

Uniwersytet Warszawski

Wydział Filozofii

Kamil Tomaszek

Nr albumu: 432044

Minimalizacja długości zależności
w strukturach współrzędnie złożonych:
badanie korpusowe na podstawie
Polish Dependency Bank

Praca licencjacka

na kierunku KOGNITYWISTYKA

Praca wykonana pod kierunkiem

prof. dr. hab. Adama Przepiórkowskiego

Uniwersytet Warszawski

Warszawa, czerwiec 2023

Streszczenie

Niniejsza praca licencjacka jest poświęcona analizie koordynacji w języku polskim. Ma ona charakter empiryczny i opiera się na danych pochodzących z korpusu Polish Dependency Bank. W pracy tej przedstawiam teorię zależności składniowej oraz teorię minimalizacji długości zależności między wyrazami, aby następnie potwierdzić hipotezę o istnieniu tendencji do umieszczania dłuższego członu koordynacji częściej ze strony prawej – niezależnie od pozycji nadrzędnika koordynacji. Omawiam także wpływ pozycji i obecności nadrzędnika na tę tendencję, porównując wyniki własnej analizy z istniejącą już literaturą. Wyjaśniam jak te wyniki, wraz z minimalizacją długości zależności, wpływają na różne reprezentacje struktury koordynacji w teorii zależności składniowej.

Słowa kluczowe

koordynacja, minimalizacja długości zależności, Polish Dependency Bank, drzewa zależnościowe, korpusy językowe

Tytuł pracy w języku angielskim

Dependency Length Minimization in coordinate structures: A corpus study based on Polish Dependency Bank

Spis treści

1. Wstęp	4
1.1. Motywacja i cel pracy	4
1.2. Zakres i struktura pracy	5
2. Podstawy teoretyczne	6
2.1. Koordynacja w języku polskim	6
2.2. Zarys teorii zależności składniowej	7
2.3. Minimalizacja długości zależności	9
2.4. Różne reprezentacje koordynacji	12
2.5. Hipotezy	16
3. Dane	17
3.1. Polish Dependency Bank	17
3.2. Preprocessing danych	18
3.3. Dane po preprocessingu	20
4. Analiza statystyczna	22
4.1. Podstawowa analiza	22
4.2. Dalsza analiza	23
5. Dyskusja wyników	27
5.1. Podsumowanie wyników badań	27
5.2. Interpretacja wyników	28
5.3. Przegląd literatury	28
5.4. Perspektywy dalszych badań	29
Bibliografia	30
Załączniki	34

Rozdział 1

Wstęp

W niniejszym rozdziale przedstawiam motywację i cel pracy, a także omawiam jej zakres oraz strukturę.

1.1. Motywacja i cel pracy

W pracy tej analizuję zjawisko minimalizacji długości zależności (DLM; ang. *Dependency Length Minimization*), czyli tendencji do szeregowania elementów wypowiedzi w sposób taki, by zmniejszyć sumę długości wszystkich zależności między wyrazami. Zależność międzywyrazowa oznacza, że jeden wyraz jest nadrzędny wobec innego. W przykładzie (1) wyraz *brata* jest wyrazem nadrzędnym wobec wyrazów *śmiesznego*, *młodszego* oraz *jej*, a długości zależności między nadrzędnikiem, a tymi trzema podrzędnikami to odpowiednio 2, 1 oraz 3 (mierzone, licząc odległości w słowach).

- (1) *Widziałem [Asię i jej śmiesznego, młodszego brata].*

Interesuje mnie, jak DLM wpływa na koordynację w języku polskim. Koordynacja to zjawisko, w którym dwa lub więcej równorzędnych elementów łączy się spójnikiem w większą strukturę o tej samej funkcji co poszczególne jej człony. Przykładem koordynacji jest (1), gdzie jej nadrzędnikiem jest słowo *widziałem*, a członami są *Asię* oraz *jej śmiesznego, młodszego brata*. Człony te złączone są spójnikiem *i*.

W pracy badam dwie hipotezy dotyczące długości członów w koordynacjach w języku polskim: 1. że dłuższy człon koordynacji jest częściej ze strony prawej i 2. że pozycja nadrzędnika wpływa na rozkład długości członów koordynacji.

Długości członów mierzę na cztery różne sposoby, licząc znaki, sylaby, tokeny¹ oraz słowa. W przykładzie (1) odpowiednie wartości wynosiłyby 4 vs 31 znaków, 2 vs 9 sylab, 1 vs 5 tokenów, 1 vs 4 słów. Niewiele wysiłku zajmuje pokazanie, że pierwsza z

¹Do tokenów zalicza się całe słowa (np. *być*, *kolor*), pewne części słów (między innymi wyrazy po oderwaniu pewnych końcówek fleksyjnych oraz same te końcówki, (np. *zrobił*, *em*)), a także interpunkcję (np. *., - , ?*).

hipotez jest prawdziwa. Następnie przechodzę do omówienia wpływu obecności i pozycji nadrzędnika oraz różnicy długości między analizowanymi członami na proporcje danych, w których prawy człon jest dłuższy. Praca ta ma charakter empiryczny, opiera się na danych pochodzących z Polish Dependency Bank (PDB), czyli korpusu języka polskiego zawierającego ponad 22 tysiące drzew zależnościowych, oraz na wcześniejszej pracy badającej te same zależności, ale dla języka angielskiego (Przepiórkowski i Woźniak, 2023).

1.2. Zakres i struktura pracy

Praca składa się z pięciu rozdziałów. Niniejszy rozdział jest rozdziałem pierwszym. W rozdziale drugim omawiam teoretyczne podstawy pracy, tj. przedstawiam czym jest koordynacja, prezentuję zarys teorii zależności składniowej, opisuję teorię minimalizacji zależności oraz wskazuję różne reprezentacje zależnościowe wraz z ich przewidywaniami. W rozdziale trzecim opisuję źródło danych, czyli Polish Dependency Bank, jak i ich preprocessing – działanie algorytmu, napisanego w języku Python, wybierającego koordynacje oraz informacje o nich z PDB, a także pokazuję format danych po preprocessingu. W rozdziale czwartym dokładniej opisuję hipotezy badawcze, ich testowanie wraz z analizami statystycznymi w języku R. W rozdziale piątym omawiam wyniki badań i ich interpretację w kontekście dotychczasowej literatury naukowej, a także podsumowuję pracę oraz proponuję perspektywy dalszych badań.

Rozdział 2

Podstawy teoretyczne

W niniejszym rozdziale omawiam teoretyczne podstawy pracy, tj. opisuję czym jest koordynacja, przedstawiam zarys teorii zależności składniowej, a także prezentuję teorię minimalizację długości zależności oraz różne reprezentacje zależnościowe wraz z ich przewidywaniami.

2.1. Koordynacja w języku polskim

Słowo koordynacja wywodzi się z łacińskiego wyrazu *coordinatio*, które składa się z przedrostka *co-* (wspólny, zgodny) i sufiksu *-ordinatio* (rządzenie, uporządkowanie). W lingwistyce pojęcie koordynacja jest używane do opisu zjawiska związanego z łączeniem równorzędnych elementów językowych w większe całości. Jest ono również znane pod nazwą struktura współrzędnie złożona. Według definicji Oxford Bibliographies² koordynacja to zjawisko, w którym dwa lub więcej elementów (nazywanych w tej pracy członami) jest ze sobą połączonych przy użyciu spójnika, np. *i*, w jeden, większy element. W przeciwieństwie do relacji podrzędnej, w której jeden element jest asymetryczny względem drugiego, koordynacja pod wieloma względami jest symetryczna – dlatego nazywamy ją strukturą współrzędną. Wszystkie jej człony należą zwykle do tej samej kategorii gramatycznej, posiadają zazwyczaj te same funkcje składniowe, a każdy z nich może pojawić się samodzielnie na tym miejscu w zdaniu. Koordynacja jednak zachowuje się czasem niesymetrycznie, mając dwa człony z różnymi funkcjami składniowymi³. Istnieją także rodzaje zdań, które są mocno związane z koordynacją – np. pomijanie (ang. *gapping*), gdzie zdanie składa się z dwóch połączonych zdań, jednak drugie nie ma czasownika (zob. (2a)), oraz podnoszenie prawego węzła (ang. *right node raising*), gdzie zdanie tworzą dwa zdania z tym samym elementem końcowym, więc jest on pomijany w tym pierwszym (zob. (2b)).

²<https://www.oxfordbibliographies.com/display/document/obo-9780199772810/obo-9780199772810-0128.xml>, dostęp z dn. 7.04.2023

³Przykładowym wyjątkiem od reguły posiadania tej samej funkcji składniowej jest *Kto i kogo kopnął?*, gdzie *kto* jest podmiotem, a *kogo* – dopełnieniem bliższym.

- (2) a. Łucja gra na pianinie, a Łukasz na gitarze.
- b. Laura kupuje, a Kuba sprzedaje stare książki.

Koordinacja jest jednym z podstawowych sposobów łączenia słów (zob. (3a)), fraz (zob. (3b)), czy zdań (zob. (3c)). Jak określa Wikipedia, każda kategoria leksykalna lub frazowa może być skoordynowana⁴.

