brak nadrzędnika (N = 6 829)												
znaki	30	40	37,41	50,99	6,9e+06	7,4e-154	znaki	31	45	39,54	57,18	9,9e+07
sylaby	8	10	9,61	13,00	6e+06	5,7e-148	sylaby	11	16	14,16	20,14	9,3e+07
słowa	6	8	7,13	9,65	5,1e+06	3,6e-159	słowa	5	7	6,13	9,08	7,1e+07
Nadrzędnik po lewej (N = 11 171)												
znaki	10	16	17,28	26,49	1,3e+07	0	znaki	10	16	19,59	27,69	2,8e+08
sylaby	3	4	4,64	6,99	9,5e+06	0	sylaby	4	6	7,28	10,10	2,4e+08
słowa	2	3	3,05	4,66	4,4e+06	0	słowa	1	2	2,93	4,14	8,7e+07
Nadrzędnik po prawej (N = 2 972)												
znaki	7	9	10,10	13,90	8,1e+05	6,5e-99	znaki	8	10	12,83	16,28	2,2e+07
sylaby	2	2	2,88	3,88	4,9e+05	5,2e-75	sylaby	3	4	4,81	6,02	1,6e+07
słowa	1	1	1,74	2,34	8e+04	1,8e-68	słowa	1	1	1,82	2,29	2,2e+06
Nadrzędnik po lewej (N = 49 341)												
znaki	10	16	19,59	27,69	2,8e+08	0	znaki	10	16	19,59	27,69	2,8e+08
sylaby	4	6	7,28	10,10	2,4e+08	0	sylaby	4	6	7,28	10,10	2,4e+08
słowa	1	2	2,93	4,14	8,7e+07	0	słowa	1	2	2,93	4,14	8,7e+07
Nadrzędnik po prawej (N = 14 279)												
znaki	8	10	12,83	16,28	2,2e+07	0	znaki	8	10	12,83	16,28	2,2e+07
sylaby	3	4	4,81	6,02	1,6e+07	3,3e-291	sylaby	3	4	4,81	6,02	1,6e+07
słowa	1	1	1,82	2,29	2,2e+06	5,1e-202	słowa	1	1	1,82	2,29	2,2e+06

Tabela 5.4: Długość członów koordynacji – języki inicjalne

	medianą	średnia		medianą	średnia	
	lewy	prawy	lewy	prawy	V	p
Język niemiecki						
Wszystkie koordynacje (N = 92 115)						
znaki	15	26	26,71	37,86	1,1e+09	0
sylaby	5	9	8,72	12,12	9,5e+08	0
słowa	2	3	3,64	5,09	4,3e+08	0
Brak nadzędnika (N = 24 029)						
znaki	46	56	51,35	64,93	9,4e+07	0
sylaby	14	17	16,03	20,02	9e+07	0
słowa	7	8	7,26	9,21	7,3e+07	0
Nadzędnik po lewej (N = 43 089)						
znaki	12	20	19,35	30,61	2e+08	0
sylaby	4	7	6,56	10,03	1,8e+08	0
słowa	1	2	2,55	3,98	6,1e+07	0
Nadzędnik po prawej (N = 23 637)						
znaki	11	16	15,29	22,72	5,9e+07	0
sylaby	4	6	5,27	7,64	5e+07	0
słowa	1	2	1,95	2,80	1,2e+07	0
Język koreański						
Wszystkie koordynacje (N = 21 506)						
znaki	10	15	15,99	24,44	4,7e+07	0
sylaby	4	6	6,35	9,08	4e+07	0
słowa	1	2	1,94	2,88	1e+07	0
Brak nadzędnika (N = 6 491)						
znaki	14	35	24,13	41,71	3,8e+06	0
sylaby	5	13	9,35	15,12	3,9e+06	0
słowa	2	4	2,86	4,86	2,4e+06	0
Nadzędnik po lewej (N = 718)						
znaki	10	15	13,31	20,92	4,8e+04	2,5e-38
sylaby	4	6	5,35	7,68	4,7e+04	3,4e-27
słowa	1	2	1,66	2,47	1,4e+04	2,3e-27
Nadzędnik po prawej (N = 14 289)						
znaki	9	11	12,42	16,77	2,3e+07	0
sylaby	4	5	5,04	6,40	1,7e+07	8e-292
słowa	1	1	1,53	2,01	2,3e+06	0

Tabela 5.5: Długość członów koordynacji – języki mieszane i finalne

5.3. Długość członów koordynacji

Tabele 5.4 i 5.5 przedstawiają średnią oraz medianę długości prawych i lewych członów koordynacji w zależności od pozycji nadzędnika w przykładowych językach.

Wybrano dwa języki inicjalne (czeski i angielski) oraz po jednym języku mieszczanym (niemiecki) i finalnym (koreański) o największej łącznej objętości korpusów. Tabele przedstawiające długości członów we wszystkich badanych językach oraz statystyki testowe stanowią Dodatek C do niniejszej pracy.

Dane potwierdzają zaobserwowaną w pracy Przepiórkowski i Woźniak (2023) tendencję do umieszczania krótszych członów po lewej stronie konstrukcji współrzędnie złożonych w języku angielskim. To zjawisko pojawia się niezależnie od pozycji nadzędnika.

Tendencja ta jest obecna również w innych językach inicjalnych, a także w językach mieszczanych i finalnych.

5.4. Różnica długości członów a pozycja krótszego członu

W celu replikacji badania z pracy Przepiórkowski i Woźniak (2023) i weryfikacji tendencji przewidywanych przez różne modele struktury składniowej koordynacji należy zbadać zależność między pozycją krótszego członu a różnicą długości członów koordynacji.

Rysunki 5.1–5.5 pokazują tę relację w przykładowych językach. Pozostałe wykresy przedstawiające tę zależność znajdują się w Dodatku D²⁰.

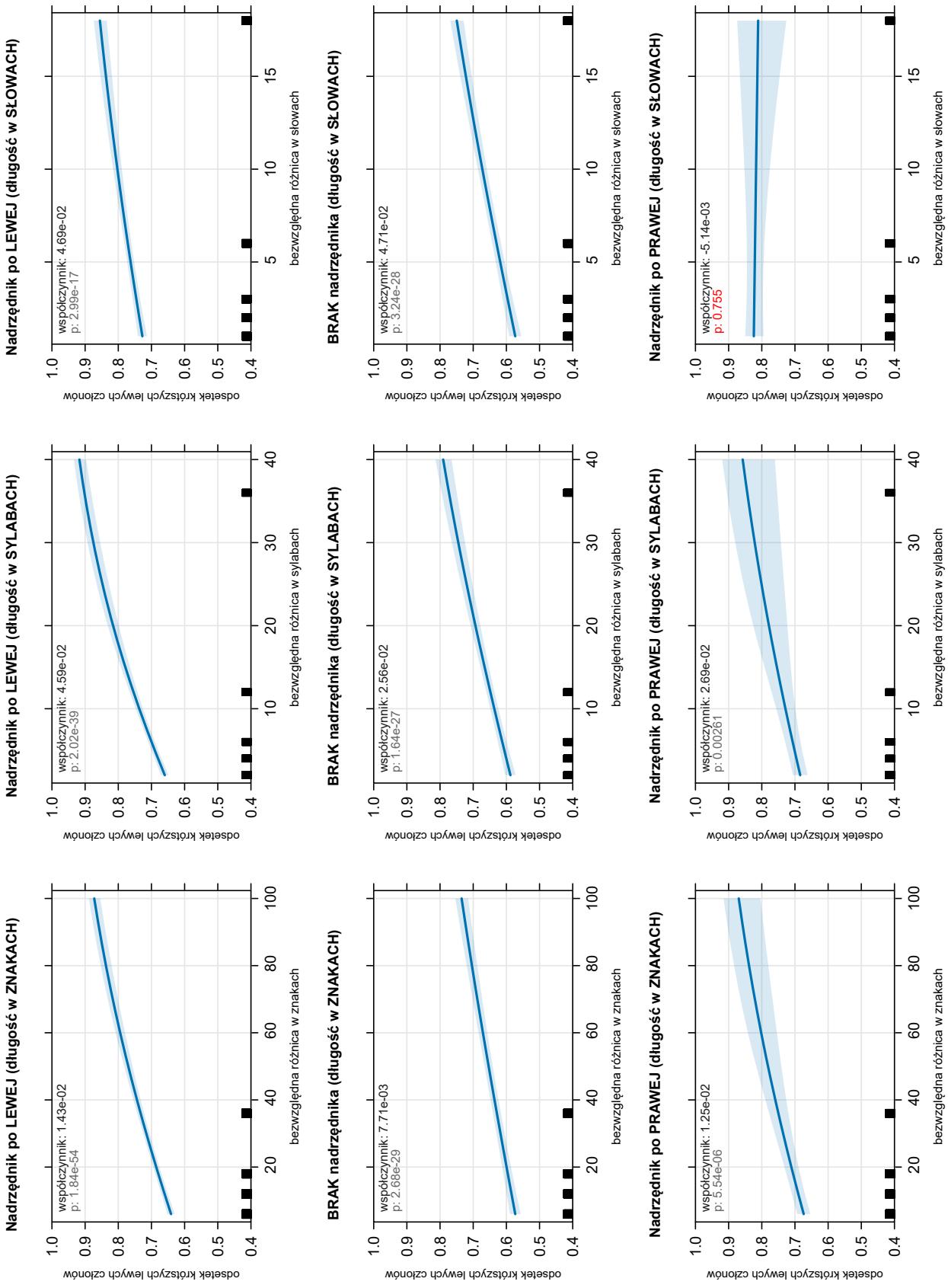
Wykresy przedstawiają modele regresji logistycznej dopasowane do danych pogrupowanych w podzbiory. Podgrupy dla wykresów dotyczących długości mierzonej w słowach zostały określone na podstawie wartości granicznych ze zbioru {1, 2, 3, 6, 18}. Dla pozostałych wykresów wartości te są przeskalowane przez 2 w przypadku sylab i 6 w przypadku znaków. Jest to podejście analogiczne do zastosowanego w pracy Przepiórkowski i Woźniak (2023) ze zmianami wartości wynikającymi z konieczności utworzenia odpowiednio dużych podgrup. Wartości graniczne są zaznaczone na wykresach.

Na wszystkich wykresach widoczna jest pozytywna relacja między wzrostem różnicy długości członów a tendencją do umieszczania krótszego członu po lewej stronie w przypadkach, gdy nadziednik jest po lewej stronie. Ta relacja nie jest zależna od tego, czy język jest inicjalny, mieszany czy finalny. Taką samą zależność można zaobserwować w przypadku koordynacji pozabawionych nadziednika. Wszystkie te relacje są istotne statystycznie, przy czym w większości przypadków istotność jest znaczna ($p < 0,001$).

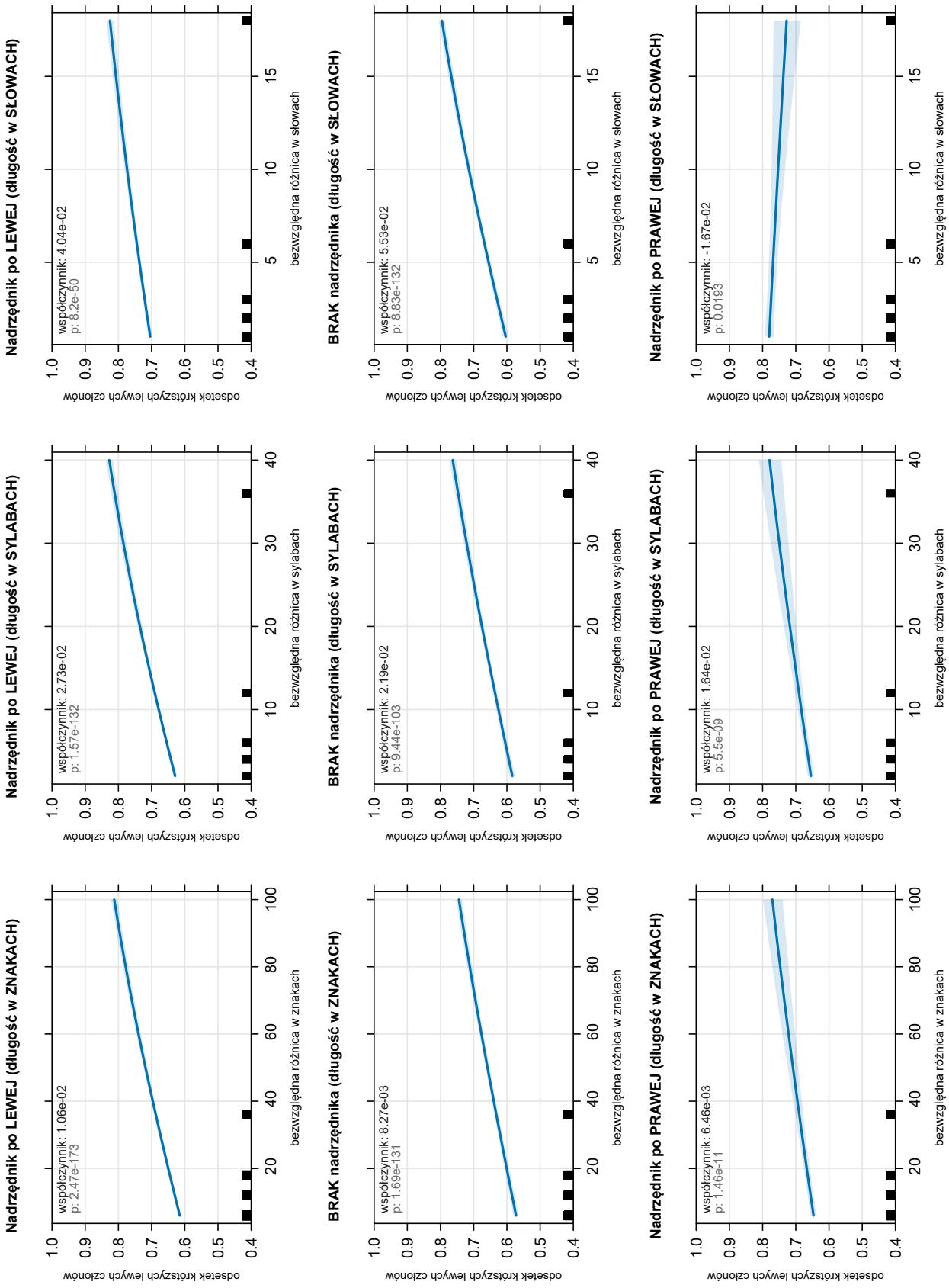
Natomiast w przypadku konstrukcji współrzędnie złożonych z nadziednikiem po prawej stronie zaobserwowane tendencje są różne. W języku niemieckim, koreańskim i tureckim opisywana tendencja jest pozytywna. To samo dotyczy wielu języków inicjalnych (m.in. języków romańskich: włoskiego, portugalskiego i hiszpańskiego). Jednak w przypadku języka angielskiego i łaciny, gdy długość członów liczona jest w słowach, widoczna jest tendencja spadkowa. W inicjalnych językach słowiańskich tendencja jest spadkowa (język czeski) lub nie jest istotna statystycznie (język polski i rosyjski).

W kolejnym rozdziale omawiam dokładnie uzyskane wyniki, kładąc szczególny nacisk na zaobserwowane różnice dotyczące koordynacji z nadziednikiem po prawej stronie.