- (3) a. [Ania **i** Julia] *idą* na spacer.
- b. [Wesoła Marysia **oraz** smutny Janek] *wybrali się* do parku.
- c. [Kuba zjadł obiad **a** Marysia poszła spać].

W przykładach prezentowanych w niniejszej pracy koordynacje otoczone są nawiasami kwadratowymi. Człony koordynacji nazywamy koniunktami, to co je łączy – spójnikiem współrzędnym (w przykładach ilustrowany pogrubionym tekstem), a wyraz nadrzędny względem obu członów – nadrzędnikiem koordynacji (w przykładach ilustrowany kursywą). Jak widać w (3c), nie zawsze istnieje nadrzędnik koordynacji. W powyższych przykładach koniunktami są: (3a) – *Ania, Julia*; (3b) – *Wesoła Marysia, smutny Janek*; (3c) – *Kuba zjadł obiad, Marysia poszła spać*.

Ze względów semantycznych zwykle wyróżnia się cztery rodzaje koordynacji: koordynacje koniunkcyjne (4a), koordynacje dysjunkcyjne (4b), koordynacje adwersatywne (4c) oraz koordynacje kauzalne (4d) (Haspelmath, 2007).

- (4) a. Marta *zjadła* [jabłko **i** gruszkę].
- b. Ona miała [szesnaście **lub** siedemnaście] *lat*.
- c. *Byli* [ładni, **ale** głupi].
- d. [Nie zrobiłem pracy domowej, **bo** nie chciałem].

Do łączenia koniunktów używają one różnych zestawów spójników. W koordynacjach koniunkcyjnych człony łączy się między innymi spójnikami *i*, *oraz*, *ani*, *tudzież*, *również*, a w koordynacjach dysjunkcyjnych – *albo*, *bądź* , *lub*, *czy*, *lecz* w obu tych kategoriach wykorzystywana jest także interpunkcja. W koordynacjach adwersatywnych używane są między innymi spójniki *ale*, *lecz*, *zaś*, *natomiast*, *jednak*, a w koordynacjach kauzalnych – *bo*, *bowiem*. W tej pracy skupię się na pierwszych trzech rodzajach koordynacji, jako że struktury z wyrazami *bo* i *bowiem* są w PDB oznaczone jako struktury podrzędne.

2.2. Zarys teorii zależności składniowej

De Marneffe i Nivre (2019) oraz Pedersen i in. (2004) zwracają uwagę na to, że teoria zależności składniowej ma długą i bogatą historię, która sięga aż starożytności.

⁴[https://en.wikipedia.org/wiki/Coordination_\(linguistics\)](https://en.wikipedia.org/wiki/Coordination_(linguistics)), dostęp z dn. 07.04.2023

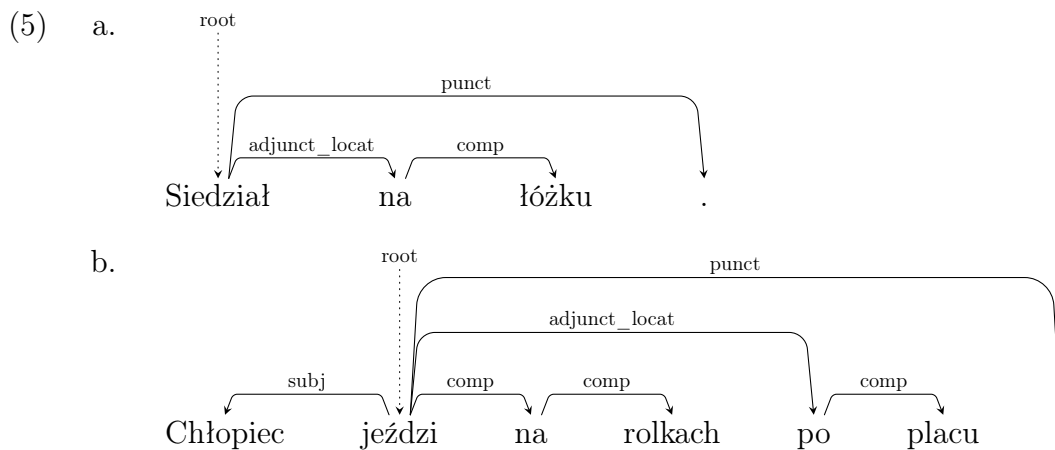
Pierwsze ślady tego podejścia można znaleźć w gramatyce sanskrytu Pāṇiniego, czy w pracach wczesnych arabskich gramatyków (Kruijff, 2002), a także w niektórych teoriach gramatycznych średniowiecza (Covington, 1984).

Tesnière (1959) podjął pierwszą próbę stworzenia kompleksowej teorii gramatyki, w której wszystko byłoby oparte na zależnościach. Przedstawił on ją, jako mającą duży potencjał do uchwycenia podobieństw, jak i różnic między językami. Wróblewska (2014) opisuje, że podstawowymi założeniami teorii Tèsnierè’a było występowanie *połączeń* (fr. *connexions*) oraz *walencji* (fr. *valence*). *Połączenia* obecnie określa się zależnościami i są one jednymi z podstawowych relacji zachodzących w składni. Łączą one dwa wyrazy współwystępujące w zdaniu i prezentują ich zależność w drzewie składniowym, któremu u Tesnière’a odpowiada *stemma*. Jeden z połączonych wyrazów określa się mianem nadrzędnika, wyrazu nadrzędnego (u Tesnière’a *terme supérieur*), a drugie – podrzędnika, wyrazu zależnego (u Tesnière’a *terme inférieur*). Relacja ta jest zawsze jednostronna, nie jest symetryczna. Teoria walencji zakłada, że w centrum zdania jest czasownik, który wymaga pewnych argumentów (u Tesnière’a *actants*), ale także mogą się przy nim znaleźć dodatkowe, niewymagane modyfikatory (u Tesnière’a *circonstants*). Przy czym czasownik z modyfikatorami połączony jest jedynie relacją zależności, natomiast tylko z argumentami jest połączony zarówno zależnością, jak i walencją. W praktyce oznacza to, że czasowniki mogą wymagać wystąpienia jakichś argumentów obok nich, np. we frazie *kupić <coś>* wyraz *kupić* nie może wystąpić sam, co znaczy, że jest on przynajmniej uniwalentny. Tak samo czasowniki mogą być biwalentne, triwalentne itd. Przepiórkowski (2017) argumentuje, że rozróżnienie podrzędników na argumenty i modyfikatory jest niepotrzebne.

Jak pisze Wróblewska (2014), istnienie *połączeń* oraz *walencji* zostało ogólnie przyjęte przez teoretyków teorii zależności. W XX wieku teoria ta mocno rozwinęła się zwłaszcza w lingwistyce klasycznej i słowiańskiej (Mel’čuk, 1988). Obecnie mówi się o kilku rodzajach reprezentacji zależności – semantycznych, morfologicznych, prozodycznych, syntaktycznych⁵, jednak w tej pracy skupiam się tylko na reprezentacji uwzględniającej czynniki morfoskładniowe oraz wymagania przez człon główny określonej formy członu zależnego.

Drzewo zależnościowe składa się z węzłów i krawędzi (graficznych reprezentacji zależności). Węzły reprezentują wyrazy w zdaniu, a krawędzie – zależności między nimi. Korzeń jest węzłem, który nie ma nadrzędnika, czyli nie jest w relacji podrzędności z żadnym z innych elementów. Zwykle uznaje się, że w zdaniu nie może być więcej niż jeden korzeń, a z korzenia da się przejść po strzałkach do każdej innej części zdania. Strzałki krawędzi są skierowane zawsze od wyrazu nadrzędnego do wyrazu podrzędnego.

⁵https://en.wikipedia.org/wiki/Dependency_grammar, dostęp z dn. 07.04.2023



Przykładowe drzewa zależnościowe z korpusu PDB

Aby odróżnić od siebie różne zależności, krawędzie mogą być etykietowane, często funkcjami gramatycznymi, jak w przykładach (5a–b). Oto objaśnienia użytych etykiet:

- *root* – korzeń zdania
- *subj* – podmiot (jeden z argumentów) zdania
- *comp* – inny argument
- *adjunct_locat* – modyfikator miejsca
- *punct* – znak interpunkcyjny

Teoria zależności składniowej jest popularnym podejściem w dziedzinie przetwarzania języka naturalnego, ponieważ umożliwia łatwe i precyzyjne analizowanie struktury zdania. Ma ona wiele zastosowań, np. w dziedzinach takich jak tłumaczenie maszynowe (ang. *Machine Translation*) czy analiza sentymentu (ang. *Sentiment Analysis*), ponieważ ułatwia przetwarzanie i rozumienie znaczenia zdań. W ostatnich latach powstały projekty takie jak Universal Dependencies (<https://universaldependencies.org/>), które mają na celu zunifikowanie reprezentacji lingwistycznych (w tym wypadku: morfosyntaktycznej i składniowej) dla różnych języków. Dla języka polskiego stworzono już kilka korpusów zgodnych z tym standardem (Przepiórkowski i Patejuk, 2020; Wróblewska, 2020) oraz cały czas powstają nowe, także dla innych języków.