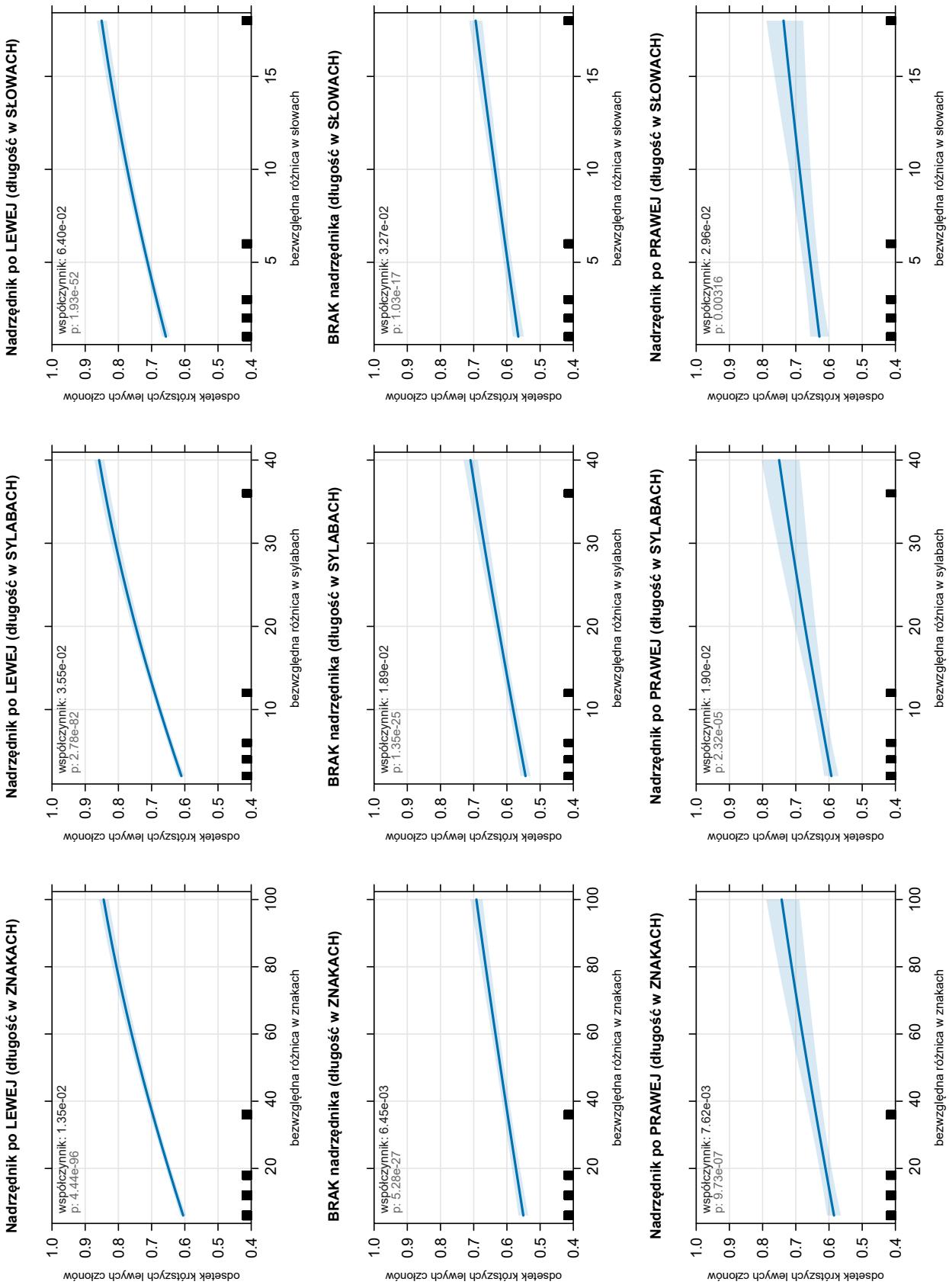
²⁰Wszystkie wykresy w niniejszej pracy zostały utworzone przy pomocy biblioteki Cairo języka R (R Core Team, 2023). W obliczeniach użyto skryptów zastosowanych w pracy Przepiórkowski i Woźniak (2023).



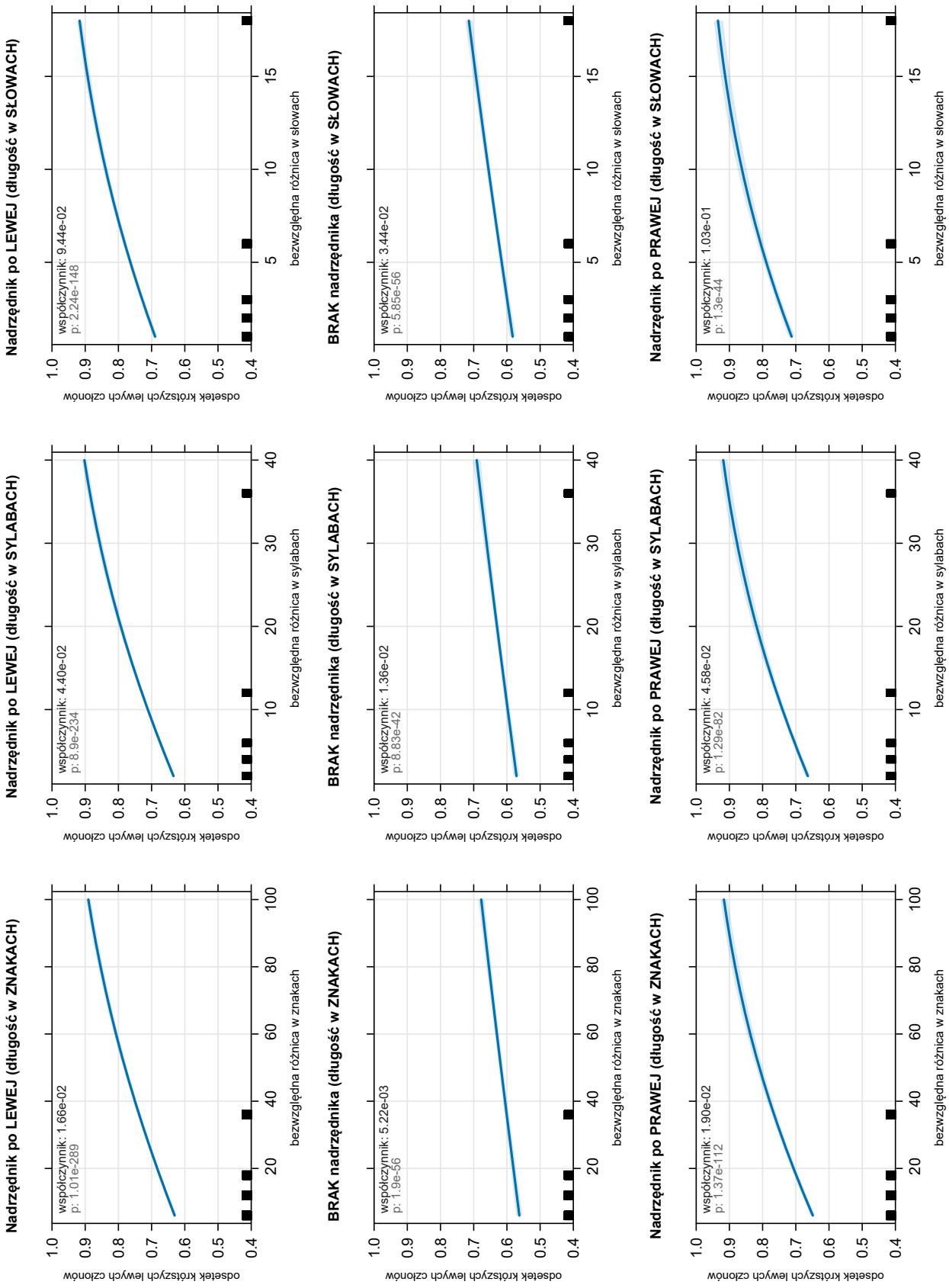
Rysunek 5.1: Różnica długości członów a występowanie krótszego członu po lewej stronie – język angielski (inicjalny, germański)



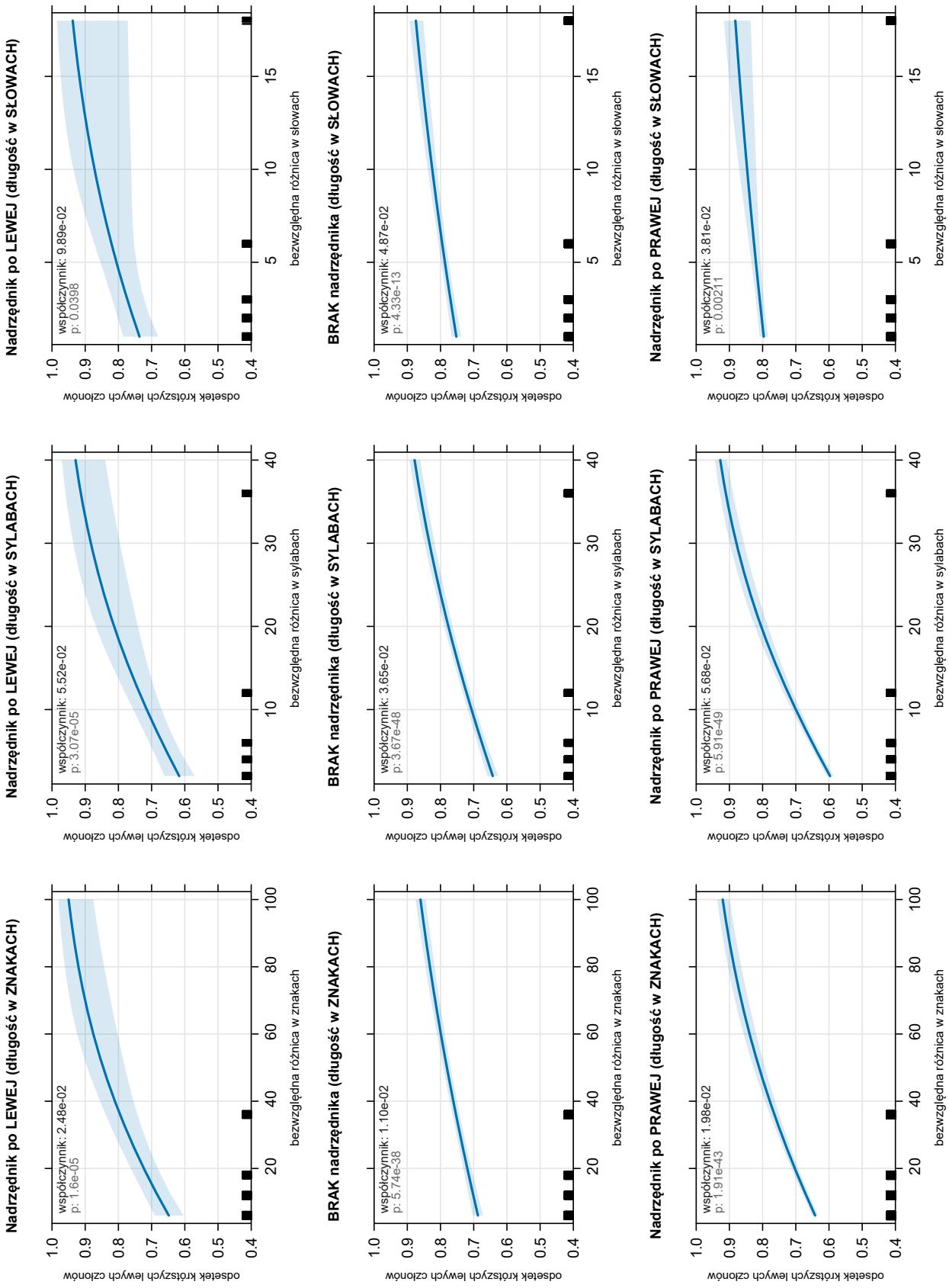
Rysunek 5.2: Różnica długości członów a występowanie krótszego członu po lewej stronie – język czeski (inicjalny, słowiański)



Rysunek 5.3: Różnica długości członów a występowanie krótszego członu po lewej stronie – język portugalski (inicjalny, romski)



Rysunek 5.4: Różnica długości członów a występowanie krótszego członu po lewej stronie – język niemiecki (mieszany)



Rysunek 5.5: Różnica długości członów a występowanie krótszego członu po lewej stronie – język koreański (finalny)

Rozdział 6

Dyskusja wyników

6.1. Replikacja poprzednich badań

6.1.1. Język angielski

Przepiórkowski i Woźniak (2023) opisali dwie zależności dotyczące zmian tendencji do umieszczania krótszego członu koordynacji na jej początku wraz ze wzrostem różnic długości między członami koordynacji w języku angielskim: tendencje pozytywną w koordynacjach z nadziedniakiem po lewej stronie (L) oraz bez nadziednika (0) oraz (nieistotną statystycznie) tendencję negatywną w koordynacjach z nadziedniakiem po prawej stronie (R).

Niniejsza analiza replikuje obserwację pozytywną dotyczącą koordynacji (L) oraz (0). W przypadku koordynacji typu (R) zależność jest pozytywna, gdy długość liczona jest w znakach ($p < 0,001$) i sylabach ($p = 0,003$). Gdy długość członów liczona jest w słowach, tendencja jest negatywna i nieistotna statystycznie ($p = 0,755$).

Te wyniki należy interpretować w kontekście efektu DLM. Polega on na minimalizacji łącznej długości relacji zależnościowych w języku. Długość zależności może być liczona na różne sposoby. Najlepszym z nich jest najprawdopodobniej złożoność syntaktyczna frazy (Lohmann, 2014). Spośród stosowanych przez mnie miar (słowa, sylaby, znaki) najbliższą złożoności syntaktycznej są słowa. W związku z tym uznaję wyniki dotyczące długości członów w słowach za istotniejsze od pozostałych.

Oznacza to, że niniejsza analiza replikuje badanie Przepiórkowski i Woźniak (2023) w zakresie opisu relacji między różnicą długości członów koordynacji a tendencją do umieszczania krótszego członu jako pierwszego w języku angielskim.

6.1.2. Języki słowiańskie

W języku czeskim zależność dotycząca koordynacji (R) jest podobna do tej w języku angielskim. Jest to tendencja spadkowa istotna statystycznie na poziomie $p = 0,019$. Tendencje w języku polskim oraz rosyjskim nie są istotne statystycznie. Może to oznaczać,

że w językach słowiańskich występuje ta sama zależność między pozycją nadrzędnika a zmianą odsetka koordynacji (R-L) względem (R), co w języku angielskim. Znaczy to, że niniejsze badanie rozszerza wyniki pracy Przepiórkowski i Woźniak (2023) na język czeski oraz nie wyklucza, że omawiane tendencje występują w pozostałych językach słowiańskich.

6.1.3. Języki romańskie

Wyniki dla analizowanych języków romańskich pokazują jednoznaczna istotną statystycznie tendencję pozytywną we wszystkich rozpatrywanych przypadkach. Oznacza to, że w językach włoskim, hiszpańskim i portugalskim krótszy człon koordynacji niezależnie od pozycji nadrzędnika znajduje się tym częściej jako pierwszy, im krótszy jest on od ostatniego członu.

W przypadku języka rumuńskiego omawiana tendencja nie jest istotna statystycznie ($p = 0,29$). Warto zauważyć, że język ten pomimo należenia do grupy języków romańskich jest pod silnym wpływem języków słowiańskich.

6.1.4. Języki mieszane

W przypadku języka niemieckiego widoczne są podobne tendencje, jak w przypadku języków romańskich, zaś w łacinie występuje taka zależność, jak w języku angielskim i w językach słowiańskich. Ponieważ metodologia niniejszej pracy nie zakładała przewidywań dotyczących tendencji występujących w tych językach, nie interpretuję tych wyników w kontekście replikacji badania Przepiórkowski i Woźniak (2023).

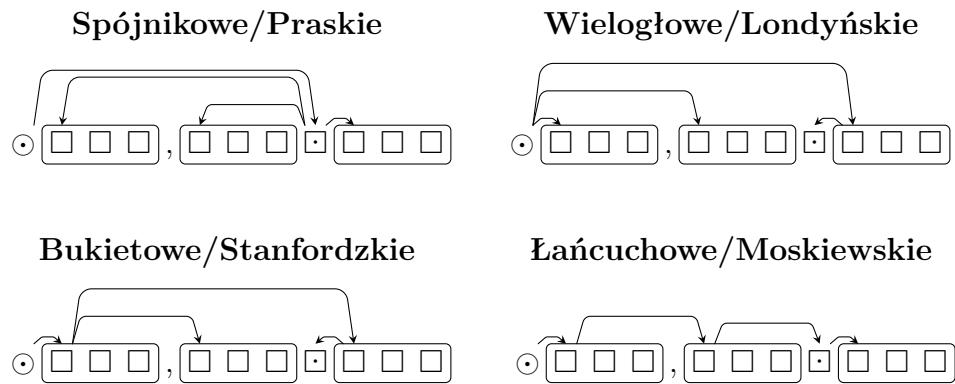
6.1.5. Języki finalne

W języku koreańskim widoczna jest ewidentna tendencja wzrostowa we wszystkich przypadkach. Jest ona znacznie istotna statystycznie ($p < 0,001$) w przypadku koordynacji pozbawionych nadrzędnika, istotna na poziomie $p = 0,002$ dla koordynacji (L) oraz na poziomie $p = 0,04$ dla koordynacji (R). Jest to podobna tendencja do tej zaobserwowanej w językach romańskich i języku niemieckim. W przypadku języka tureckiego tendencja jest wyraźnie słabsza. W przypadku koordynacji (L) nie jest ona istotna statystycznie ($p = 0,29$).

6.2. Przewidywania modeli struktury zależnościowej koordynacji

6.2.1. Języki inicjalne

Jak zostało to omówione w punkcie 3.1.2, istnieją cztery przyjmowane podejścia do struktury zależnościowej koordynacji:



Podejścia praskie i londyńskie nazywane są symetrycznymi, zaś stanfordzkie i moskiewskie – asymetrycznymi. Przepiórkowski i Woźniak (2023) na podstawie przeprowadzonej analizy korpusowej języka angielskiego argumentują, że jedynie podejścia symetryczne (praskie i londyńskie) mogą poprawnie opisywać strukturę zależnościową konstrukcji współrzędnie złożonej. Opierają swoje rozumowanie na predykcjach modeli dotyczących zmiany tendencji do umieszczania krótszego członu koordynacji na jej początku wraz ze wzrostem różnicy długości członów.

Wszystkie cztery podejścia, zgodnie z faktami, przewidują tendencję pozytywną w przypadku koordynacji (L). Dla koordynacji (0) podejście praskie, stanfordzkie i moskiewskie poprawnie przewidują tendencje pozytywne, zaś podejście londyńskie (wbrew faktom) przewiduje brak zależności. Kluczowa jest zależność dotycząca koordynacji (R). Podejścia symetryczne przewidują tendencję spadkową, zaś podejścia asymetryczne tendencję wzrostową.

Ponieważ w języku angielskim omawiana tendencja dla koordynacji (R) jest negatywna, Przepiórkowski i Woźniak (2023) odrzucają podejścia asymetryczne. Bronią jednakże podejścia londyńskiego na podstawie argumentu o gramatykalizacji. Zgodnie z nim fakt, że koordynacje (L) występują w języku angielskim znacznie częściej niż (R) powoduje, że umieszczanie pierwszego członu jako pierwszego mogło stać się regułą.

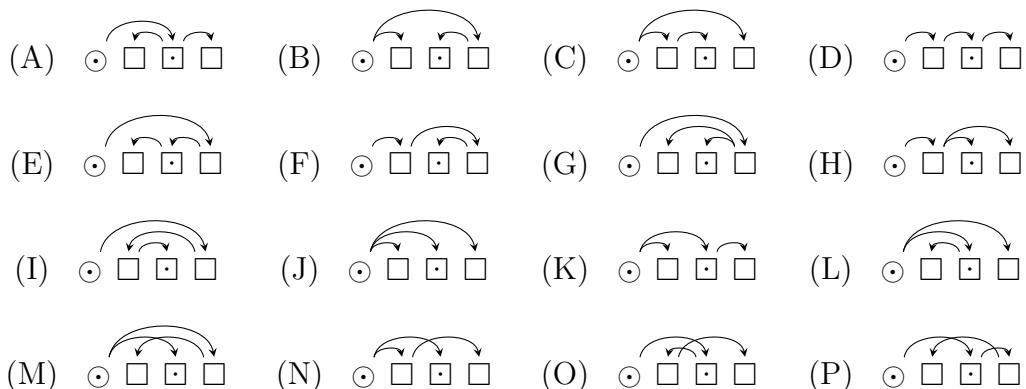
Wyniki przeprowadzonej przeze mnie analizy w zakresie języka angielskiego i języków słowiańskich nie zaprzeczają przywołanej wyżej argumentacji. Niemniej jednak wyniki dotyczące pozostałych języków inicjalnych wskazują na występowanie zupełnie innych tendencji w językach romańskich. Istnieją przynajmniej dwa możliwe wyjaśnienia takiego zjawiska.

Zależności widoczne w językach romańskich są zgodne z przewidywaniami modeli asymetrycznych. Istnieje więc możliwość, że języki romańskie posiadają odrębną strukturę zależnościową koordynacji niż ta występująca w języku angielskim i w językach słowiańskich.

Drugie wyjaśnienie takiego stanu rzeczy jest oparte na argumencie o gramatykalizacji. Możliwe że gramatykalizacja umieszczania krótszego członu po lewej stronie koordynacji jest w przypadku języków romańskich tak silna, że niezależnie od faktycznej struktury zależnościowej koordynacji efekt DLM nie ma większego wpływu na ustalenie członów. Innymi słowy, krótszy człon zawsze występuje częściej z lewej strony tym częściej, im większa jest różnica długości członów niezależnie od pozycji nadziednika.

6.2.2. Języki finalne

Tendencje zaobserwowane w językach finalnych są znacznie trudniejsze do wyjaśnienia na podstawie efektu DLM. Jak zostało to pokazane w punkcie 3.2.2, istnieje 16 możliwych podejść do struktury zależnościowej koordynacji:



Jak pokazuję w rozdziale 3, żadne z nich nie przewiduje zaobserwowanej tendencji pozytywnej dla koordynacji (R). Żadne z podejść nie przewiduje też, żeby tendencja dla koordynacji (L) była bardziej pozytywna, niż tendencja dla koordynacji (R). Oznacza to, że nie istnieje struktura zależnościowa koordynacji, która może tłumaczyć uzyskane wyniki. Jednocześnie nie jest możliwe, żeby koordynacja nie posiadała żadnej struktury zależnościowej.

Uzyskanie niewyjaśnialnych wyników może oznaczać, że metodologia niniejszej pracy jest niewystarczająca w celu poprawnego i jednoznacznego opisu struktury zależnościowej koordynacji.

6.3. Wyjaśnienia wyników

6.3.1. Gramatykalizacja krótszego prawego członu

Z Tabel 5.4 i 5.5 wynika, że we wszystkich badanych językach, niezależnie od pozycji nadziednika krótszy człon koordynacji znajduje się częściej na jej początku. Może to oznaczać, że umieszczanie najkrótszego jako pierwszego uległo gramatykalizacji.

W przypadku języków inicjalnych, przyczyny tej gramatykalizacji są proste w wyjaśnieniu. Każde z czterech omawianych podejść zakłada, że w przypadku koordynacji z nadziedniem po lewej stronie wzrost różnicy długości członów przekłada się na częstsze umieszczanie krótszego członu na początku koordynacji. Ponieważ w językach inicjalnych koordynacje z nadziedniem po lewej stronie są znacznie częstsze od pozostałych (a w wielu przypadkach stanowią ponad połowę konstrukcji współrzędnie złożonych), umieszczanie krótszego członu po lewej stronie stało się w tych językach regułą.

Niemniej jednak zastosowanie analogicznego rozumowania nie wyjaśnia tendencji uzyskanych w językach finalnych. Żadne z analizowanych podejść nie zakłada częstszego umieszczania krótszego członu na początku koordynacji wraz ze wzrostem różnicy długości członów niezależnie od pozycji nadziednika. Nawet jeśli umieszczanie krótszego członu po lewej stronie wynika z tego, że koordynacje (R) są częstsze w językach finalnych, to nie wiadomo, dlaczego krótszy człon znajduje się częściej po lewej stronie w koordynacjach (R).

Jeśli w językach finalnych umieszczanie krótszego członu po lewej stronie koordynacji uległo gramatykalizacji, musi wynikać to z innych przyczyn, niż z powodu działania efektu DLM.

6.3.2. Monotoniczność tendencji

Dependency Length Minimisation jest zjawiskiem polegającym na możliwym skracaniu relacji zależnościowych występujących w zdaniach. Jego wpływ na ustawianie kolejności członów koordynacji można traktować jako funkcję bezwzględnej różnicy długości członów w prawdopodobieństwo wystąpienia krótszego członu po lewej stronie. Nie wiadomo nic na temat działania tej funkcji, oprócz tego, że jest ona monotoniczna.

Niemniej jednak istnieje możliwość, że pomimo, że efekt DLM ma charakter funkcji monotonicznej, tendencja do zmiany proporcji umieszczania krótszego członu po lewej stronie koordynacji wraz ze zmianą różnicy długości członów nie jest monotoniczna.

Przepiórkowski i in. (2024) przeprowadzili replikację analizy z pracy Przepiórkowski i Woźniak (2023) na większym korpusie języka angielskiego. Przeanalizowali 11 502 053 koordynacji. Dla porównania, Przepiórkowski i Woźniak (2023) zbadali 21 825 konstrukcji współrzędnie złożonych, a w niniejszej pracy przeanalizowano 21 013 koordynacji

języka angielskiego. Wyniki dotyczące koordynacji (L) i (0) potwierdziły tendencje zaobserwowane z pracy Przepiórkowski i Woźniak (2023).

W przypadku koordynacji (R) Przepiórkowski i in. (2024) zaobserwowali niemonotonową zależność między zmianą odsetka występowania krótszych członów na początku koordynacji a różnicą długości członów. Dla niewielkiej różnicy długości (nie przekraczającej 4 słów lub 30 znaków) dla koordynacji (R) zaobserwowano tendencję rosnącą, zaś dla większych różnic tendencję malejącą.

Możliwe, że w przypadku badanych przeze mnie języków istnieją podobne zależności. Jeśli tak jest, taki kształt relacji może wynikać z interakcji między dwiema przeciwnymi tendencjami. W koordynacjach (R) efekt DLM składa się na tendencję spadkową, zaś inne przyczyny na tendencje wzrostową. W przypadku konstrukcji z małą różnicą długości członów efekt DLM jest słaby i nie jest widoczny. Dopiero gdy różnica długości przekracza pewien próg (na przykład czterech słów) efekt DLM staje się silniejszy niż pozostałe tendencje wzrostowe.

Niestety, weryfikacja powyższych spekulacji za pomocą korpusów zależnościowych UD jest niemożliwa. W analizowanych danych znajduje się zbyt mało koordynacji (R) o dużej różnicy długości członów, żeby można było na ich podstawie utworzyć istotny statystycznie model.

6.3.3. Przyczyny tendencji wzrostowych

Przepiórkowski i Woźniak (2023) twierdzą, że tendencje wzrostowe dla koordynacji (0) mogą wynikać z gramatykalizacji umieszczania krótszego członu po lewej stronie. Jak pokazuję w punkcie 6.3.1, wyjaśnienie to może opisywać jedynie wyniki uzyskane dla języków inicjalnych. Gramatykalizacja ta nie może zachodzić w językach finalnych lub może zachodzić z innych powodów, niż twierdzą Przepiórkowski i Woźniak (2023).

W niniejszym punkcie przedstawiam argument mogący stanowić wyjaśnienie występowania różnych tendencji dotyczących koordynacji (R) w badanych przeze mnie językach.

W punkcie 2.5 opisuję różne przyczyny, dla których w języku występuje naturalna tendencja do umieszczania krótszego członu koordynacji po lewej stronie. Należą do nich czynniki pragmatyczne, psycholingwistyczne (argument o częstym używaniu krótszych słów) oraz wynikające z prozodii (argument o rozłożeniu sylab akcentowanych).

Wright i in. (2005) wskazują, że czynniki wynikające z prozodii mają szczególnie mocny wpływ na koordynacje krótkie, tj. o członach jedno- i dwusylabowych. W przypadku takich koordynacji różnica długości członów jest zwykle niewielka, więc efekt DLM może być pomijalny.

Jak wskazują McDonald i in. (1993), koordynacje o krótkich członach są konstruowane tak, żeby akcentowane sylaby tworzyły rytm. Pokazuję to poniższe przykłady:

- (63) [Jan] i [Maria] | zjedli | obiad.
 (64) [Maria] i [Jan] zjedli obiad.
 (65) Obiad | zjedli | [Jan] i [Maria].
 (66) Obiad zjedli [Maria] i [Jan].

W przykładach (63) i (64) nadrzędnik koordynacji *zjedli* znajduje się po prawej stronie, zaś w przykładach (65) i (66) po lewej stronie. Rytm oparty o stopy metryczne powstaje tylko w zdaniach, w których krótszy człon *Jan* występuje jako pierwszy, tj. w (63) i (65). W tych przykładach rozkład sylab akcentowanych jest niezależny od pozycji nadrzędnika. Zgodnie z argumentacją przedstawioną w pracy McDonald i in. (1993) to właśnie zdania (63) i (65) mają większą szansę na pojawienie się w języku naturalnym niż zdania (64) i (66).

To samo rozumowanie nie ma zastosowania w przypadku koordynacji z większą różnicą długości członów.

- (67) [Wysoki brunet Jan] i [kobieta w średnim wieku imieniem Maria] zjedli obiad.
 (68) [Kobieta w średnim wieku imieniem Maria] i [wysoki brunet Jan] zjedli obiad.
 (69) Obiad zjedli [wysoki brunet Jan] i [kobieta w średnim wieku imieniem Maria].
 (70) Obiad zjedli [kobieta w średnim wieku imieniem Maria] i [wysoki brunet Jan].

Dodatkowe podrzędniki głów członów *Jan* oraz *Maria* w przykładach (67)–(70) zaburzają rytm we wszystkich czterech przypadkach. Niezależnie od pozycji krótszego członu osiągnięcie pożądanej sekwenacji naprzemiennie występujących sylab krótkich i długich jest niemożliwy²¹. Powoduje to, że omawiany efekt prozodyczny nie ma w tych przykładach zastosowania. W takiej sytuacji większe znaczenie dla ustalenia kolejności członów może mieć efekt DLM. Jeśli prawidłowe są podejścia symetryczne (jak argumentują Przepiórkowski i Woźniak 2023), to zdanie (68) ma większą szansę pojawić się w języku, niż zdanie (67), zaś zdanie (69) ma większą szansę niż (70).

Kluczową różnicą między zdaniami (63)–(66) i (67)–(70) jest długość członów występujących w nich koordynacji. Możliwe, że w przypadku krótszych członów rytm występuje częściej niż w przypadku dłuższych, ponieważ każdy dodatkowy podrzędnik głowy członu ma szansę na jego zaburzenie. W takiej sytuacji w koordynacjach z krótszymi członami większy wpływ ma efekt ustalenia sylab akcentowanych, zaś w przypadku zdań z dłuższymi koordynacjami efekt DLM.

Warto zauważyć, że przykład zdań (63)–(66) stanowi szczególny przypadek, w którym jeden z członów koordynacji posiada jedną, zaś drugi dwie sylaby. W przypadku

²¹ Rytm zawsze występuje w części zdania – np. fraza Wysoki brunet Jan dzieli się na trzy stopy jambiczne. Kluczowe jest ustalenie sylab na krawędziach członu: we fragmencie Maria i wysoki nie występuje rytm.

koordynacji, w której człony miałyby odpowiednio dwie i trzy sylaby, tendencje byłyby odwrotne. Podane przykłady nie mają na celu wykazania, że prozodia ma większy wpływ na ustawienie członów koordynacji niż efekt DLM ani udowodnienia częstości występowania koordynacji z członami jedno- i dwusylabowymi. Zdania (63)–(66) mają na celu jedynie pokazać, że istnieją sytuacje, w których rozkład sylab akcentowanych ma większy wpływ na ustawienie członów koordynacji niż efekt DLM. Do określenia, jak częste jest to zjawisko i jaka jest jego faktyczna interakcja z efektem DLM, należałyby przeprowadzić osobne badanie.

Koordynacje z dużą różnicą długości członów (przekraczającą kilka słów) są możliwe tylko wtedy, kiedy jeden z członów jest długi. Możliwe, że w przypadku koordynacji z większą różnicą długości członów jest więcej koordynacji z długimi członami, niż w przypadku koordynacji z małą różnicą długości członów. Jeśli to prawda, zgodnie z omówioną powyżej hipotezą efekt DLM ma większy wpływ od efektu wynikającego z prozodii tylko w przypadku koordynacji z różnicą długości członów przekraczającą kilka słów. W takim przypadku interakcja tych dwóch efektów jest jednym z możliwych wyjaśnień wyników uzyskanych w badaniu (Przepiórkowski i in., 2024)

To zjawisko może również tłumaczyć zaobserwowane w niniejszej pracy różnice między językami. Typowe rozkłady sylab akcentowanych różnią się w zależności od języka. Ponadto, mogą być one zależne w większym stopniu od przynależności do danej rodziny językowej niż związane z inicjalnością lub finalnością języka. Jeśli w językach romańskich i języku koreańskim rytm zdania ma większy wpływ na ułożenie kolejności słów w zdaniu i nie jest tak łatwo zaburzany jak w języku angielskim i językach słowiańskich, może to być wyjaśnieniem otrzymanych wyników.

Weryfikacja postawionej w niniejszym punkcie hipotezy nie jest możliwa na podstawie metod i danych użytych w niniejszej pracy. W następnym rozdziale 7 omawiam przyczyny tego faktu oraz propozycje badań mogących sprawdzić opisaną przeze mnie hipotezę.

Rozdział 7

Zakończenie

7.1. Ograniczenia

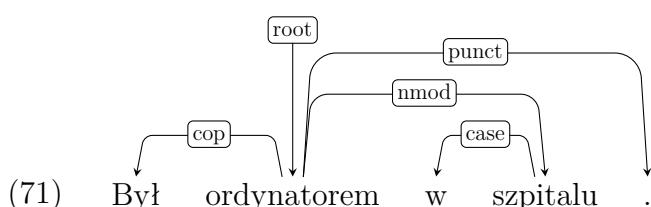
7.1.1. Korpusy zależnościowe UD

Jednym z głównych ograniczeń niniejszej pracy jest korzystanie wyłącznie z korpusów zależnościowych Universal Dependencies. Dzięki standardowemu opisowi relacji składniowej korpusy UD umożliwiają miarodajną analizę wielu różnych języków naraz. Jednak ze względu na wymuszanie przez UD konkretnego sposobu opisu struktur, korpusy mogą tracić na jakości.

Ograniczenia dla języków finalnych Choi i Palmer (2011) oraz Kanayama i in. (2018) wskazują, że tworzenie korpusów zależnościowych zgodnie z narzuconym przez UD podejściem stanfordzkim tworzą drzewa niezgodne z teorią lingwistyczną. Z tego powodu z japońskich korpusów Universal Dependencies zostały usunięte wszystkie koordynacje, a w korpusach koreańskich występują równolegle dwa różne standardy opisu struktur.

Powoduje to, że analiza koordynacji w korpusach japońskich jest niemożliwa, zaś w przypadku korpusów koreańskich może być w znacznym stopniu niemiarodajna.