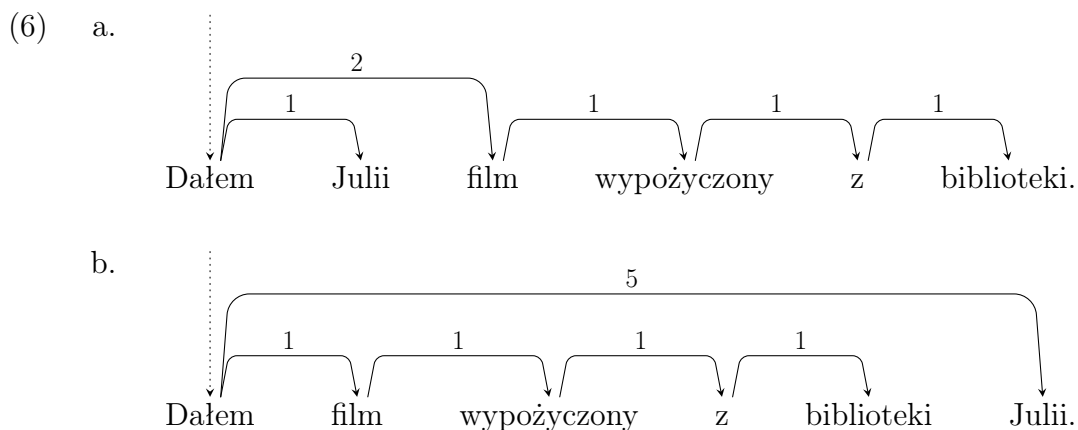
2.3. Minimalizacja długości zależności

Minimalizacja długości zależności (DLM – ang. *Dependency Length Minimization*) to zasada, według której języki naturalne dążą do zmniejszania odległości między słowami, które są od siebie zależne syntaktycznie. Ułatwia to przetwarzanie informacji i redukuje obciążenie pamięci roboczej. Zasada ta jest odnotowywana w lingwistyce

już od długiego czasu i pozwala nam na bardziej efektywne analizowanie i generowanie języka naturalnego.

Jednym ze sposobów na badanie DLM jest tworzenie sztucznych języków losowych oraz porównywanie długości zależności w tych językach z długościami zależności w językach naturalnych. Futrell i in. (2015) przedstawili wyniki badań na dużym korpusie tekstów z 37 języków, w których zmierzili średnią długość zależności w zdaniach. Długość zależności definiowali jako liczbę słów między słowem nadrzędnym a podrzędnym w drzewie składniowym zdania. Porównali średnie długości zależności w tekstach naturalnych z długością zależności w tekstach losowo przestawionych i stwierdzili, że we wszystkich badanych językach długość zależności w tekstach naturalnych była znacząco mniejsza niż w tekstach losowych, co świadczy o uniwersalnej tendencji do minimalizacji długości zależności. Zauważyli również, że różne języki mają różne strategie minimalizacji długości. Wnioskowali, że minimalizacja długości zależności jest wspólną cechą języków naturalnych i wynika z ograniczeń pamięci roboczej ludzkiego mózgu.

W przykładach (6a–b) możemy zauważyć, że zgodnie z DLM zdanie (6a) jest bardziej naturalne, ponieważ suma długości wszystkich zależności wynosi 6, podczas gdy w (6b) wynosi ona 9 (dla uproszczenia pominąłem zależność między korzeniem zdania, a kropką na jego końcu; wliczając ją obie wartości byłyby większe o 6).



Przepiórkowski i Woźniak (2023) zwracają uwagę, że Futrell i in. (2020) oraz Hawkins (1994) twierdzą, że występowanie DLM można rozróżnić na poziom gramatyczny (*grammar*), jak i użycie języka (*use*). Hawkins (1994) wskazuje, że na poziomie gramatyki, pewne skonwencjonalizowane szyki składników okazują się minimalizować średnią długość zależności. Jako przykład podaje sytuację w języku angielskim, gdy NP oraz PP są zależne od V⁶. Wtedy szyk V-NP-PP miałby średnio krótszą długość zależności, niż V-PP-NP, jako że frazy rzeczownikowe są w języku angielskim średnio krótsze niż frazy przyimkowe. Jak dodają Przepiórkowski i Woźniak (2023), Hawkins (1994) argumentuje, że tendencja ta jest skonwencjonalizowana – występuje w gramatyce, a nie

⁶Wyjaśnienia użytych skrótów: V – czasownik (ang. *verb*), NP – fraza rzeczownikowa (ang. *nominal phrase*), PP – fraza przyimkowa (ang. *prepositional phrase*).

tylko w użyciu. Jako powód wskazuje, że w języku angielskim szyk V-NP-PP występuje częściej niż V-PP-NP nie tylko gdy NP jest krótsze od PP, ale i wtedy, gdy są podobnej długości – ilustrują to przykłady (7a–b), gdzie zdanie (7a) z szykiem V-NP-PP jest bardziej naturalne niż zdanie (7b) z szykiem V-PP-NP, mimo ich podobnej długości. Gdy jednak wydłużymy NP, to szyk V-PP-NP (zob. (7c)) staje się bardziej naturalny, co znów jest zgodne z hipotezą DLM, ale już na poziomie użycia.

- (7) a. I gave <a book> <to John> .
 Dałem⁷ <książkę> <Johnowi> .
- b. I gave <to John> <a book> .
 Dałem <Johnowi> <książkę> .
- c. I gave <to John> <the most interesting book I've read in years>
 Dałem <Johnowi> <najbardziej interesującą książkę, jaką przeczy-
 tałem od lat> .

DLM jest również powiązana z innymi właściwościami języków naturalnych, między innymi z pozycyjnością głowy. Głowa (centrum składniowe) frazy to jej główny element, który decyduje o jej kategorii gramatycznej i znaczeniu. Dopełnienie to element zależny od głowy, który uzupełnia jej znaczenie. Na przykład we frazie *jeść czerwone jabłko* czasownik *jeść* jest głową, a fraza rzeczownikowa *czerwone jabłko* – dopełnieniem. Głową frazy *czerwone jabłko* jest rzeczownik *jabłko*, a przymiotnik *czerwone* jest jej modyfikatorem. Pozycyjność głowy jest jednym z kryteriów klasyfikacji języków naturalnych i ma wpływ na ich strukturę syntaktyczną i semantyczną. Oznacza ona pozycję głowy względem reszty frazy. W zależności od pozycyjności głowy, języki można podzielić na inicjalne (*head-initial*) oraz finalne (*head-final*). Na przykład, w języku angielskim, który jest inicjalny, głowa frazy zazwyczaj (ale nie zawsze) znajduje się przed jej dopełnieniem (*eat a red apple*), natomiast w języku japońskim, który jest finalny, głowa zazwyczaj jest za dopełnieniem, a ta sama fraza zapisana by była jako *czerwone (a dokładniej: być czerwonym)_[NPAST] jabłko_[ACC] jeść_[NPAST] (aka-i ringo-o tabe-ru)*⁸.

Badania wykazały, że istnieje związek między pozycyjnością głowy a długością zależności, przy czym języki finalne mają średnio krótszą długość zależności niż języki inicjalne (Futrell i in., 2015).

DLM nie jest jedynym czynnikiem kształtującym strukturę syntaktyczną języków naturalnych. Istnieją również inne ograniczenia i preferencje, które mogą wpływać

⁷Frazy *I gave* oraz *I've read* można przetłumaczyć również jako odpowiednio *dałem* i *przeczytałam*, nie zawierają one informacji o rodzaju; dla uproszczenia wszystkie przykłady tłumaczę używając rodzaju męskiego.

⁸https://en.wikipedia.org/wiki/Head-directionality_parameter, dostęp z dn. 08.04.2023 oraz https://www.uni-bamberg.de/fileadmin/aspra/01_Studium/sample_termpaper_ma_generallinguistics.pdf, dostęp z dn. 19.04.2023

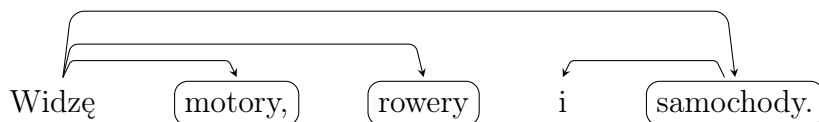
na kolejność słów i długość zależności, między innymi wskazana pozycyjność głowy, czy preferencje semantyczne. Niektóre z tych czynników mogą być sprzeczne lub komplementarne względem DLM. Dlatego DLM należy rozumieć nie jako jedyny, a jeden z wielu czynników wpływających na organizację języka naturalnego.

2.4. Różne reprezentacje koordynacji

Jeśli chodzi o przedstawienie drzew zależnościowych dla struktury koordynacji, to możemy wyróżnić 4 podstawowe podejścia, wraz z ich wariacjami (Popel i in., 2013; Przepiórkowski i Woźniak, 2023). Zilustrowane są one przykładami (8)–(11), stworzonymi na podstawie przykładowego zdania „Widzę motory, rowery i samochody”. Popel i in. (2013) wskazują na trudności związane z wyborem jednego podejścia oraz przedstawiają przegląd trzech rodzin modeli – nie znajduje się u nich model *londyński* wyróżniony w pracy Przepiórkowskiego i Woźniaka. Oto wszystkie 4 podejścia:

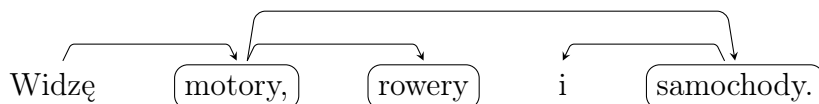
- **Podejście londyńskie** – jak wskazują Przepiórkowski i Woźniak (2023), podejście to nazwać możemy londyńskim, w duchu nazywania podejść od nazw miast, w których zostały one stworzone. Jest ono kojarzone z Word Grammar (Hudson, 1984, 1990, 2010). W angielskiej nomenklaturze możemy znaleźć je również pod nazwą *multi-headed*. Zakłada ono, że głowa każdego członu jest głową koordynacji, a zatem koordynacja posiada więcej niż jedną głowę.

(8)



- **Podejście stanfordzkie** – w angielskiej nomenklaturze określane także mianem *bouquet*. Jest ono używane w stanfordzkim parserze zależnościowym⁹ (de Marneffe i in., 2006). Zakłada ono, że głową koordynacji jest jej pierwszy człon, a reszta członów koordynacji jest od niego bezpośrednio zależna. Jego wariacją jest także model, w którym głową koordynacji jest jej ostatni człon. Spójnik zazwyczaj oznacza się jako zależny albo od jednego z dwóch otaczających go członów, albo od głowy koordynacji.

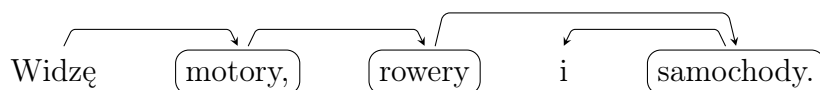
(9)



⁹<https://nlp.stanford.edu/software/lex-parser.shtml>

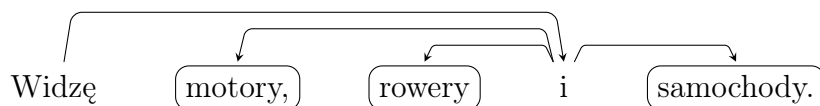
- **Podejście moskiewskie** – w angielskiej nomenklaturze spotkać się możemy również z określeniem *chain*. Jest używane w moskiewskim Meaning-Text Theory (Mel’čuk, 1974, 1988, 2009). Zakłada ono, że zależności w koordynacji są ustawione szeregowo, gdzie każdy człon jest zależny od poprzedniego. Głową koordynacji w tym przypadku jest jej pierwszy człon, a spójnik jest zależny od jednego z dwóch otaczających go członów. Jego wariacje obejmują modele, w których głową koordynacji jest jej ostatni człon i wtedy każdy człon jest zależny od tego następującego po nim, ale również takie, w których w skład szeregu wchodzi nie tylko człony, ale i spójniki.

(10)



- **Podejście praskie** – w angielskiej nomenklaturze znane również pod nazwą *conjunction-headed*. Jest ono używane w Prague Dependency Treebank (Hajič i in., 2006). Zakłada, że głową koordynacji jest jej spójnik i każdy z jej elementów jest zależny bezpośrednio od niego. To właśnie to podejście wykorzystywane jest w PDB.

(11)



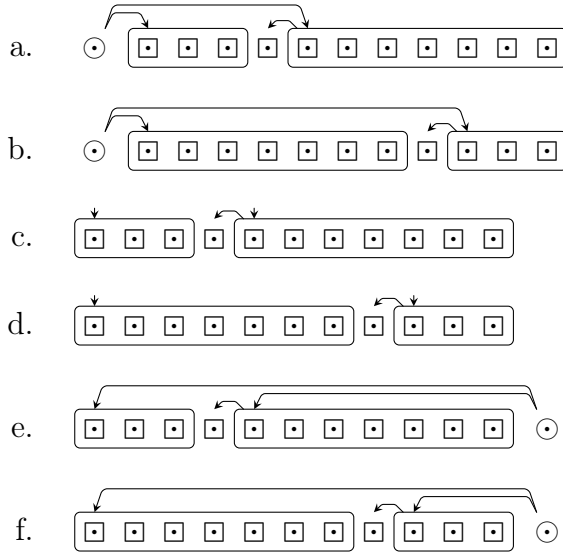
Aby opisać różnice między tymi podejściami w kontekście DLM, przedstawię założenie, które opisują Przepiórkowski i Woźniak (2023). Mówi ono, że w języku angielskim głowy wszystkich członów koordynacji są średnio umieszczone w tej samej – zwykle krótkiej – odległości od lewej krawędzi frazy. W przypadku PP, VP oraz CP¹⁰ zazwyczaj będzie to pierwsze słowo od lewej strony. W przypadku NP, przyjmując, że jego głową jest rzeczownik, średnio będzie to drugie słowo – zwykle rzeczownik jest poprzedzony przedimkiem.

W podejściu londyńskim, zakładając pozycję nadrzędnika z lewej strony (zob. (12a–b)), suma długości zależności jest zminimalizowana, gdy lewy człon jest krótszy.¹¹ Symetrycznie, gdy nadrzędnik jest z prawej strony (zob. (12e–f)), suma długości zależności jest zminimalizowana, gdy to prawy człon jest krótszy. Wartość zminimalizowanej sumy w obu przypadkach jest różna od wyższej sumy o różnicę długości członów. Gdy nadrzędnik nie występuje (zob. (12c–d)), suma długości nie zależy w ogóle od długości członów.

¹⁰Wyjaśnienia użytych skrótów: VP – fraza czasownikowa (ang. *verb phrase*); CP – fraza zdaniowa podrzędna (ang. *complementizer phrase*), np. *że on przyszedł*.

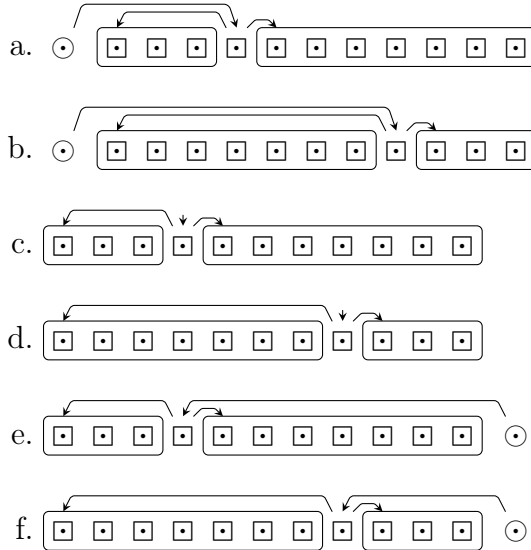
¹¹Rysunki (12)–(17) pochodzą z pracy Przepiórkowskiego i Woźniaka (2023).

(12) **Londyńskie:**



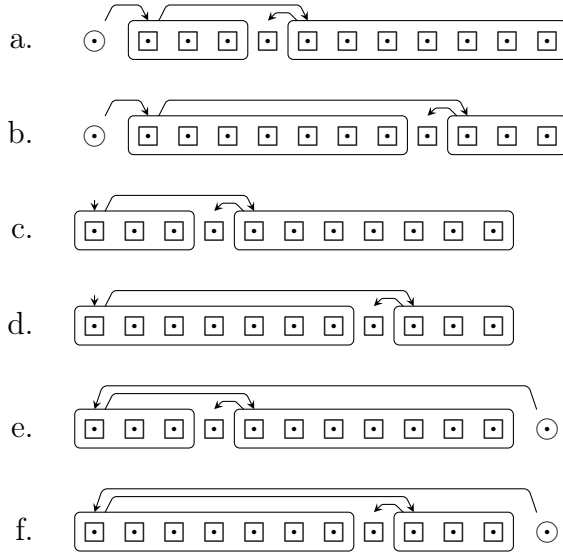
W podejściu praskim krótszy człon z lewej strony jest zgodny z hipotezą DLM, gdy nadrzędnik jest z lewej strony (zob. (13a–b)). Takie samo ustawienie członów minimalizuje sumę długości zależności, gdy nie ma nadrzędnika (zob. (13c–d)). W przypadku nadrzędnika z prawej strony (zob. (13e–f)), możemy zauważyć, że suma długości zależności nie zależy od tego, który człon będzie krótszy, a który dłuższy.