Nieprawidłowy opis głów fraz Standard Universal Dependencies nie jest wyłącznie standardem składniowym; przy ustalaniu relacji zależnościowych stosuje również kryteria semantyczne. Oznacza to, że według wytycznych tego standardu głową członu nie jest faktyczna, składniowa główna członu, lecz temat lub najważniejszy semantycznie token w obrębie frazy.



W przykładzie (71) głową frazy *w szpitalu* jest *szpitalu*, zaś głową całego zdania (czyli korzeniem zdania) *ordynatorem*. Zgodnie z teorią lingwistyczną głową frazy przyimkowej *w szpitalu* powinien być przyimek *w*, zaś korzeniem zdania z orzeczeniem imiennym – łącznik *Był* (Hoeksema, 1992). Traktowanie rzeczowników jako głów fraz przyimkowych i orzeczników jako korzeni zdań jest znamienne dla drzew UD. Efektem tych rozwiązań jest w szczególności „przesuwanie” głowy członu na jej koniec w językach inicjalnych.

Niemniej jednak efekty tych rozwiązań nie powinny być na tyle powszechnne, żeby mieć znaczący wpływ na wynik mojej analizy. Powtórzona poniżej Tabela 5.1 pokazuje względną pozycję głów członów koordynacji w badanych językach. Gdyby w korpusach UD faktycznie istniała częsta tendencja do nieprawidłowego oznaczania późniejszych elementów fraz jako ich główny, średnia pozycja główny w językach inicjalnych byłaby większa niż 0,5. Wobec tego należy uznać znakowanie UD za zasadniczo poprawne.

Język	lewy człon				prawy człon			
	N	średnia	t	p	N	średnia	t	p
Języki inicjalne								
angielski	12326	0,36	-37	1.74e-288	15155	0,40	-31	1.99e-199
czeski	51416	0,34	-88	0	62872	0,39	-70	0
hiszpański	19685	0,30	-74	0	22137	0,31	-80	0
islandzki	31929	0,18	-196	0	36967	0,18	-225	0
polski	9976	0,23	-74	0	12049	0,33	-50	0
portugalski	19732	0,30	-71	0	22349	0,30	-84	0
rosyjski	36608	0,38	-56	0	46050	0,41	-47	0
rumuński	27224	0,37	-53	0	31024	0,37	-60	0
włoski	17728	0,43	-22	1,2e-107	20330	0,37	-49	0
Języki mieszane								
łaciński	19766	0,36	-47	0	25755	0,48	-5,8	7,98e-09
niemiecki	51068	0,53	17	9,27e-62	64616	0,52	14	2,61e-46
Języki finalne								
koreański	6801	0,78	59	0	12951	0,65	47	0
turecki	7994	0,64	28	7,97e-167	12763	0,69	55	0

Tabela 5.1: Względna pozycja głów członów koordynacji.

7.1.2. Dependency Length Minimisation

Założenie o monotoniczności tendencji Badając wpływ DLM na kolejność ustawnienia członów koordynacji, zakładam, że jest on wyrażony jako monotoniczna funkcja.

Z tego powodu, obliczając tendencję, korzystam z dwumianowej regresji logistycznej.

Dokładniejsze zbadanie właściwości efektu DLM może pozwolić na stworzenie dokładniejszych modeli, które są w stanie estymować proporcje ustawiania członów lepiej, niż przez samo określenie, że są one opisywane przez rosnącą, malejącą lub stałą funkcję.

Metoda określania złożoności Lohmann (2014) stwierdza, że najlepszym sposobem na określenie długości relacji jest złożoność syntaktyczna rozumiana jako liczba węzłów w relacjach składnikowych łączących słowa w frazy. Ponieważ korzystam z korpusów zależnościowych, w analizie nie mogę obliczyć złożoności syntaktycznej. Estymacja złożoności składniowej za pomocą liczby słów nie jest idealnym rozwiązaniem i może wpływać na wyniki badania.

Jest to poważne ograniczenie badania, tym bardziej, że w wielu badanych przeze mnie językach wyniki dotyczące słów są istotnie inne, niż wyniki korzystające z innych miar długości.

7.1.3. Ograniczenia metodologiczne

W przeprowadzonym przeze mnie badaniu analizuję koordynacje wyciągnięte automatycznie na podstawie algorytmu opartego o heurystyki. Każda z heurystyk została opracowana na podstawie wielu kompromisów i założeń oraz ze świadomością, że nie zawsze będą one działać poprawnie. Ponadto, większość heurystyk została przypisana językom bez znajomości języków, na zasadzie analogii z językiem angielskim, polskim i tureckim.

Ewaluacja losowo wyciągniętych koordynacji wykazała, że większość z nich została wyciągnięta poprawnie. Niemniej jednak sam algorytm liczenia długości członów nie został poddany ewaluacji.

7.2. Przyszłe badania

7.2.1. Surface Syntactic Universal Dependencies (SUD)

Surface Syntactic Universal Dependencies²² to projekt mający na celu stworzenie alternatywnego dla UD standardu opisu relacji zależnościowych. Jest on z założenia bardziej wierny powierzchniowym relacjom składniowym, zaś na jego wytyczne nie mają wpływ argumenty semantyczne. Dzięki temu korpusy SUD są bardziej wiarygodne, jeśli chodzi o opis relacji zależnościowych i unikają problemów obecnych w zwykłych korpusach UD.

Istnieje możliwość, że przeprowadzenie analogicznego badania na korpusach SUD pozwoliłoby uzyskać bardziej wiarygodne wyniki. Niestety niska objętość korpusów

²²<https://surfacesyntacticud.github.io/>

SUD nie pozwala na wykrycie istotnych statystycznie tendencji dotyczących konstrukcji współrzędnie złożonych. W celu uzyskania wystarczającej ilości danych niezbędne jest automatyczne parsowanie zdań – takie rozwiązanie zostało zastosowane w pracy Borysiak (2024).

7.2.2. Analiza koordynacji różnych długości

Przepiórkowski i in. (2024) pokazują, że omawiane tendencje dotyczące koordynacji w języku angielskim mogą być niemonotoniczne. Z ich badania wynika, że proste modele statystyczne, takie jak regresja logistyczna, mogą nie wystarczyć w analizie tych zależności.

W rozdziale 6 przedstawiam hipotezę, zgodnie z którą uwzględnienie w analizie tylko koordynacji o dłuższych członach może pozwolić uzyskać bardziej wiarygodne wyniki. Hipoteza ta jest zgodna z wynikami uzyskanymi w pracy Przepiórkowski i in. (2024).

Niestety, konstrukcje współrzędnie złożone o dłuższych członach występują w języku naturalnym bardzo rzadko. W przypadku analizowanych przeze mnie danych uwzględnienie w analizie wyłącznie koordynacji, których oba człony składają się z co najmniej czterech tokenów nie pozwala na stworzenie istotnych statystycznie modeli logistycznych.

Oznacza to, że korpusy UD są zbyt małe, żeby można było przeprowadzić na nich takie badanie. Analiza koordynacji różnych długości powinna zostać przeprowadzona na korpusach zawierających miliony zdań.

Dodatek A

Rozszerzone zależności składniowe

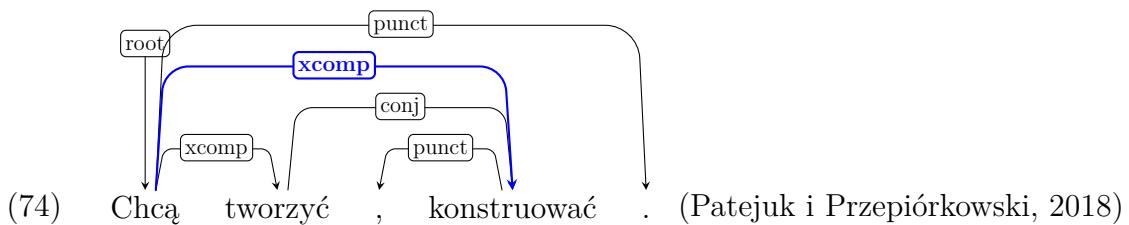
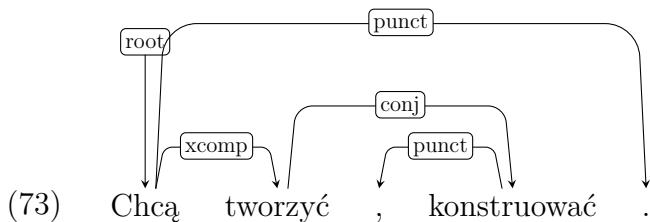
A.1. Enhanced Dependencies

W niniejszej pracy analizuję relacje składniowe w formie drzew zależnościowych. Oznacza to, że zakładam, że relacje składniowe w zdaniach tworzą drzewa. Jest to dość powszechnie w językoznawstwie założenie, stojące u podstaw gramatyki zależnościowej. Jednak nie musi być ono prawdziwe.

Jak zauważają De Marneffe i in. (2021), w języku istnieje wiele konstrukcji, w przypadku których opis relacji zależnościowych za pomocą drzew nie jest optymalny. Należą do nich m.in. koordynacje.

Z tego powodu twórcy Universal Dependencies stworzyli rozszerzenie standardu o nazwie Enhanced Dependencies (ED)²³. Przykład (73) pokazuje drzewo zależnościowe zdania (72) opisane według zwykłego UD. (74) jest drzewem zdania (72) opisanym według ED. Rozszerzona zależność xcomp została wyróżniona.

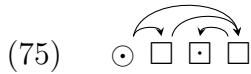
- (72) Chcą tworzyć, konstruować.



Przepiórkowski i Woźniak (2023) zauważają, że uwzględnienie rozszerzonych relacji zależnościowych w strukturze koordynacji zmienia przewidywanie modeli dotyczące

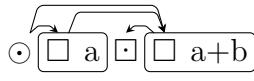
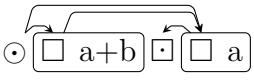
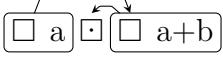
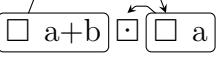
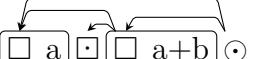
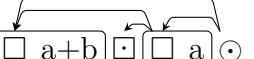
²³<https://universaldependencies.org/u/overview/enhanced-syntax.html>

omawianych tendencji. Prezentują model struktury zależnościowej koordynacji oparty na założeniach ED:



A.2. Języki inicjalne

W językach inicjalnych model (75) przewiduje następujące długości relacji:

(L-L)		$S = 2a$	(L-R)		$S = 2a + 2b$
(0-L)		$S = a$	(0-R)		$S = a + b$
(R-L)		$S = 4a + 2b$	(R-R)		$S = 4a + 2b$

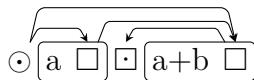
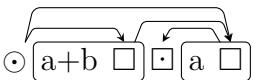
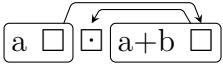
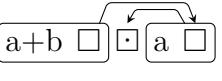
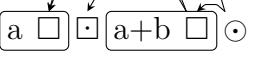
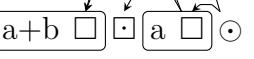
Wraz ze wzrostem różnicy długości członów (b):

- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **znacznie rośnie**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **rośnie**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **nie zmienia się**.

Przewidywania modelu (75) pokrywają się z tendencjami zaobserwowanymi przez Przepiórkowski i Woźniak (2023) oraz przeze mnie w przypadku języka angielskiego i języków słowiańskich.

A.3. Języki finalne

W językach finalnych model (75) przewiduje następujące długości relacji:

(L-L)		$S = 5a + 3b$	(L-R)		$S = 5a + 2b$
(0-L)		$S = 2a + 2b$	(0-R)		$S = 2a$
(R-L)		$S = 3a + 3b$	(R-R)		$S = 3a$

To podejście przewiduje, że wraz ze wzrostem różnicy długości członów (b):

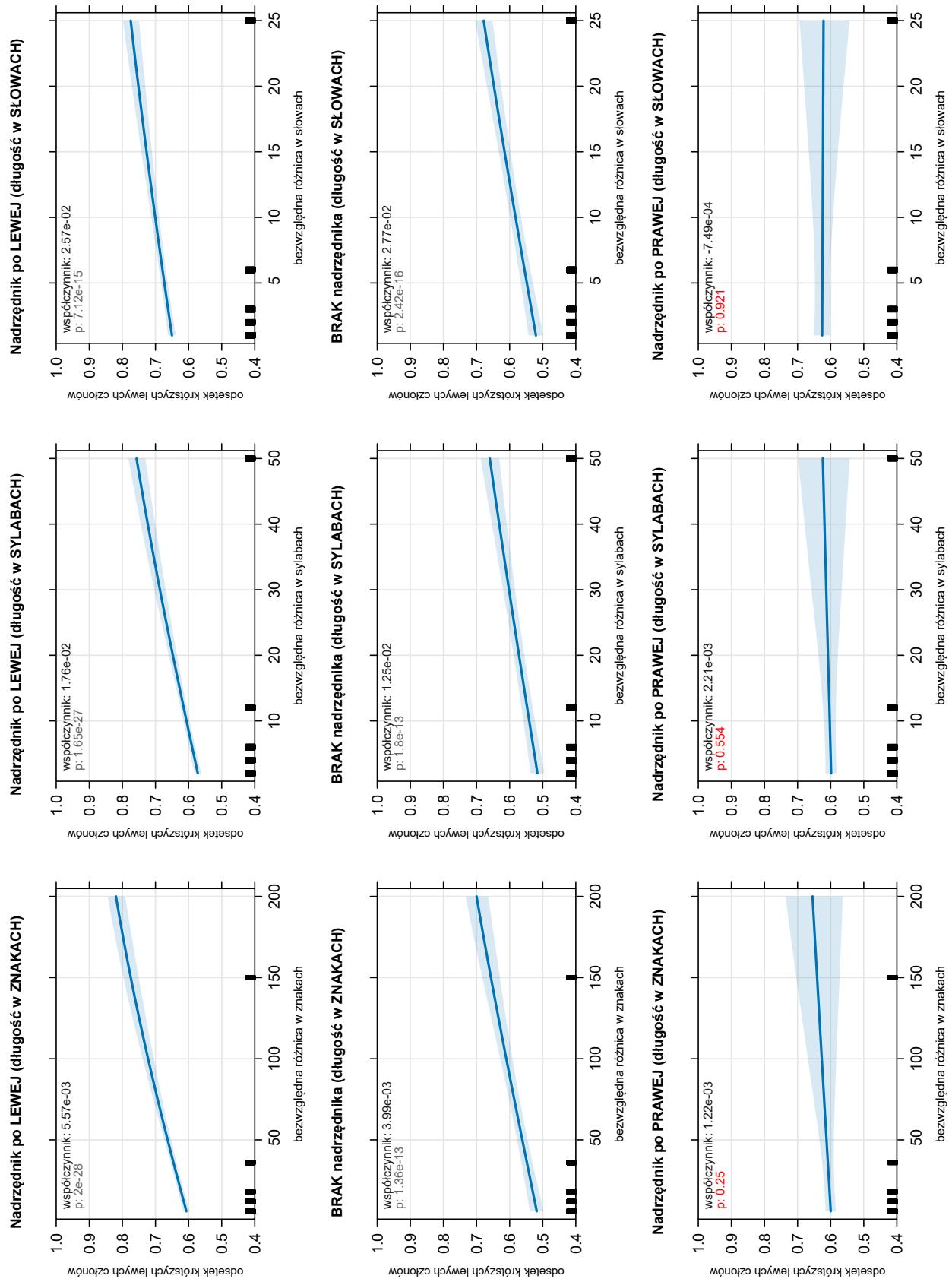
- odsetek koordynacji (L-L) względem wszystkich koordynacji (L) **spada**;
- odsetek koordynacji (0-L) względem wszystkich koordynacji (0) **znacznie spada**;
- odsetek koordynacji (R-L) względem wszystkich koordynacji (R) **bardzo znacznie spada**.

Przewidywania modelu (75) nie pokrywają się z uzyskanymi wynikami dla języka koreańskiego i tureckiego. Oznacza to, że model oparty na rozszerzonych zależnościach składniowych nie wyjaśnia uzyskanych wyników²⁴.

²⁴W przypadku języków finalnych należy rozpatrzyć też podejście zakładające, że spójnik jest podrzędnikiem prawego członu. W przypadku tego wariantu predykcje również nie pokrywają się z uzyskanymi wynikami.