(13) **Praskie:**



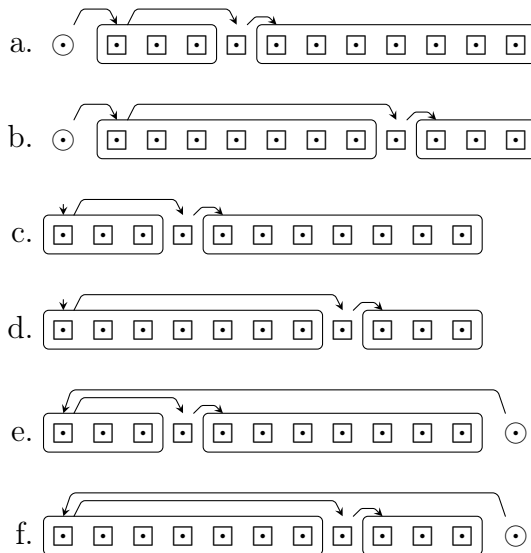
W standardowym (zakładającym, że głową jest pierwszy człon) podejściu stanfordzkim, jeśli krótszy człon będzie z lewej strony, minimalizuje to sumę długości zależności zarówno w przypadku, gdy nadrzędnik jest z lewej strony (zob. (14a–b)), w przypadku bez nadrzędnika (zob. (14c–d)), jak i wtedy, gdy jest z prawej (zob. (14e–f)). W każdym przypadku zminimalizowana suma jest różna od wyższej sumy o różnicę długości członów.

(14) **Stanfordzkie:**



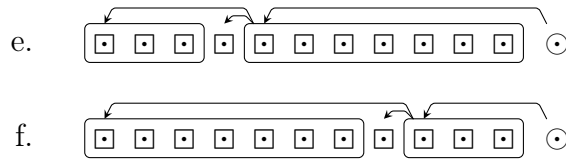
W standardowym (zakładającym, że głową jest pierwszy człon) podejściu moskiewskim, zależności są dokładnie takie same, jak w klasycznym podejściu stanfordzkim. Krótszy człon z lewej strony minimalizuje długość zależności zarówno gdy nadrzędnik jest z lewej (zob. (15a–b)), jak i z prawej strony (zob. (15e–f)), a także, gdy nadrzędnik nie występuje (zob. (15c–d)). Zminimalizowana suma jest różna od wyższej sumy o różnicę długości członów.

(15) **Moskiewskie:**



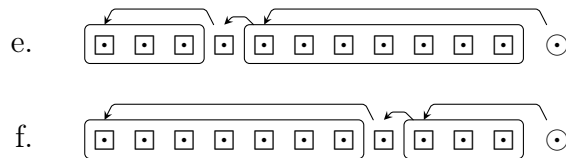
W wariacji podejścia stanfordzkiego zakładającej, że głową jest najbliższy nadrzędnikowi człon, przy pozycji nadrzędnika z lewej strony nic się nie zmienia – dalej zgodne z DLM będzie wystąpienie krótszego członu z lewej strony. W przypadku pozycji nadrzędnika z prawej strony (zob. (16e–f)), podobnie jak w podejściu praskim, suma długości zależności nie zależy w ogóle od tego, który człon będzie krótszy, a który dłuższy.

(16) **Stanfordzkie z głową bliżej nadrzędnika:**



W podejściu moskiewskim, zakładając, że to człon najbliższy nadrzędnika jest głową koordynacji, przewidywanie modelu jest dokładnie takie samo, jak w przypadku wyżej (zob. (17e–f)). Gdy nadrzędnik jest z prawej strony, to suma długości zależności nie zależy od sposobu ustawienia członów.

(17) **Moskiewskie z głową bliżej nadrzędnika:**



2.5. Hipotezy

Przepiórkowski i Woźniak (2023) pokazali, że w języku angielskim:

- 1) pierwszy (lewy) człon koordynacji jest w większości przypadków krótszy niż ostatni (prawy), niezależnie od pozycji nadrzędnika,
- 2) pozycja nadrzędnika istotnie wpływa na zmianę tej tendencji wraz ze zmianą różnicy w długości członów koordynacji. Kiedy nadrzędnik jest z lewej strony, bądź go nie ma, tendencja do umieszczania krótszego członu na pierwszej pozycji rośnie wraz ze wzrostem modułu (wartości bezwzględnej) z różnicy długości pierwszego i ostatniego członu. Gdy nadrzędnik jest z prawej strony, efektu takiego już nie ma.

Spodziewam się, że wyniki dla języka polskiego będą podobne, tj. pierwszy człon koordynacji będzie częściej krótszy od ostatniego członu oraz proporcja ta będzie rosła wraz ze wzrostem modułu z różnicy długości między członami – ale tylko gdy nadrzędnik nie występuje lub znajduje się po lewej stronie. Gdy nadrzędnik jest po prawej stronie, spodziewam się, że wpływ różnicy długości członów na proporcję członów będzie istotnie mniejszy lub nie będzie go wcale.

Rozdział 3

Dane

W tym rozdziale przedstawiam korpus będący źródłem danych użytych w mojej pracy, kryteria ich wyodrębniania oraz sposób ich przygotowania do analizy statystycznej.

3.1. Polish Dependency Bank

Polish Dependency Bank (PDB; Wróblewska, 2014) to jeden z największych korpusów języka polskiego zawierających drzewa zależnościowe. Wróblewska (2020) opisuje, że zdania w PDB pochodzą z wielu różnych źródeł, którymi są: (1) NKJP1M¹², (2) równoległe korpusy polsko-angielskie: *Europarl* (Koehn, 2005), *Pelcra Parallel Corpus* (Pęzik i in., 2011), *DGT-Translation Memory* (Steinberger i in., 2012), *OPUS* (Tiedemann, 2012), (3) *CDSCorpus* (Wróblewska i Krasnowska-Kieraś, 2017) i (4) nowoczesna literatura i korpus NKJP z wyłączeniem NKJP1M. Wróblewska (2020) przedstawia także zawartość PDB – składa się on z ponad 22 tysięcy drzew zależnościowych (350 tysięcy tokenów). Zdanie z tego korpusu posiada średnio 15,8 tokenów. 34% wszystkich zdań ma długość od 1 do 10 tokenów, 42% – między 11, a 20 tokenów, a 24% – powyżej 20 tokenów. Wszystkie drzewa zależnościowe w PDB były ręcznie anotowane.

Dane z PDB zostały umieszczone w 9 plikach. Sam korpus został podzielony na 3 części – *train*, *dev* oraz *test*¹³. Każda z tych części znajduje się w 3 oddzielnych plikach – jeden z nich, z rozszerzeniem ‘.txt’, to zbiór wszystkich zdań w danej części korpusu, drugi to zbiór tych samych zdań, ale już podzielonych na tokeny oraz z zaznaczeniem zależności, jest on w formacie ‘.conll’ i na jego podstawie można wyświetlić zdania te jako drzewa zależnościowe, a trzeci plik to zbiór metadanych o tych zdaniach, zawierający między innymi informacje skąd one pochodzą i jest on w formacie ‘.json’.

¹²NKJP – Narodowy Korpus Języka Polskiego (zob. Przepiórkowski i in. 2012). Część tego korpusu, którą znakowano ręcznie, nazywa się NKJP1M.

¹³Części *train*, *dev*, *test* to zwyczajowe nazwy na trzy zbiory danych w przetwarzaniu języka naturalnego, w których część *train* służy do uczenia modelu, część *dev* do jego bieżącej ewaluacji i *test* do ostatecznej oceny.

3.2. Preprocessing danych

Preprocessing danych ma na celu wyodrębnienie z korpusu PDB tylko tych zdań, które zawierają koordynację, oraz wydobywanie informacji na temat tych koordynacji. Interesują mnie tylko dwa człony koordynacji – pierwszy i ostatni, zatem to informacje właśnie o nich są tu kluczowe. Preprocessing robię w języku Python, w czterech osobnych plikach, które znajdują się w Załączniku A. Najpierw wczytuję opisane wcześniej dane z PDB, zapisując je w postaci list, a następnie wyszukuję w nich koordynacji (szukając wyrazów, które są nadrzędne zależnością o etykiecie *conjunct* dla przynajmniej 2 innych wyrazów – są to spójniki współrzędne) i tworzę osobną listę składającą się tylko z tych koordynacji, zapisując w niej informację o obu członach, o spójniku i o nadrzędniku. Na przykładzie zdań 18–19, pochodzących z PDB, pokażę format danych oraz w jaki sposób się go odczytuje.

(18) [Mały, jasny] *ptak* pochyla głowę w stronę leżącego obok okruszka.

(19) Boguś [mieszka] tu [i pracuje]¹⁴.

W (20)–(21) przedstawiam tabele z tymi zdaniami w lekko uproszczonym formacie ‘.con11’, a w (22)–(23) – te same informacje przetłumaczone na drzewa zależnościowe¹⁵. Na przykładzie tych dwóch zdań zilustruję, jakie informacje o koordynacji zostają wyciągnięte w trakcie preprocessingu.