Dodatek B

Wyniki poprzednich badań



Rysunek B.1: Różnica długości członów a występowanie krótszego członu po lewej stronie w języku angielskim – wyniki uzyskane w pracy Przepiorkowski i Woźniak (2023)

Dodatek C

Długość członów koordynacji

C.1. Języki inicjalne

	medianą	średnią				
	lewy	prawy	lewy	prawy	V	p
Język angielski						
Wszystkie koordynacje (N = 21 013)						
znaki	13	20	22,79	32,69	5,3e+07	0
sylaby	4	5	6,00	8,51	4e+07	0
słowa	2	4	4,19	5,96	2,3e+07	0
Brak nadziednika (N = 6 829)						
znaki	30	40	37,41	50,99	6,9e+06	7,4e-154
sylaby	8	10	9,61	13,00	6e+06	5,7e-148
słowa	6	8	7,13	9,65	5,1e+06	3,6e-159
Nadziednik po lewej (N = 11 171)						
znaki	10	16	17,28	26,49	1,3e+07	0
sylaby	3	4	4,64	6,99	9,5e+06	0
słowa	2	3	3,05	4,66	4,4e+06	0
Nadziednik po prawej (N = 2 972)						
znaki	7	9	10,10	13,90	8,1e+05	6,5e-99
sylaby	2	2	2,88	3,88	4,9e+05	5,2e-75
słowa	1	1	1,74	2,34	8e+04	1,8e-68

	medianą	średnią				
	lewy	prawy	lewy	prawy	V	p
Język czeski						
Wszystkie koordynacje (N = 90 566)						
znaki	13	21	24,44	34,64	1e+09	0
sylaby	5	8	8,93	12,43	8,7e+08	0
słowa	2	3	3,70	5,31	3,8e+08	0
Brak nadziednika (N = 26 688)						
znaki	31	45	39,54	57,18	9,9e+07	0
sylaby	11	16	14,16	20,14	9,3e+07	0
słowa	5	7	6,13	9,08	7,1e+07	0
Nadziednik po lewej (N = 49 341)						
znaki	10	16	19,59	27,69	2,8e+08	0
sylaby	4	6	7,28	10,10	2,4e+08	0
słowa	1	2	2,93	4,14	8,7e+07	0
Nadziednik po prawej (N = 14 279)						
znaki	8	10	12,83	16,28	2,2e+07	0
sylaby	3	4	4,81	6,02	1,6e+07	3,3e-291
słowa	1	1	1,82	2,29	2,2e+06	5,1e-202

	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język hiszpański							
Wszystkie koordynacje (N = 28 666)							
znaki	18	26	30,70	43,79	1,1e+08	0	
sylaby	6	9	10,84	15,08	9,6e+07	0	
słowa	3	5	5,20	7,50	5,1e+07	0	
Brak nadziednika (N = 6 300)							
znaki	45	65	55,91	79,19	5,7e+06	2e-169	
sylaby	16	22	19,59	27,08	5,5e+06	1,3e-154	
słowa	8	11	9,64	13,77	4,6e+06	5,5e-176	
Nadziednik po lewej (N = 19 557)							
znaki	14	20	23,83	34,71	4,9e+07	0	
sylaby	5	7	8,46	12,00	4,2e+07	0	
słowa	2	3	3,98	5,88	1,8e+07	0	
Nadziednik po prawej (N = 2 751)							
znaki	11	14	21,88	26,91	1e+06	2,6e-34	
sylaby	4	5	7,74	9,38	8,6e+05	1,8e-29	
słowa	2	2	3,69	4,59	3,6e+05	1,9e-29	
	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język polski							
Wszystkie koordynacje (N = 16 684)							
znaki	15	21	22,52	32,73	3,4e+07	0	
sylaby	5	7	7,57	10,80	2,8e+07	0	
słowa	2	3	3,26	4,77	1,4e+07	0	
Brak nadziednika (N = 6 023)							
znaki	25	35	31,93	46,85	4,8e+06	2,2e-189	
sylaby	8	11	10,50	15,18	4,4e+06	3,1e-182	
słowa	4	5	4,80	7,10	3,3e+06	2,2e-206	
Nadziednik po lewej (N = 8 407)							
znaki	11	17	18,20	26,57	8,3e+06	1,7e-264	
sylaby	4	6	6,26	8,93	6,5e+06	2,5e-247	
słowa	1	2	2,52	3,70	2,3e+06	3e-242	
Nadziednik po prawej (N = 2 219)							
znaki	8	11	13,21	17,59	5,3e+05	1,7e-68	
sylaby	3	4	4,55	5,95	3,8e+05	1,5e-58	
słowa	1	1	1,90	2,48	9,4e+04	4,2e-40	
	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język islandzki							
Wszystkie koordynacje (N = 43 852)							
znaki	24	28	36,25	42,06	3,5e+08	1,2e-266	
sylaby	6	7	9,31	10,60	3e+08	2,1e-173	
słowa	4	5	6,40	7,39	2,1e+08	3,2e-191	
Brak nadziednika (N = 23 877)							
znaki	38	42	52,54	57,05	1,2e+08	1,1e-46	
sylaby	10	11	13,38	14,32	1,2e+08	2,8e-30	
słowa	7	8	9,38	10,24	1,1e+08	3,1e-53	
Nadziednik po lewej (N = 16 986)							
znaki	10	13	17,65	25,91	3,7e+07	0	
sylaby	3	4	4,65	6,58	2,8e+07	6,8e-285	
słowa	1	2	2,99	4,29	1,2e+07	2,2e-250	
Nadziednik po prawej (N = 2 928)							
znaki	7	9	11,30	14,00	1,1e+06	1,1e-54	
sylaby	2	3	3,11	3,76	8e+05	2,2e-35	
słowa	1	1	1,83	2,19	2e+05	4,1e-23	
	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język portugalski							
Wszystkie koordynacje (N = 29 255)							
znaki	15	22	25,71	36,14	1,2e+08	0	
sylaby	6	8	9,64	13,31	1e+08	0	
słowa	3	4	4,42	6,22	5,2e+07	0	
Brak nadziednika (N = 8 364)							
znaki	30	39	40,04	52,84	1,1e+07	4,3e-132	
sylaby	11	14	14,63	19,03	1,1e+07	1,2e-124	
słowa	5	7	6,95	9,19	8e+06	2e-133	
Nadziednik po lewej (N = 17 661)							
znaki	12	18	20,35	30,64	3,7e+07	0	
sylaby	5	7	7,79	11,46	3,1e+07	0	
słowa	2	3	3,45	5,22	1,3e+07	0	
Nadziednik po prawej (N = 3 157)							
znaki	10	13	17,72	22,38	1,4e+06	3,1e-41	
sylaby	4	5	6,72	8,37	1,2e+06	1,3e-37	
słowa	2	2	3,08	3,91	4,3e+05	1,5e-37	

	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język rosyjski							
Wszystkie koordynacje (N = 61 004)							
znaki	15	23	24,46	35,36	4,8e+08	0	
sylaby	5	8	8,79	12,45	4,1e+08	0	
słowa	2	3	3,49	5,04	2,1e+08	0	
Brak nadziednika (N = 20 556)							
znaki	27	37	35,85	49,22	6,4e+07	0	
sylaby	10	13	12,75	17,13	5,9e+07	0	
słowa	4	6	5,25	7,24	4,5e+07	0	
Nadziednik po lewej (N = 31 679)							
znaki	12	20	20,41	31,30	1,2e+08	0	
sylaby	5	7	7,38	11,10	9,9e+07	0	
słowa	2	3	2,85	4,37	4,4e+07	0	
Nadziednik po prawej (N = 8 485)							
znaki	9	12	12,14	16,98	7,1e+06	0	
sylaby	3	4	4,49	6,18	5,2e+06	1,6e-268	
słowa	1	1	1,64	2,24	8,5e+05	2,1e-224	
Język rumuński							
Wszystkie koordynacje (N = 37 247)							
znaki	17	25	25,66	38,92	1,7e+08	0	
sylaby	6	9	9,28	13,83	1,5e+08	0	
słowa	3	5	4,43	6,88	8,8e+07	0	
Brak nadziednika (N = 11 993)							
znaki	28	44	37,49	56,80	1,9e+07	0	
sylaby	10	15	13,26	19,74	1,8e+07	0	
słowa	5	8	6,68	10,44	1,5e+07	0	
Nadziednik po lewej (N = 21 873)							
znaki	13	20	20,50	31,84	5,4e+07	0	
sylaby	5	7	7,55	11,52	4,6e+07	0	
słowa	2	3	3,45	5,45	2,3e+07	0	
Nadziednik po prawej (N = 3 088)							
znaki	10	13	17,31	20,63	1,4e+06	2,7e-36	
sylaby	4	5	6,40	7,59	1,2e+06	1,7e-32	
słowa	2	2	2,81	3,43	4,8e+05	4,8e-36	

	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język włoski							
Wszystkie koordynacje (N = 25 426)							
znaki	15	25	25,91	40,83	7,4e+07	0	
sylaby	6	9	9,64	14,84	6,2e+07	0	
słowa	3	4	4,47	7,03	3,3e+07	0	
Brak nadziednika (N = 5 992)							
znaki	30	45	41,48	61,40	4,8e+06	3,3e-189	
sylaby	11	16	15,05	21,90	4,5e+06	2,4e-179	
słowa	5	8	7,15	10,72	3,7e+06	7,2e-190	
Nadziednik po lewej (N = 17 014)							
znaki	13	21	21,63	35,97	3e+07	0	
sylaby	5	8	8,17	13,19	2,5e+07	0	
słowa	2	4	3,72	6,15	1,1e+07	0	
Nadziednik po prawej (N = 2 345)							
znaki	10	13	17,15	23,54	6,3e+05	1,3e-62	
sylaby	4	5	6,48	8,79	4,6e+05	1,8e-60	
słowa	2	2	2,99	4,04	1,6e+05	9,7e-49	

C.2. Języki mieszane

	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język łaciński							
Wszystkie koordynacje (N = 39 510)							
znaki	11	17	23,77	31,66	2,2e+08	0	
sylaby	4	6	7,94	9,92	2,1e+08	0	
słowa	1	3	3,57	4,74	8,4e+07	0	
Brak nadziednika (N = 9 112)							
znaki	24	34	40,79	52,62	1,4e+07	1,3e-122	
sylaby	7	10	12,81	14,97	1,5e+07	7,5e-42	
słowa	4	5	6,16	7,98	8,9e+06	3,6e-135	
Nadrzędnik po lewej (N = 19 635)							
znaki	10	17	20,69	28,66	4,9e+07	0	
sylaby	4	6	7,00	9,25	4,8e+07	1,9e-269	
słowa	1	2	3,08	4,26	1,7e+07	0	
Nadrzędnik po prawej (N = 9 264)							
znaki	8	10	14,27	17,59	1,1e+07	4,8e-176	
sylaby	3	4	5,35	6,41	9,6e+06	9,1e-116	
słowa	1	1	2,15	2,57	2,7e+06	6,3e-93	
Język niemiecki							
Wszystkie koordynacje (N = 92 115)							
znaki	15	26	26,71	37,86	1,1e+09	0	
sylaby	5	9	8,72	12,12	9,5e+08	0	
słowa	2	3	3,64	5,09	4,3e+08	0	
Brak nadziednika (N = 24 029)							
znaki	46	56	51,35	64,93	9,4e+07	0	
sylaby	14	17	16,03	20,02	9e+07	0	
słowa	7	8	7,26	9,21	7,3e+07	0	
Nadrzędnik po lewej (N = 43 089)							
znaki	12	20	19,35	30,61	2e+08	0	
sylaby	4	7	6,56	10,03	1,8e+08	0	
słowa	1	2	2,55	3,98	6,1e+07	0	
Nadrzędnik po prawej (N = 23 637)							
znaki	11	16	15,29	22,72	5,9e+07	0	
sylaby	4	6	5,27	7,64	5e+07	0	
słowa	1	2	1,95	2,80	1,2e+07	0	

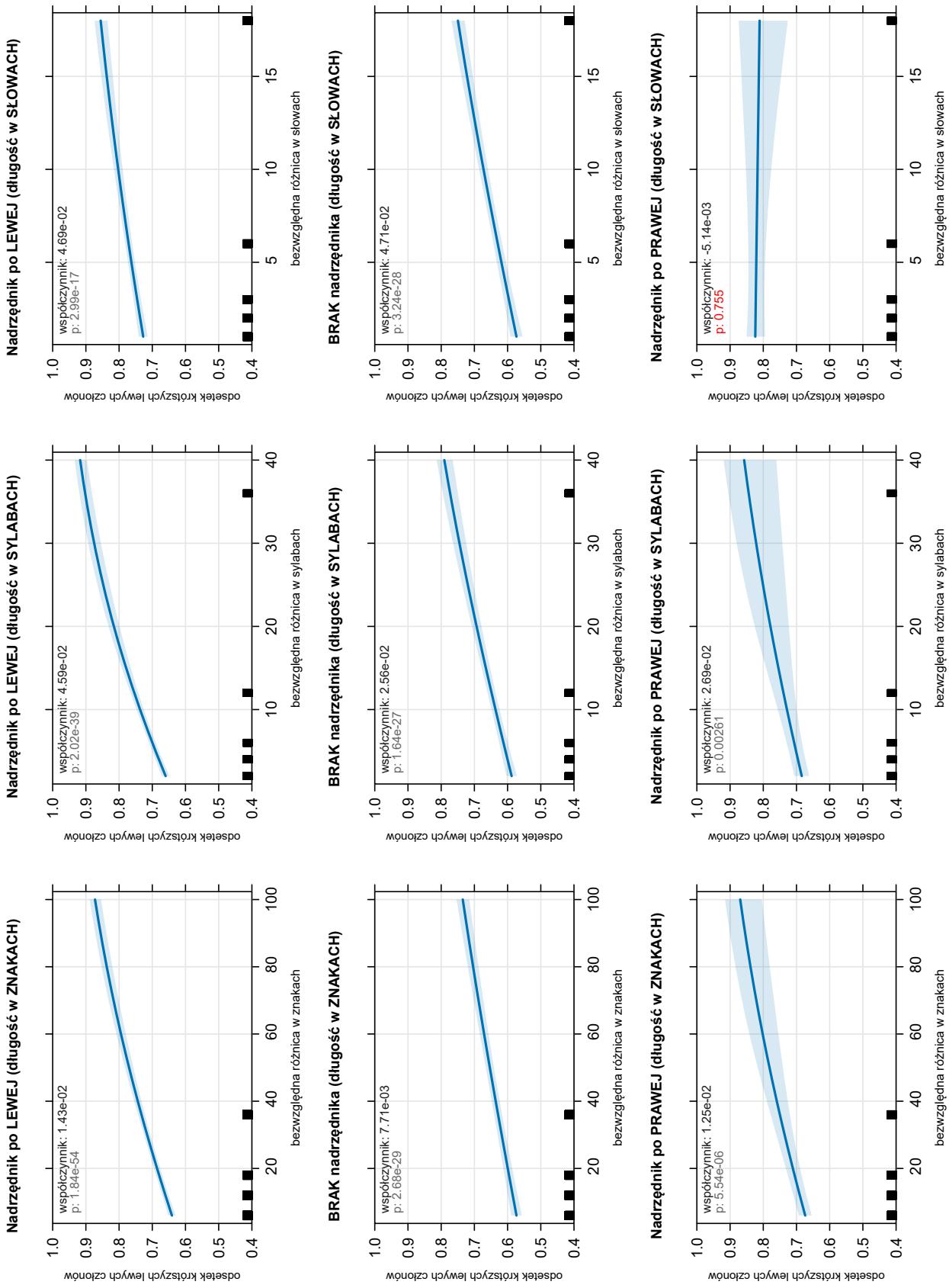
C.3. Języki finalne

	medianą	średnia					
	lewy	prawy	lewy	prawy	V	p	
Język koreański							
Wszystkie koordynacje (N = 21 506)							
znaki	10	15	15,99	24,44	4,7e+07	0	
sylaby	4	6	6,35	9,08	4e+07	0	
słowa	1	2	1,94	2,88	1e+07	0	
Brak nadziednika (N = 6 491)							
znaki	14	35	24,13	41,71	3,8e+06	0	
sylaby	5	13	9,35	15,12	3,9e+06	0	
słowa	2	4	2,86	4,86	2,4e+06	0	
Nadrzędnik po lewej (N = 718)							
znaki	10	15	13,31	20,92	4,8e+04	2,5e-38	
sylaby	4	6	5,35	7,68	4,7e+04	3,4e-27	
słowa	1	2	1,66	2,47	1,4e+04	2,3e-27	
Nadrzędnik po prawej (N = 14 289)							
znaki	9	11	12,42	16,77	2,3e+07	0	
sylaby	4	5	5,04	6,40	1,7e+07	8e-292	
słowa	1	1	1,53	2,01	2,3e+06	0	
Język turecki							
Wszystkie koordynacje (N = 19,598)							
znaki	9	15	14,56	22,60	3,4e+07	0	
sylaby	4	6	5,71	8,69	2,7e+07	0	
słowa	1	2	2,07	3,14	1e+07	0	
Brak nadziednika (N = 5,758)							
znaki	12	27	22,36	35,81	3,3e+06	1,6e-301	
sylaby	5	10	8,54	13,42	3e+06	6,9e-282	
słowa	2	4	3,15	4,91	2e+06	3,1e-237	
Nadrzędnik po lewej (N = 1,760)							
znaki	8	15	11,06	20,35	2e+05	1,3e-135	
sylaby	3	6	4,35	7,74	1,5e+05	3,8e-125	
słowa	1	2	1,57	2,86	4,8e+04	1,3e-112	
Nadrzędnik po prawej (N = 11,936)							
znaki	8	11	11,37	16,59	1,2e+07	0	
sylaby	3	5	4,57	6,56	9,1e+06	0	
słowa	1	2	1,63	2,32	2,1e+06	0	