(20)

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Mały	mały	adj	adj:sg:nom:m2:pos	sg nom m2 pos	2	conjunct
2	,	,	interp	interp	—	4	adjunct
3	jasny	jasny	adj	adj:sg:nom:m2:pos	sg nom m2 pos	2	conjunct
4	ptak	ptak	subst	subst:sg:nom:m2	sg nom m2	5	subj
5	pochyla	pochylać	fin	fin:sg:ter:imperf	sg ter imperf	0	root
6	głowę	głowa	subst	subst:sg:acc:f	sg acc f	5	obj
7	w	w	prep	prep:acc:nwok	acc nwok	5	adjunct_adl
8	stronę	strona	subst	subst:sg:acc:f	sg acc f	7	mwe
9	leżącego	leżeć	pact	pact:sg:gen:m3:imperf:aff	sg gen m3 imperf aff	11	adjunct
10	obok	obok	adv	adv	—	9	adjunct_locat
11	okruszka	okruszek	subst	subst:sg:gen:m3	sg gen m3	8	comp
12	.	.	interp	interp	—	5	punct

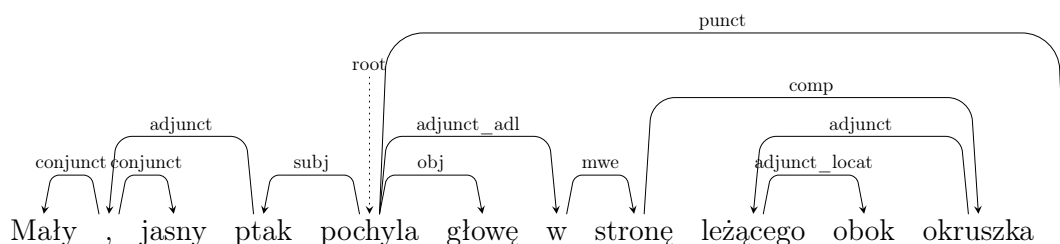
¹⁴Jak widać, w przykładzie (19), między członami, poza spójnikiem, występuje także wyraz *tu*. Według PDB jest on podrzędny względem spójnika *i*, ale relacją *adjunct_locat*, a nie *conj*, zatem nie jest on częścią koordynacji. Z tego powodu nawias kwadratowy jest tutaj rozbity na dwie części.

¹⁵Oryginalne pliki z danymi z PDB znajdują się w Załączniku B.

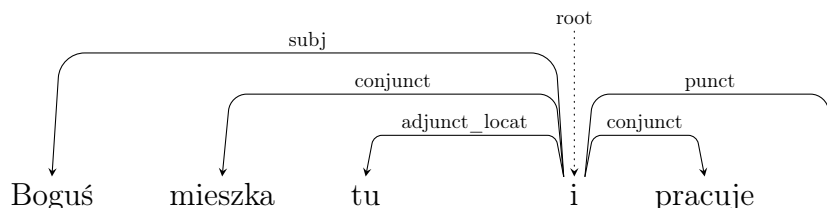
(21)

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Boguś	Boguś	subst	subst:sg:nom:m1	sg nom m1	4	subj
2	mieszka	mieszkać	fin	fin:sg:ter:imperf	sg ter imperf	4	conjunct
3	tu	tu	adv	adv	—	4	adjunct_locat
4	i	i	conj	conj	—	0	root
5	pracuje	pracować	fin	fin:sg:ter:imperf	sg ter imperf	4	conjunct
6	.	.	interp	interp	—	4	punct

(22)



(23)



Dla każdej koordynacji tworzę wiersz w tabeli, w którym umieszczam konkretne informacje. Oznaczam je tutaj małymi rzymskimi numerami, które odwołują się do tabeli z przykładu (24). Dla koordynacji, które mają nadrzędnik, do tabeli wstawiam kolejno: (i) pozycję nadrzędnika, (ii) słowo będące nadrzędnikiem, (iii) pełny tag¹⁶ nadrzędnika, (iv) skrócony tag nadrzędnika, (v) informacje morfosyntaktyczne o nadrzędniku, a dla koordynacji bez nadrzędnika wstawiam tam puste wartości (poza pozycją nadrzędnika – w tym przypadku wstawiam tam wartość 0). Następnie niezależnie od obecności nadrzędnika umieszczam w tabeli (vi) etykietę koordynacji, (vii) spójnik współrzędny, (viii) tag spójnika, (ix) liczbę koniunktów, oraz następujące informacje o pierwszym i ostatnim członie koordynacji: (x, xxi) pełny człon, (xi, xxii) człon podzielony na sylaby¹⁷, (xii, xxiii) głowa tego członu, (xiii, xxiv) pełny oraz (xiv, xxv) skrócony tag głowy członu, (xv, xxvi) informacje morfosyntaktyczne o głowie członu, (xvi, xxvii) liczbę słów danego członu, (xvii, xxviii) liczbę jego tokenów, (xix, xxx) liczbę jego sylab, (xx, xxxi) liczbę jego znaków (wliczając spacje) oraz (xx, xxxi) informację o tym, czy jest on ciągły, tj. czy wszystkie jego tokeny występują kolejno po sobie,

¹⁶Jako *tag* rozumie się oznaczenie danej części mowy, tutaj wraz z jej odmianą.

¹⁷Aby policzyć sylaby w członach, użyłem bibliotek: *num2words* (<https://pypi.org/project/num2words/>) do zamieniania liczb na tekst oraz *pyphen* (<https://pypi.org/project/pyphen/>) do dzielenia fraz na sylaby. Oba pakiety mogły popełniać małe błędy, jednak statystycznie powinny to robić w takim samym stopniu w członie lewym co prawym, więc nie powinno to zaburzać wyników analiz.

czy między nimi znajduje się jakiś token niebędący częścią tego członu. Na sam koniec dodaję do tabeli (xxxii) całe zdanie, w którym występuje koordynacja, (xxxiii) jego identyfikator oraz (xxxiv) informację o tym, czy jest ono w zbiorze treningowym, walidacyjnym czy testowym.

3.3. Dane po preprocessingu

Dane po preprocessingu zawarte są w Załączniku C. Tutaj zilustruję je na podstawie wcześniejszych przykładów.

(24) Przykład danych dla dwóch koordynacji wyciągniętych z korpusu PDB

governor.positionⁱ	governor.wordⁱⁱ	governor.tagⁱⁱⁱ	governor.pos^{iv}
R	ptak	subst:sg:nom:m2	subst
0			

governor.ms^v	coordination.label^{vi}	conjunction.word^{vii}	conjunction.tag^{vii}
sg nom m2	adjunct	,	interp
	root	i	conj

no.conjuncts^{ix}	L.conjunct^x	L.conj.syllabified^{xi}	L.head.word^{xii}
2	Mały	Ma~ły	Mały
2	mieszka	miesz~ka	mieszka

L.head.tag^{xiii}	L.head.pos^{xiv}	L.head.ms^{xv}	L.words^{xvi}	L.tokens^{xvii}
adj:sg:nom:m2:pos	adj	sg nom m2 pos	1	1
fin:sg:ter:imperf	fin	sg ter imperf	1	1

L.syllables^{xviii}	L.chars^{xix}	L.is.continuous^{xx}	R.conjunct^{xxi}	R.conj.syllabified^{xxii}
2	4	1	jasny	jas~ny
2	7	1	pracuje	pra~cu~je

R.head.word^{xxiii}	R.head.tag^{xxiv}	R.head.pos^{xxv}	R.head.ms^{xxvi}	R.words^{xxvii}
jasny	adj:sg:nom:m2:pos	adj	sg nom m2 pos	1
pracuje	fin:sg:ter:imperf	fin	sg ter imperf	1

R.tokens^{xxviii}	R.syllables^{xxix}	R.chars^{xxx}	R.is.continuous^{xxxi}
1	2	5	1
1	3	7	1

sentence^{xxxii}

Mały, jasny ptak pochyła głowę w stronę leżącego obok okruszka.
Boguś mieszka tu i pracuje.

sent.id^{xxxiii}

CDScorpus_6721_B#1673
200-2-000000212_morph_9.61-s#6421

sent.file^{xxxiv}

test
test

W tabeli po preprocessingu znajduje się łącznie 13247 koordynacji, w tym w 3828 nie występuje nadrzędnik (np. w przypadku (26) poniżej), w 7730 występuje on po lewej stronie (jak wyżej w (18)), w 44 pomiędzy członami (np. w (25) poniżej), a w 2045 po prawej stronie (jak w (19) wyżej). Koordynacje zagnieżdżone (jak w (27) poniżej) również są uwzględniane, więc mamy pewność, że wszystkie koordynacje występujące w tym korpusie zostały wyciągnięte. Koordynacji dwuczłonowych jest 11635, trzyczłonowych – 1171, jest także 265 czteroczłonowych, 90 pięcioczłonowych, 47 sześcioczłonowych, 16 siedmioczłonowych, 10 ośmioczłonowych, 3 dziewięcioczłonowe, 3 dziesięcioczłonowe, 2 jedenastoczłonowe, 2 dwunastoczłonowe, 2 trzynastoczłonowe i jedna czternastoczłonowa.

- (25) – [Nie agitujemy ani za, ani przeciw ociepleniom] – *zapewniał* prezes – [ale informujemy o warunkach]¹⁸.
- (26) Słoń z pomalowanymi [uszami i trąbą] idzie po bruku.
- (27) – Byłem świadkiem, *kiedy* góra lodowa dryfująca po zatoce, [pękła i rozkruszyła się na [trzy, **albo** cztery części]] – opowiada pan Tadeusz.

Z oczyszczonych danych, możemy odczytać jakie spójniki występują w koordynacjach w PDB i są to: *a, albo, ale, ani, bądź, co, czy, czyli, ewentualnie, i, ile, inaczej, jak, jednak, jednakże, lecz, lub, miast, natomiast, ni, niemniej, oraz, przy, to, tyle, tylko, tymczasem, względnie, zaś* oraz znaki interpunkcyjne: *-, -, —, , , ; , : , ! , . , . . . , & ,* a także znaki matematyczne (i ich słowne określenia): */, +, x, minus, plus, razy*. Poza nimi, pojawiły się także dwa wystąpienia angielskiego *and* oraz jedno wystąpienie francuskiego *et*.

¹⁸Podobnie jak w przykładzie (19), koordynacja jest podzielona na dwie części, ponieważ między członami znajdują się części nienależące do tej struktury.

Rozdział 4

Analiza statystyczna

W rozdziale tym przedstawiam wyniki analizy statystycznej dla wyodrębnionych koordynacji, prezentując też istotne tabele oraz wykresy.

4.1. Podstawowa analiza

Jak widać w tabeli (28), średnia długość lewego członu koordynacji jest krótsza od średniej długości prawego członu, patrząc zarówno na znaki, jak i sylaby, słowa oraz tokeny. Mediana albo jest taka sama, albo również mniejsza dla lewego członu. Efekty te widać dla każdej z czterech grup koordynacji – koordynacji bez nadrzędnika, tych z nadrzędnikiem po lewej stronie, tych z nadrzędnikiem pomiędzy członami oraz, co najważniejsze, tych z nadrzędnikiem po prawej stronie. Ten ostatni wynik od razu potwierdza, że hipoteza o tym, że lewy człon koordynacji jest krótszy od prawego, jest prawdziwa, w opozycji do hipotezy o tym, że to człon bliższy nadrzędnika jest krótszy. Wszystkie 12 efektów jest istotnych statystycznie $p < 0,05$), co potwierdza test Wilcozona, zatem możemy uznać, że różnice w długościach są istotne. W dalszych etapach analizy wszystkie efekty badam tylko dla koordynacji bez nadrzędnika lub z nadrzędnikiem po którejś ze stron, ponieważ koordynacje z nadrzędnikiem pomiędzy członami są rzadkie i nie ma ich wystarczającej liczby, aby móc przeprowadzić dokładną analizę statystyczną.

(28)

m e d i a n a		ś r e d n i a		V	p	
lewy	prawy	lewy	prawy			
Wszystkie koordynacje (N = 13 247)						
słowa	3	3	3,90	5,08	7,72e07	5,44e-67
tokeny	3	3	4,19	5,54	7,71e07	1,20e-67
sylaby	6	8	9,04	11,69	7,56e07	4,46e-85
znaki	18	23	27,00	35,14	7,57e07	4,96e-84
Koordynacje bez nadrzędnika (N = 3 828)						
słowa	5	6	6,41	8,27	6,00e06	4,73e-43
tokeny	5	7	6,95	9,10	6,01e06	1,74e-42
sylaby	11	14	14,13	18,04	6,13e06	3,09e-35
znaki	34	42	43,09	55,17	6,11e06	3,92e-36
Koordynacje z nadrzędnikiem po lewej stronie (N = 7 330)						
słowa	2	2	2,98	4,00	2,35e07	3,00e-43
tokeny	2	2	3,18	4,33	2,35e07	3,23e-43
sylaby	5	6	7,26	9,66	2,27e07	3,34e-60
znaki	14	18	21,26	28,63	2,27e07	3,82e-58
Koordynacje z nadrzędnikiem pomiędzy członami (N = 44)						
słowa	4	5,5	5,48	6,89	7,17e02	3,57e-02
tokeny	4,5	6	6,18	7,52	7,48e02	6,52e-02
sylaby	9,5	12	12,50	14,32	7,68e02	9,45e-02
znaki	27,5	34,5	37,45	44,45	7,41e02	5,86e-02
Koordynacje z nadrzędnikiem po prawej stronie (N = 2 045)						
słowa	1	2	2,51	2,97	1,92e06	1,49e-06
tokeny	1	2	2,63	3,18	1,92e06	5,99e-07
sylaby	4	4	5,84	7,00	1,83e06	1,49e-12
znaki	10	12	17,23	20,77	1,82e06	1,11e-13

4.2. Dalsza analiza

Wiemy już, że w języku polskim, podobnie jak w angielskim, lewy człon koordynacji jest średnio krótszy niż prawy człon. W tej sekcji skupiam się na hipotezie, że nadrzędnik jednak „przyciąga” do siebie krótszy człon – co osłabia ogólną tendencję do umieszczania krótszego członu po lewej stronie, gdy nadrzędnik jest z prawej strony. Możemy to częściowo zauważyć w tabeli (29), pokazującej proporcje rozkładu krótszego członu z lewej strony względem krótszego członu z prawej strony (człony o równej długości odpowiednie dla każdej z miar długości zostały usunięte). W przypadku znaków oraz sylab, efekt ten nie jest istotny statystycznie, jednak przy słowach i tokenach, jest już istotny ($p < 0,05$), mierzone, używając dwustronnego testu proporcji (chi-kwadrat).

(29)

	n a d r z ę d n i k					
	z lewej str.			z prawej str.		
	prop.	<i>N</i>	prop.	<i>N</i>	$\chi^2(1)$	<i>p</i>
słowa	0,674	4249	0,635	979	5,256	0,0219
tokeny	0,671	4309	0,633	986	5,040	0,0248
sylaby	0,645	6145	0,636	1592	0,463	0,4962
znaki	0,637	6750	0,623	1872	1,208	0,2716

Założenie, że bliskość nadrzędnika wpływa na tendencję do umieszczania krótszego członu z lewej strony, osłabiając ją w przypadku, gdy nadrzędnik jest z prawej strony, może również objawiać się tym, że w przypadku braku nadrzędnika wartości proporcji będą się znajdowały pomiędzy wartościami dla nadrzędnika z lewej i prawej strony. Tak się jednak nie dzieje, co widać w tabeli (30). Proporcje te są niższe niż dla koordynacji z nadrzędnikiem z lewej lub prawej strony, niezależnie od przyjętej miary długości członów.

(30)

	brak nadrzędnika		vs z lewej		vs z prawej	
	prop.	<i>N</i>	$\chi^2(1)$	<i>p</i>	$\chi^2(1)$	<i>p</i>
słowa	0,621	3369	23,015	1,61e-06	4,450	0,445
tokeny	0,619	3390	22,301	2,33e-06	2,175	0,448
sylaby	0,604	3641	16,320	5,35e-05	4,490	0,034
znaki	0,604	3757	11,143	8,44e-04	1,802	0,179

Efekt ten można wyjaśnić tym, że koordynacje bez nadrzędnika to zazwyczaj koordynacje zdań lub fraz czasownikowych, a jak wskazują Przepiórkowski i Woźniak (2023) wiele z tych koordynacji ma strukturę zdania lub frazy czasownikowej i komentarza do niej (zob. (31)), czy też długiego zdania, po którym następuje następne, znacznie krótsze, bo używające elipsy (zob. (32)).

(31) [— Aaa — powiedziała Margie Tallworth; zatkało ją].

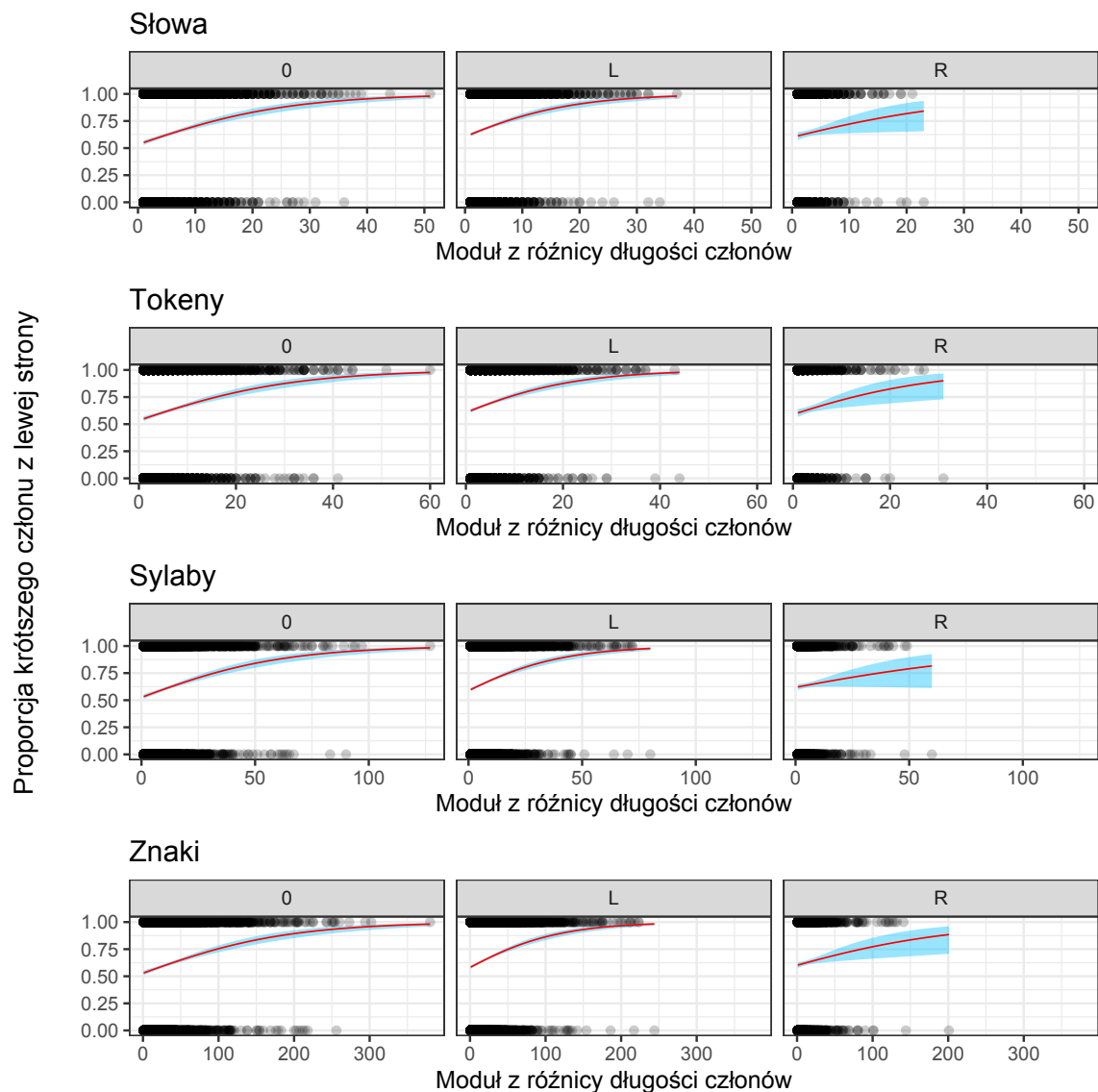
(32) [Stary pan Przypkowski był kolekcjonerem, naukowcem i konstruktorem zegarów, a młody — rzeźbiarzem i germanofilem].¹⁹