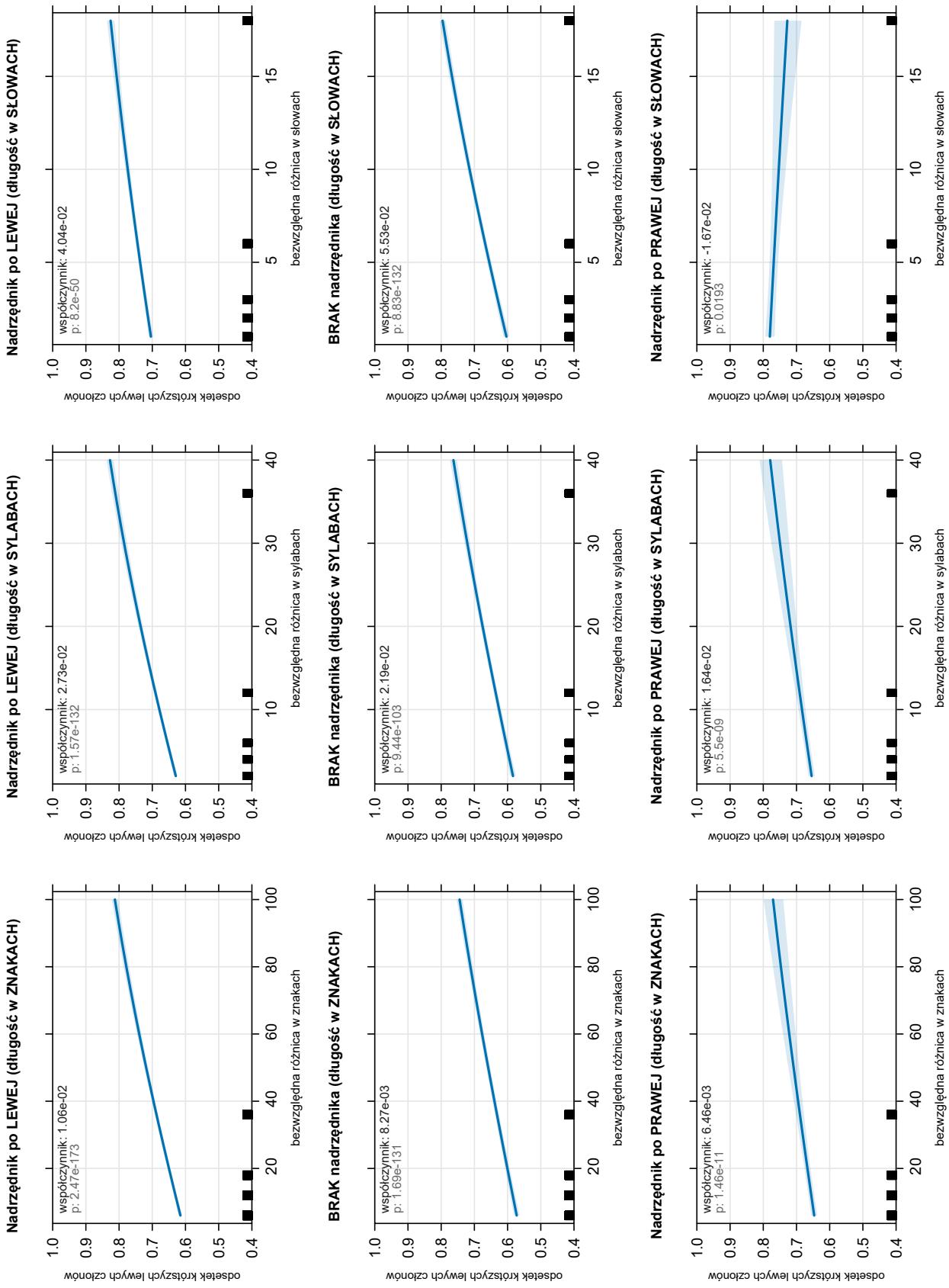
Dodatek D

Różnica długości członów a pozycja krótszego członu

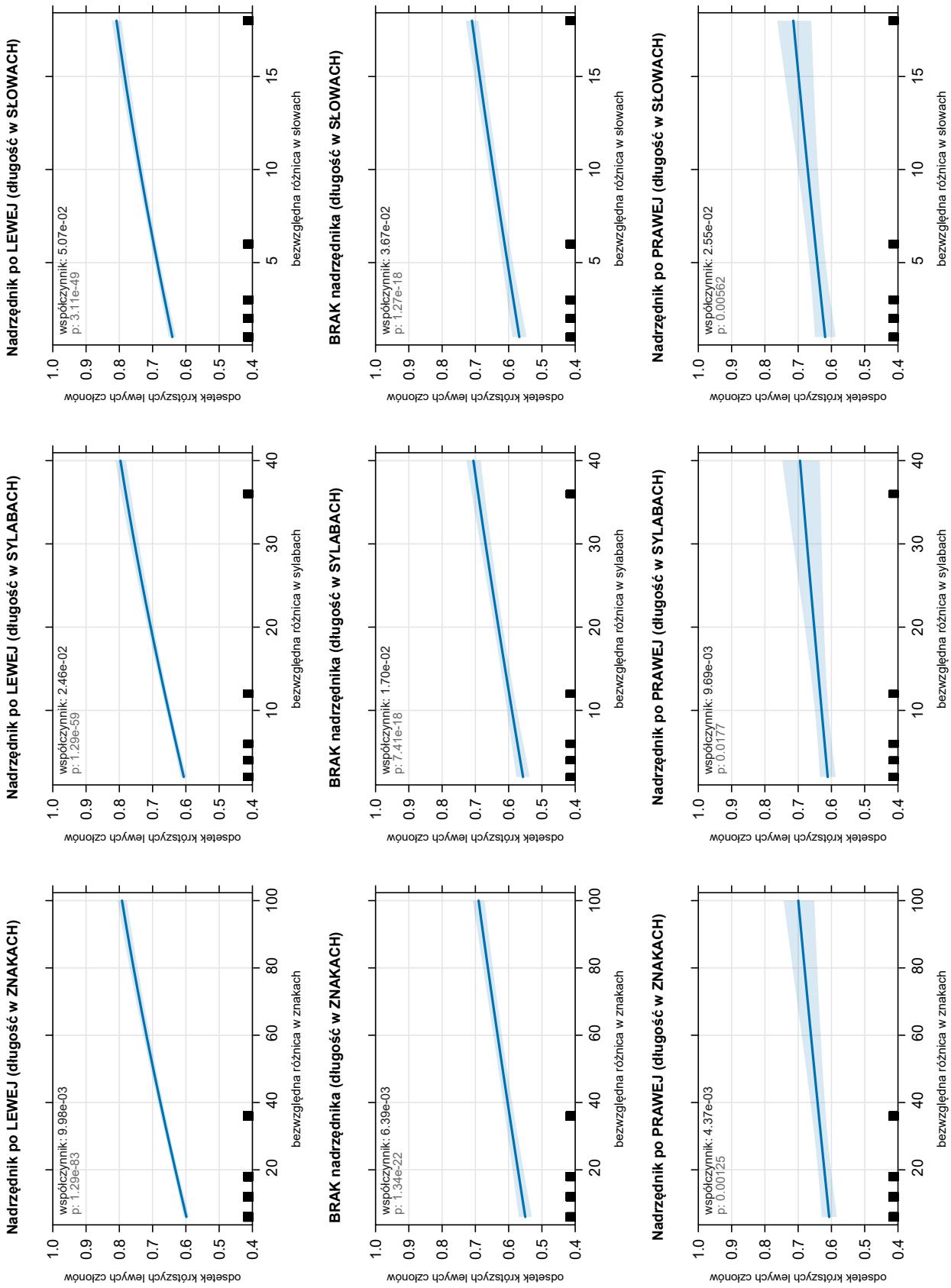
D.1. Języki inicjalne



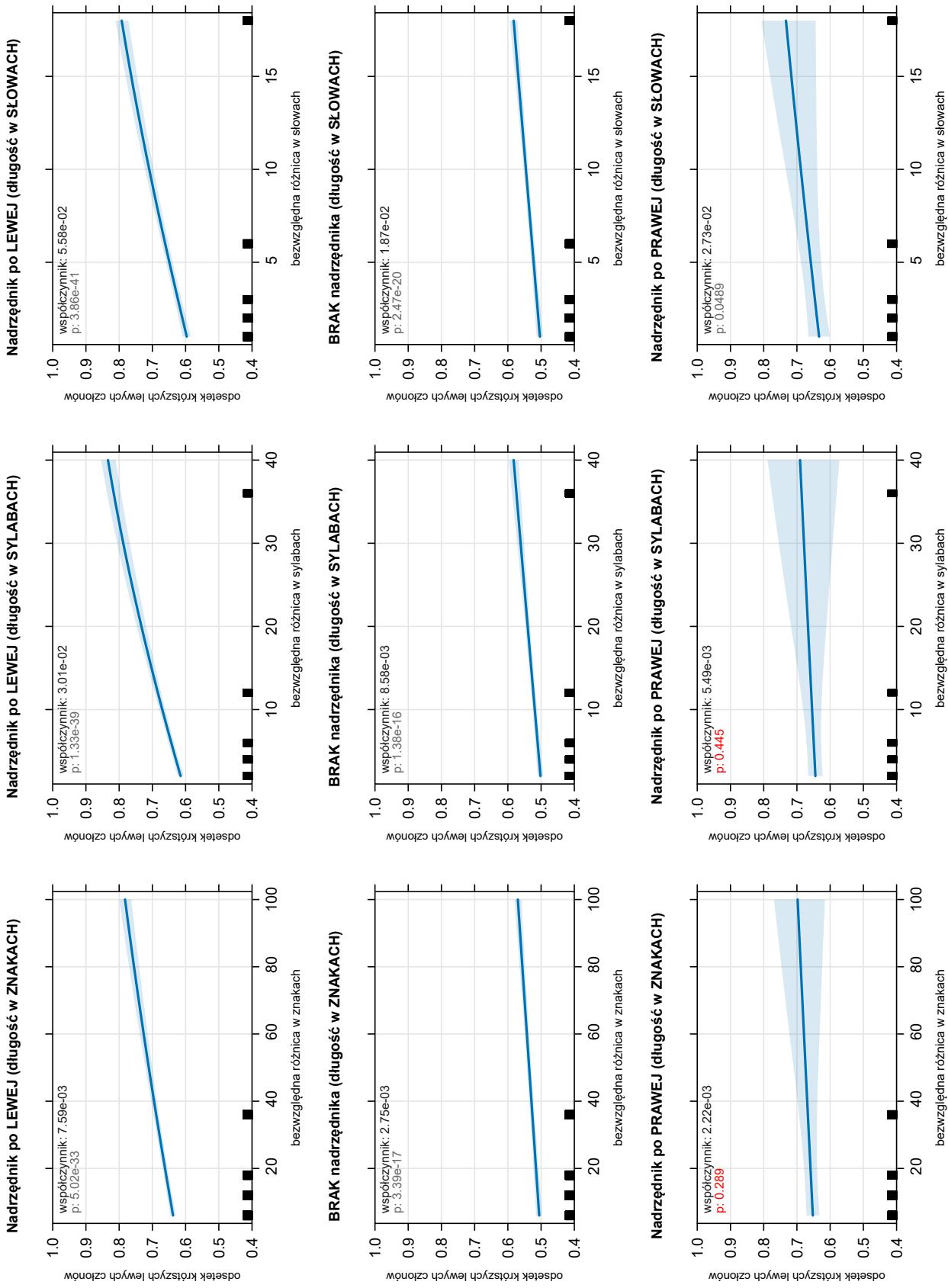
Rysunek D.1: Różnica długości członów a występowanie krótszego członu po lewej stronie – język angielski



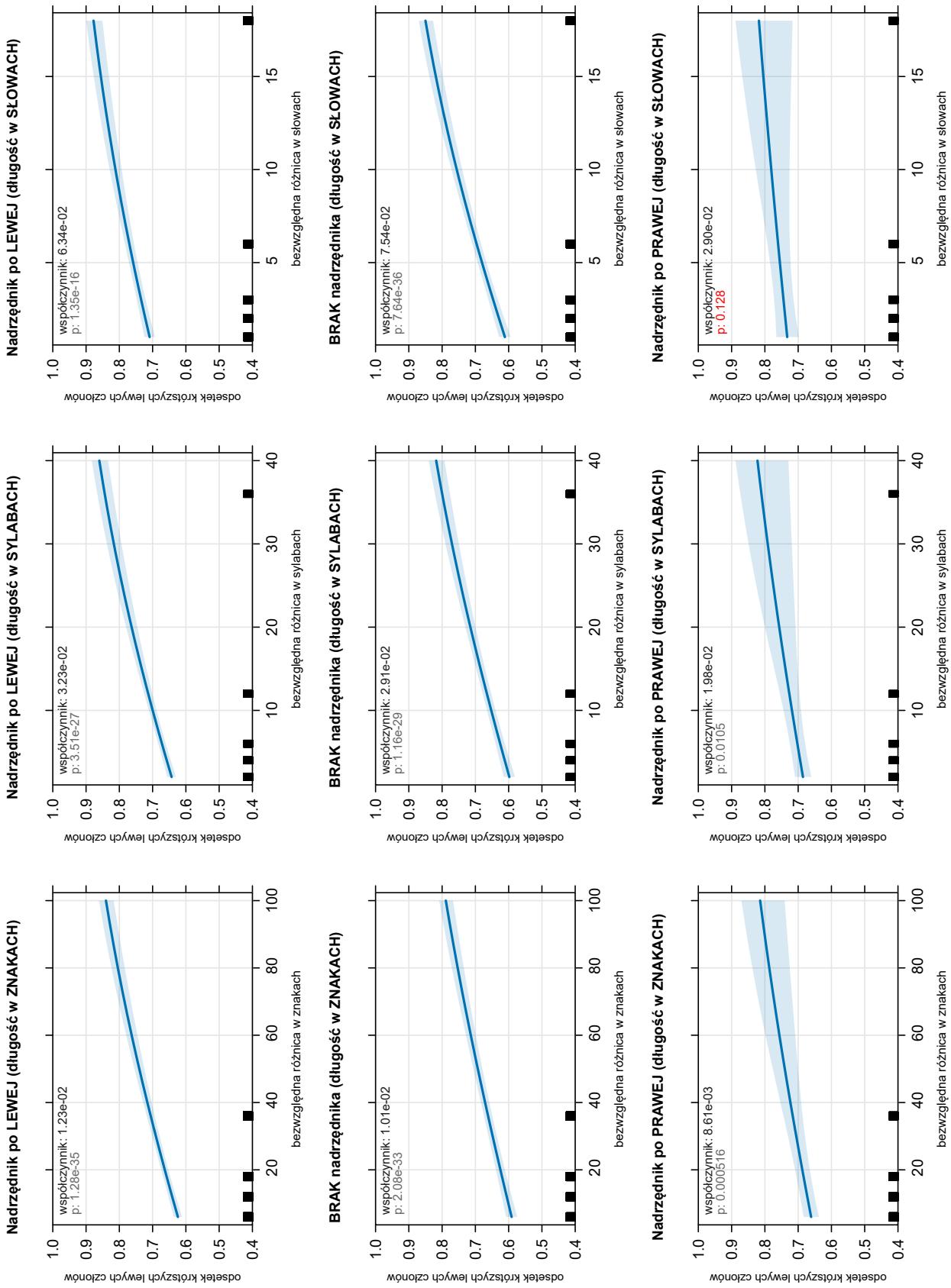
Rysunek D.2: Różnica długości członów a występowanie krótszego członu po lewej stronie – język czeski



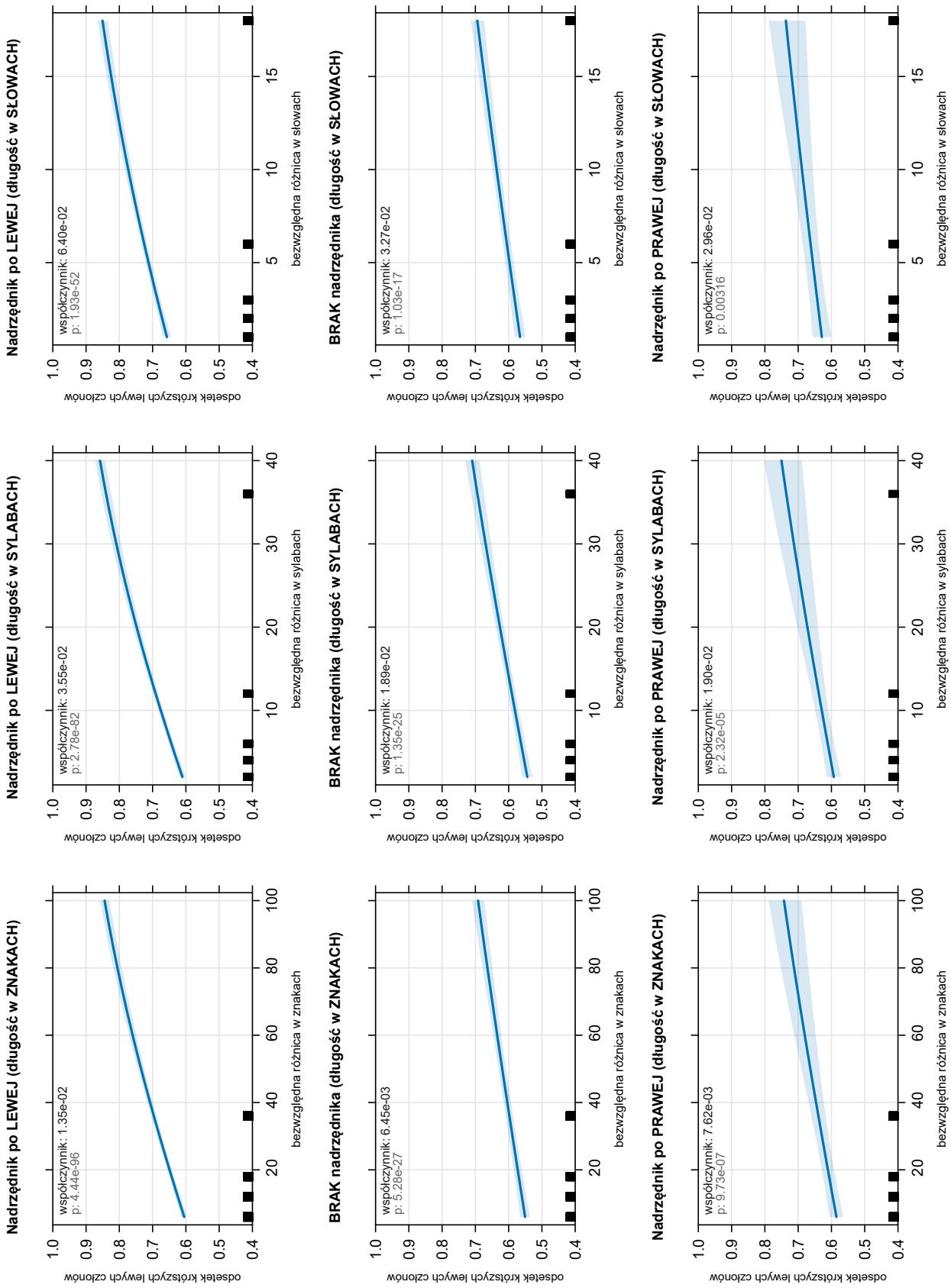
Rysunek D.3: Różnica długości członów a występowanie krótszego członu po lewej stronie – język hiszpański



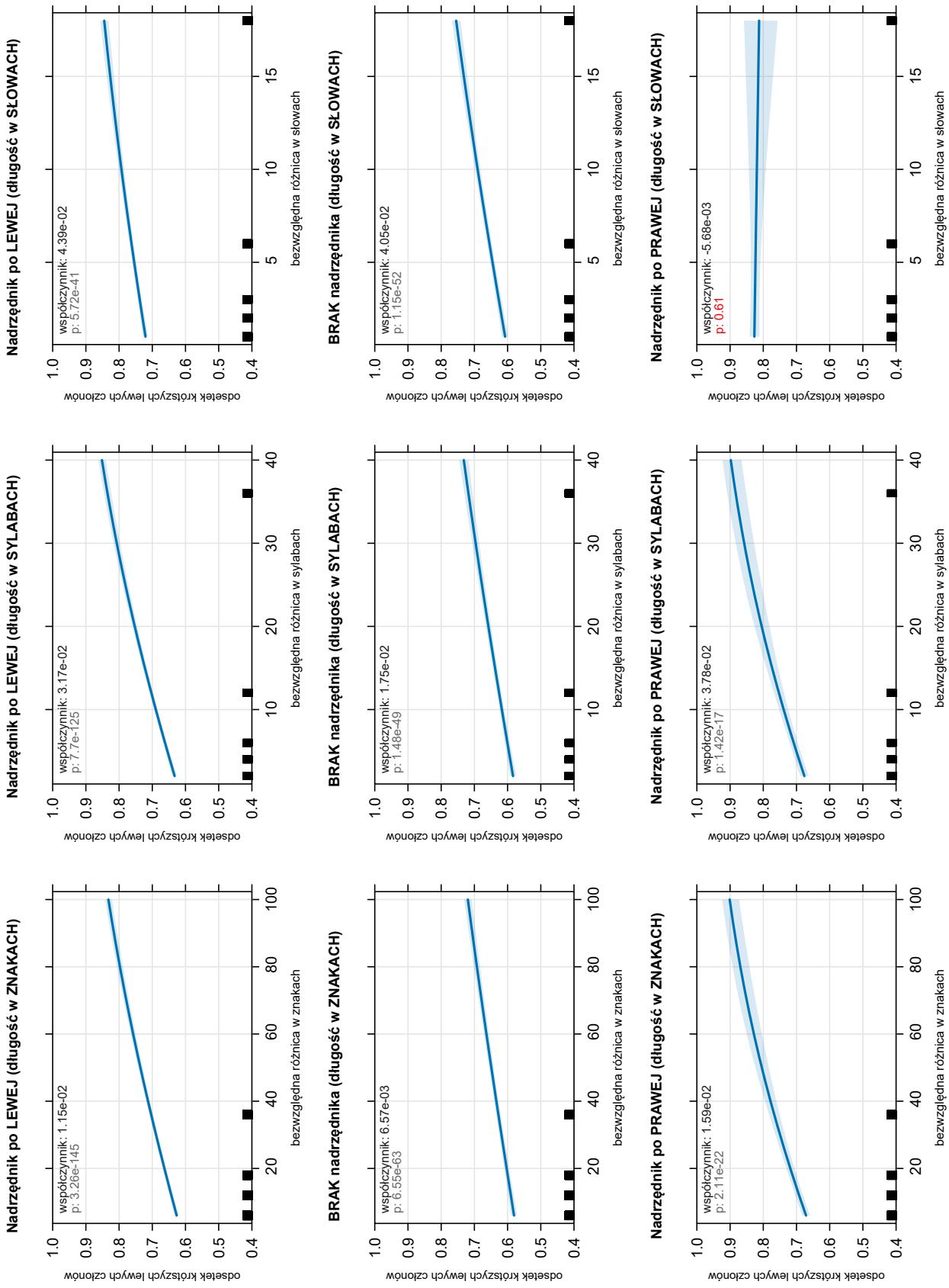
Rysunek D.4: Różnica długości członów a występowanie krótkiego członu po lewej stronie – język islandzki



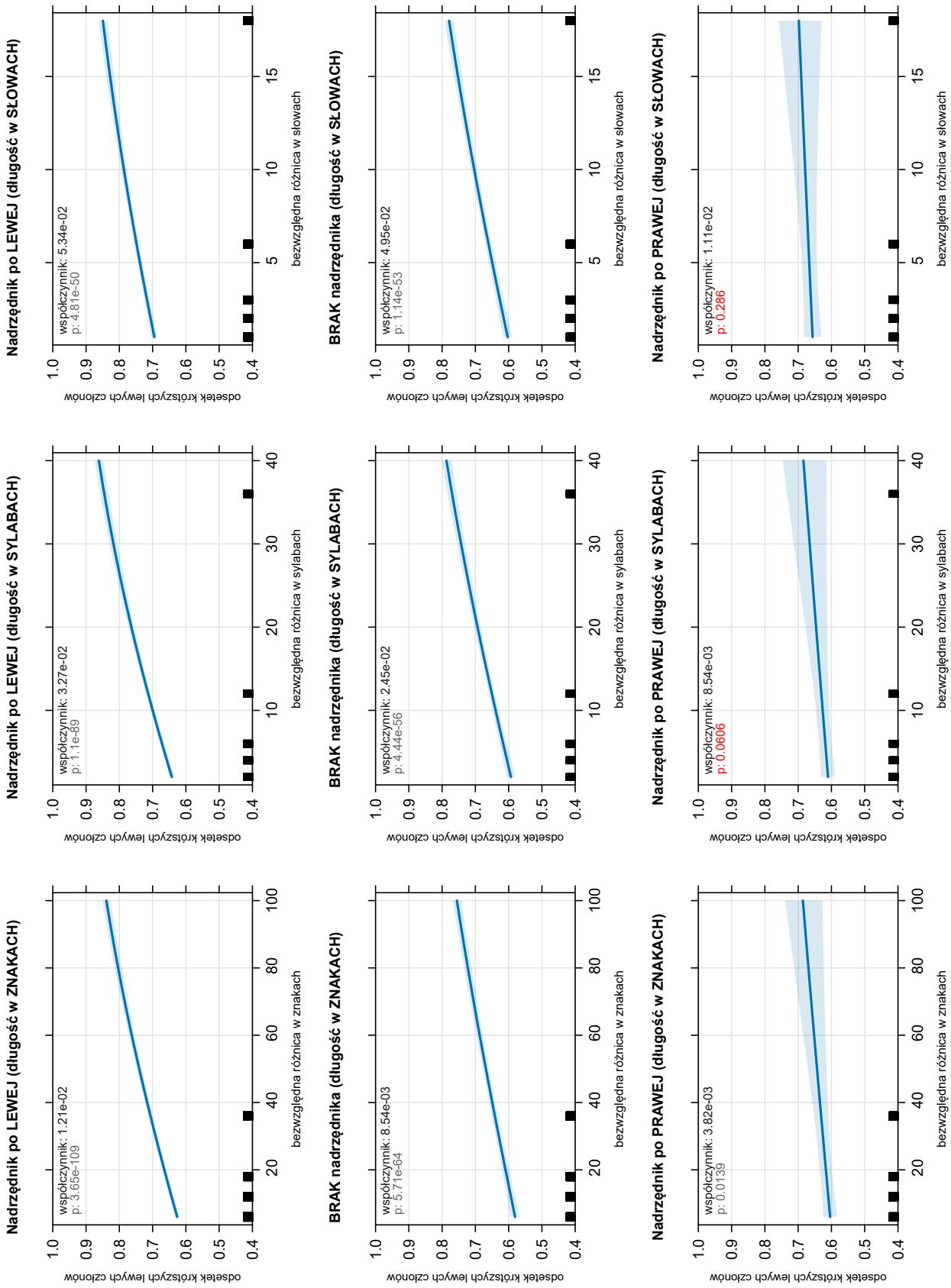
Rysunek D.5: Różnica długości członów a występowanie krótszego członu po lewej stronie – język polski



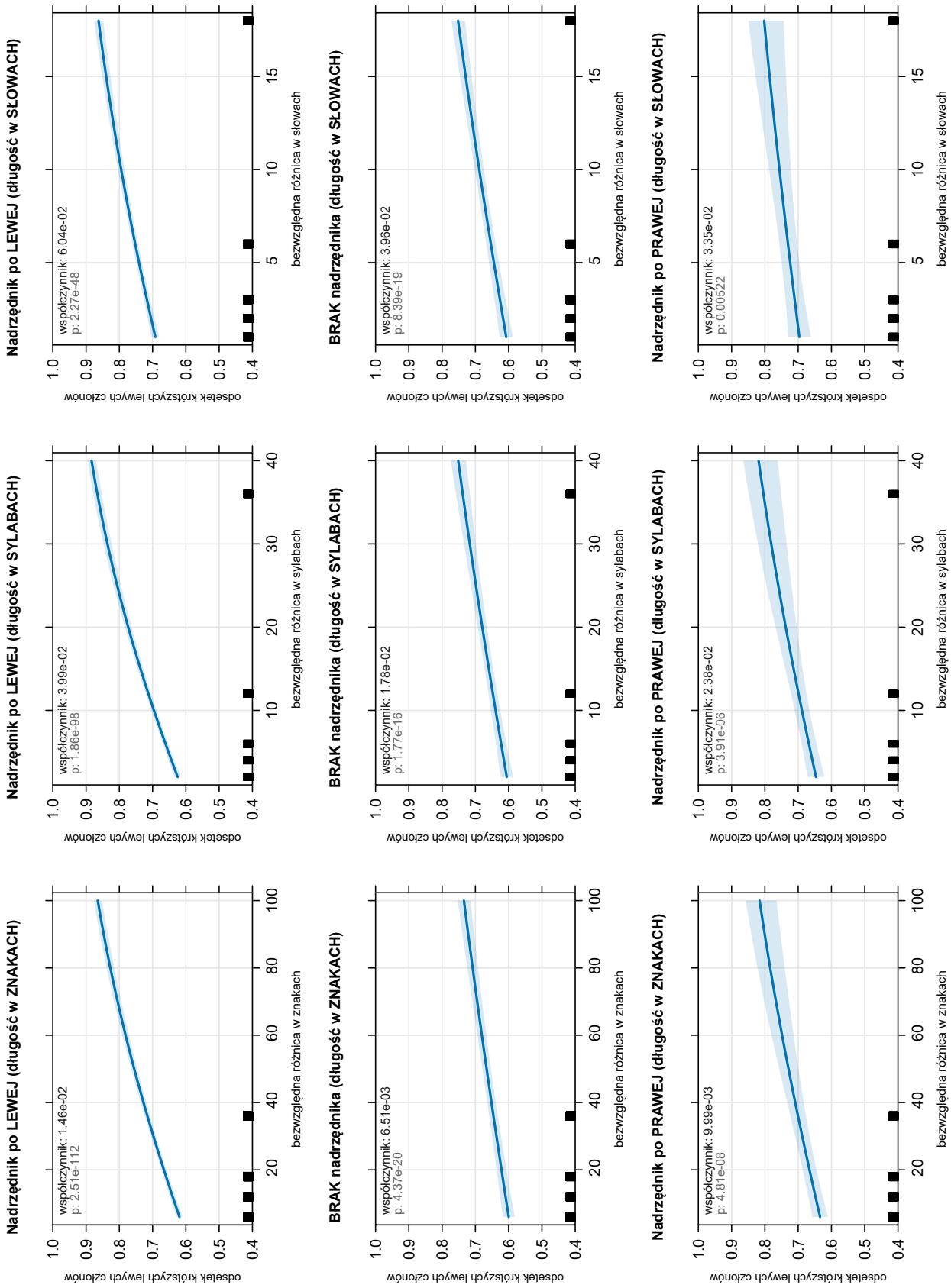
Rysunek D.6: Różnica długości członów a występowanie krótszego członu po lewej stronie – język portugalski



Rysunek D.7: Różnica długości członów a występowanie krótszego członu po lewej stronie – język rosyjski

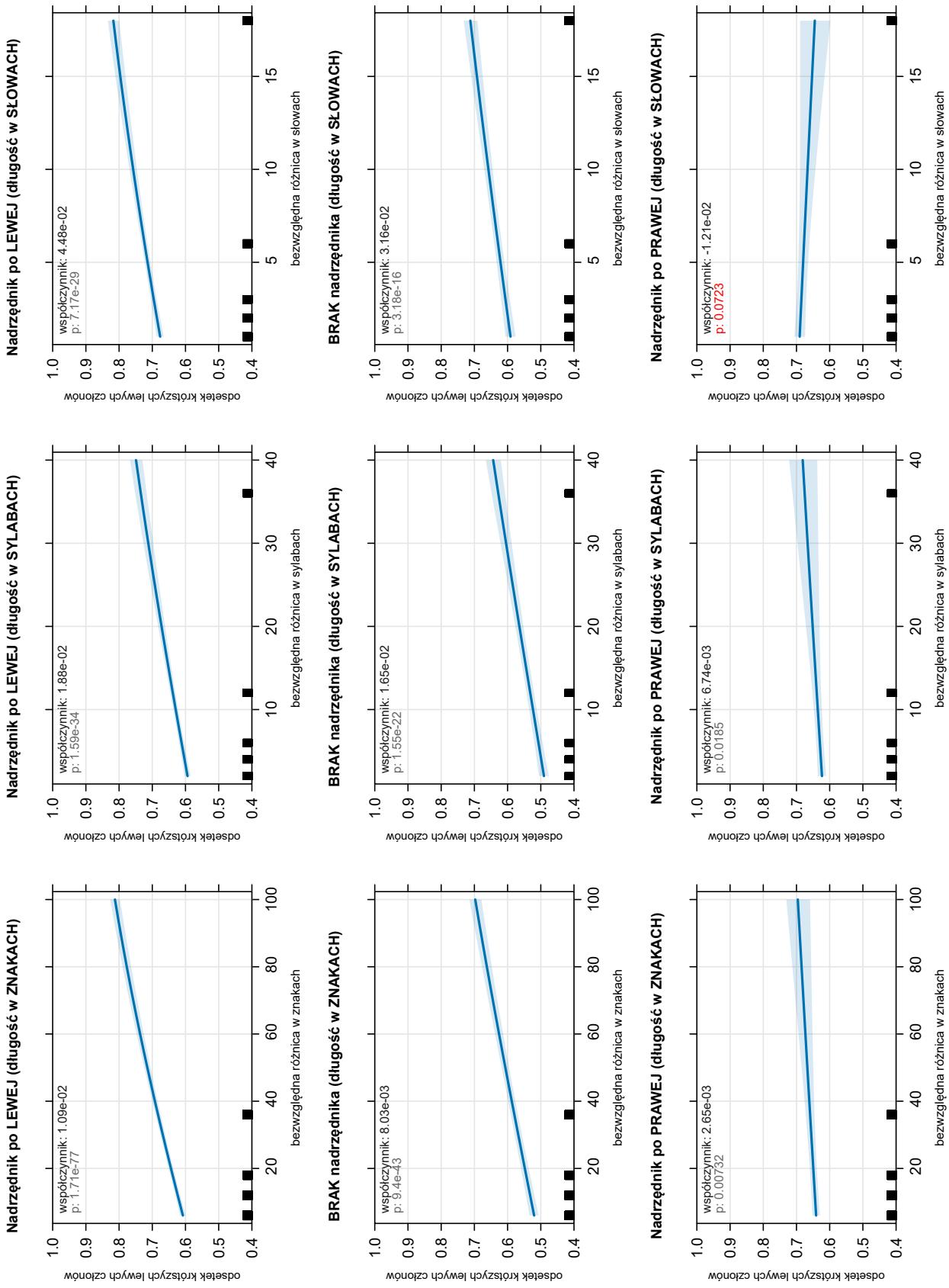


Rysunek D.8: Różnica długości członów a występowanie krótszego członu po lewej stronie – język rumuński

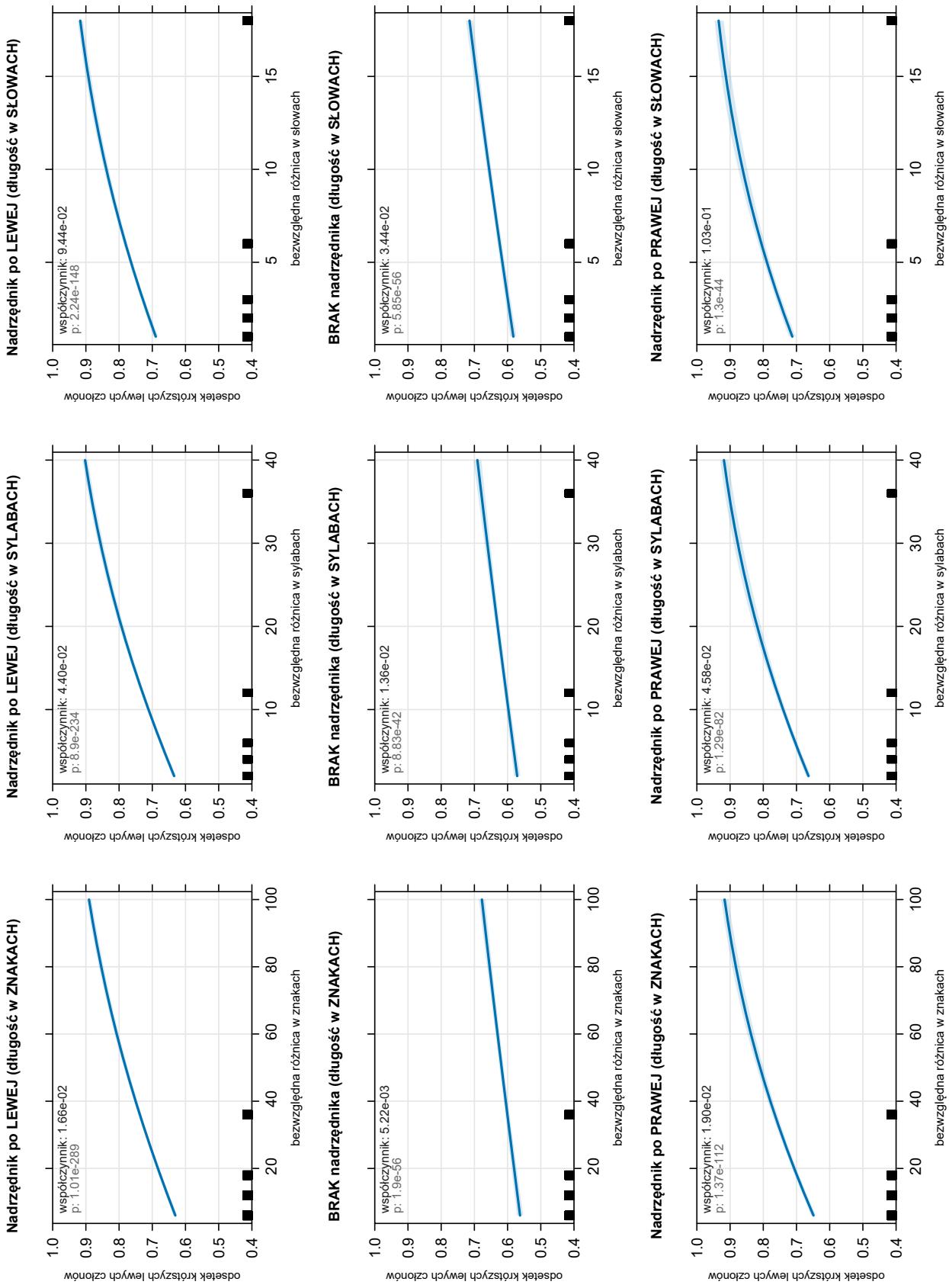


Rysunek D.9: Różnica długości członów a występowanie krótszego członu po lewej stronie – język włoski

D.2. Języki mieszane

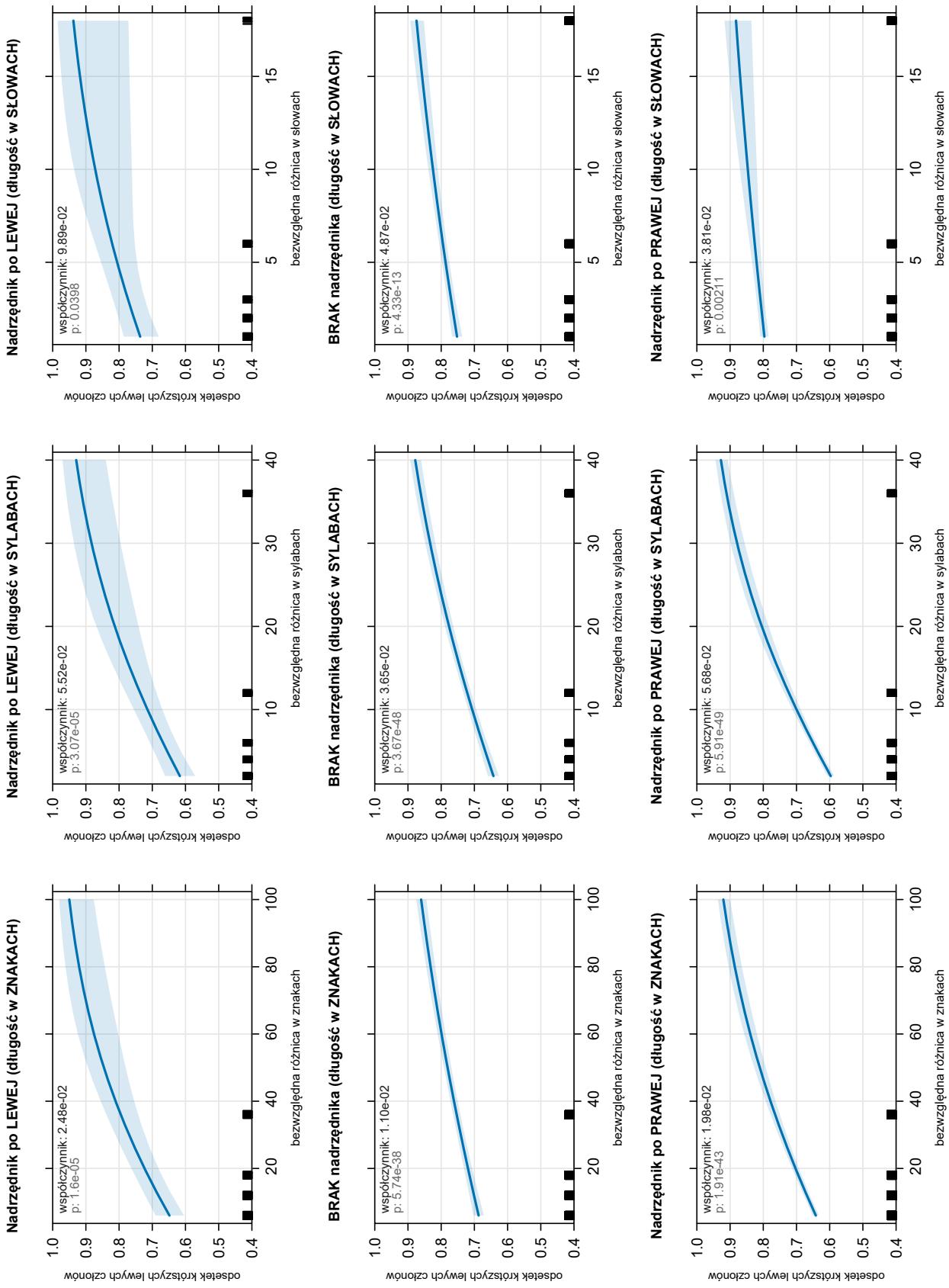


Rysunek D.10: Różnica długości członów a występowanie krótszego członu po lewej stronie – język łaciński

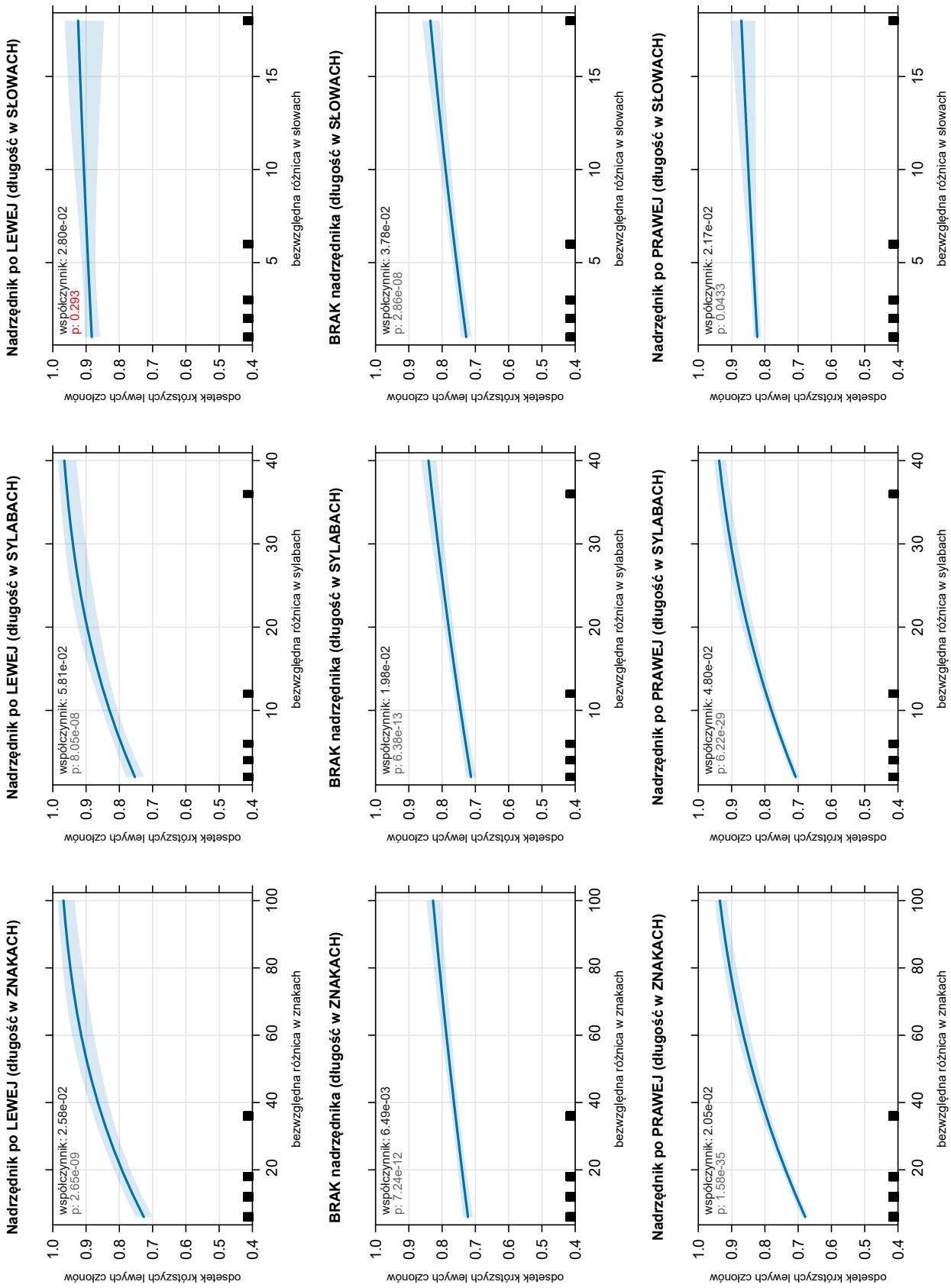


Rysunek D.11: Różnica długości członów a występowanie krótszego członu po lewej stronie – język niemiecki

D.3. Języki finalne



Rysunek D.12: Różnica długości członów a występowanie krótszego członu po lewej stronie – język koreański



Rysunek D.13: Różnica długości członów a występowanie krótszego członu po lewej stronie – język turecki

Bibliografia

- Bedir, T., Şahin, K., Güngör, O., Uskudarli, S., Özgür, A., Güngör, T., i Başaran, B. Ö. (2021). Overcoming the challenges in morphological annotation of Turkish in Universal Dependencies framework. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, strony 112–122.
- Borysiak, M. (2024). Dependency structure of English coordination: a surface-syntactic approach. Praca licencjacka na kierunku kognitywistyki (nieopublikowana). Uniwersytet Warszawski, Wydział Filozofii.
- Choi, J. D. i Palmer, M. (2011). Statistical dependency parsing in Korean: From corpus generation to automatic parsing. In *Proceedings of the second workshop on statistical parsing of morphologically rich languages*, strony 1–11.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.
- Cöltekin, C. (2010). A freely available morphological analyzer for Turkish. *LREC*, 2:19–28.
- De Marneffe, M.-C., Manning, C. D., Nivre, J., i Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, 47(2):255–308.
- Fenk-Oczlon, G. (1989). *Word frequency and word order in freezes*. Walter de Gruyter, Berlin/Nowy Jork.
- Futrell, R., Mahowald, K., i Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- Haspelmath, M. (2014). The Leipzig style rules for linguistics. *Max Planck Institute for Evolutionary Anthropology, Leipzig*, URL http://www.uni-regensburg.de/sprache-literatur-kultur/sprache-literatur-kultur/allgemeine-vergleichende-sprachwissenschaft/medien/pdfs/haspelmath_2014_style_rules_linguistics.pdf.

- Hoeksema, J. (1992). The head parameter in morphology and syntax. *Language and Cognition*, 2:119–132.
- Kanayama, H., Han, N.-R., Asahara, M., Hwang, J. D., Miyao, Y., Choi, J. D., i Matsumoto, Y. (2018). Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, strony 75–84.
- King, J. i Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5):580–602.
- Lohmann, A. (2014). *English coordinate constructions*. Cambridge University Press.
- McDonald, J. L., Bock, K., i Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, 25(2):188–230.
- Nivre, J. (2006). Constraints on non-projective dependency parsing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, strony 73–80.
- Patejuk, A. i Przepiórkowski, A. (2018). *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.
- Polinsky, M. (2012). Headedness, again. *UCLA Working Papers in Linguistics 17 (Theories of Everything)*, strony 348–359.
- Polinsky, M. i Magyar, L. (2020). Headedness and the lexicon: The case of verb-to-noun ratios. *Languages*, 5(1):9.
- Popel, M., Mareček, D., Štěpánek, J., Zeman, D., i Žabokrtský, Z. (2013). Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, strony 517–527.
- Przepiórkowski, A. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN.
- Przepiórkowski, A., Borysiak, M., i Głowacki, A. (2024). An argument for symmetric coordination from dependency length minimization: A replication study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, strony 1021–1033.
- Przepiórkowski, A. i Woźniak, M. (2023). Conjunct lengths in English, dependency length minimization, and dependency structure of coordination. In *Proceedings of*

the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), strony 15494–15512.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wiedeń, Austria.

Riedl, M. i Biemann, C. (2018). Using semantics for granularities of tokenization. *Computational Linguistics*, 44(3):483–524.

Simpson, G. B. i Kang, H. (2004). Syllable processing in alphabetic Korean. *Reading and Writing*, 17:137–151.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Türk, U., Atmaca, F., Özateş, Ş. B., Köksal, A., Başaran, B. Ö., Güngör, T., i Özgür, A. (2019). Turkish treebanking: Unifying and constructing efforts. In *Proceedings of the 13th linguistic annotation workshop*, strony 166–177.

Van Nguyen, M., Lai, V. D., Veyseh, A. P. B., i Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. *arXiv preprint arXiv:2101.03289*.

Wright, S. K., Hay, J., i Bent, T. (2005). Ladies first? Phonology, frequency, and the naming conspiracy.

Zipf, G. K. (1946). The psychology of language. In *Encyclopedia of psychology*, strony 332–341. Philosophical Library.