Oznacza to, że porównywanie całkowitych proporcji nie jest właściwym argumentem za wpływem bliskości nadrzędnika.

Jednakże, jeśli spojrzymy na Rys. 1, możemy zauważyć, że w przypadku koordynacji z nadrzędnikiem z prawej strony, tendencja do umieszczania krótszego członu z lewej strony wydaje się rosnać wraz ze wzrostem modułu z długości różnicy członów wolniej, niż w pozostałych dwóch przypadkach. Przy koordynacjach bez nadrzędnika, jak i tych z nadrzędnikiem z lewej strony, wpływ bezwzględnej wartości z różnicy długości

¹⁹Oba przykłady pochodzą z PDB.

członów jest silnie zauważalny – proporcje krótszego lewego członu rosną wraz z jej wzrostem, z wartością p poniżej 0,001 dla wszystkich 8 przypadków (2 pozycje nadrzędnika razy 4 miary długości)²⁰. W przypadku koordynacji z nadrzędnikiem z prawej strony, wartości p dla tej samej zależności są równe 0,0089, 0,0743, 0,0319 oraz 0,0066 odpowiednio dla znaków, sylab, słów i tokenów. Przyjmując standardowe kryterium istotności statystycznej ($p < 0,05$), nachylenie wykresu jest istotnie dodatnie dla słów, tokenów oraz znaków.



Rys. 1: Wykresy proporcji krótszego lewego członu w zależności od modułu z różnicy długości członów, mierzone za pomocą odpowiednich miar długości tekstu.

W tabeli (33) znajdują się porównania parami nachyleń wykresów dla różnych pozycji nadrzędnika, z podziałem na miary długości tekstu. Zawiera ona różnice nachyleń, współczynniki t oraz wartości p . Jak możemy zauważyć, okazuje się, że jedynie dwie

²⁰Liczone funkcją `emtrends` z pakietu `emmeans` (<https://github.com/rvnlenth/emmeans>)

wartości są istotne statystycznie, jest to różnica między koordynacjami bez nadrzędnika, a tymi z nadrzędnikiem z lewej, patrząc na znaki oraz różnica między koordynacjami z nadrzędnikiem z prawej strony, a tymi z nadrzędnikiem z lewej, jeśli mierzymy długości członów w sylabach.

(33)

	p o z y c j a n a d r z ę d n i k a (p a r a m i)								
	b r a k – l e w a			b r a k – p r a w a			l e w a – p r a w a		
	różnica	z	p	różnica	z	p	różnica	z	p
słowa	−0,0202	−1,486	0,2977	0,0177	0,658	0,7875	0,0379	1,363	0,3606
tokeny	−0,0156	−1,378	0,3522	0,0008	0,035	0,9993	0,1642	0,694	0,7673
sylaby	−0,0114	−2,082	0,0936	0,0143	1,422	0,3296	0,0257	2,487	0,0345
znaki	−0,0051	−2,890	0,0108	0,0020	0,601	0,8193	0,0071	2,097	0,0904

Można by zastanawiać się nad zastosowaniem do otrzymanych wyników poprawki Bonferroniego, czy Holma–Bonferroniego²¹, jednak zdecydowałem się tego nie robić, ponieważ każda z analiz służy analizie tego samego zjawiska, jedynie używając różnych miar długości (ponadto dla każdego wiersza tabeli jest różny zbiór danych). Dlatego logicznym jest, że nawet gdyby wyniki wyszły istotne statystycznie (np. $p = 0,04$) dla każdego z przypadków, chcielibyśmy zachować tę istotność statystyczną, a nie zmniejszyć jej górną granicę z 0,05 na 0,0125 (0,05/4), przez co wyniki mogłyby się okazać nieistotne statystycznie (np. $p = 0,04 > 0,0125$). W przypadku zastosowania którejkolwiek z tych poprawek oraz przy wartościach p poniżej 0,05 silniejszą istotność dawałoby zrobienie jednej analizy, niż kilku, co jest nieintuicyjne. Ponadto, przy obecnych wynikach, tylko dwie wartości mają istotność statystyczną. Jedna z nich, czyli porównanie w znakach koordynacji bez nadrzędnika i koordynacji z nadrzędnikiem z lewej strony, nie wpływa na hipotezę. Natomiast druga, czyli porównanie w sylabach koordynacji z nadrzędnikiem z prawej strony i koordynacji z nadrzędnikiem z lewej strony, nie potwierdza hipotezy, ponieważ 3 pozostałe wyniki (dla innych miar długości) nie są istotne statystycznie. Jest tak, ponieważ hipoteza nie zostanie potwierdzona, gdy tylko jedna z analiz będzie miała istotność statystyczną. O jej potwierdzeniu można by mówić jedynie gdyby zauważono te istotności w każdym z przypadków. Nawet po zastosowaniu którejś z poprawek, wartości (i kierunek) trendu pozostałyby bez zmian, jedynie jego istotność statystyczna by się zmniejszyła. W związku z tym, zdecydowałem się na pozostanie przy standardowym kryterium istotności statystycznej $p < 0,05$.

²¹zob. https://pl.wikipedia.org/wiki/Poprawka_Bonferroniego, https://pl.wikipedia.org/wiki/Poprawka_Holma-Bonferroniego

Rozdział 5

Dyskusja wyników

W tym rozdziale omawiam wyniki analizy z poprzedniego rozdziału, interpretuję je w kontekście istniejącej literatury oraz wskazuję możliwe kierunki dalszych badań.

5.1. Podsumowanie wyników badań

Zgodnie z oczekiwaniami, lewy człon koordynacji jest krótszy niezależnie od pozycji nadrzędnika oraz od przyjętej miary długości tekstu. Co do drugiej hipotezy, mówiącej o tym, że istnieje tendencja do preferowania szeregu z krótszym członem z lewej strony wraz ze wzrostem różnicy długości członów oraz że tendencja ta jest silniejsza w przypadku braku nadrzędnika lub obecności nadrzędnika z lewej strony, wyniki są niejednoznaczne. Możemy zauważyć wspomnianą tendencję, jednak jest ona obecna: (1) w przypadku braku nadrzędnika – niezależnie od przyjętej miary długości, (2) w przypadku nadrzędnika z lewej strony – również niezależnie od przyjętej miary długości, (3) w przypadku nadrzędnika z prawej strony – tylko w przypadku mierzenia długości członów w znakach, tokenach oraz słowach. Zgodnie z hipotezą, tendencja ta powinna być widoczna tylko w przypadku braku nadrzędnika oraz nadrzędnika z lewej strony. Jedynie przyjmując za miarę długości sylaby możemy zauważyć, że tendencja ta spełnia nasze oczekiwania. Niezależnie od przyjętej miary długości, widać jednak pewne trendy, które nie są istotne statystycznie (poza miarę długości w sylabach), ale wszystkie są nakierowane w tym samym kierunku. Dla każdej miary długości wynik wskazuje na to, że tendencja do silniejszego preferowania krótszego członu z lewej strony wraz ze wzrostem modułu z długości różnicy członów jest najsilniejsza w przypadku nadrzędnika z lewej strony, słabsza w przypadku braku nadrzędnika oraz najslabsza w przypadku u nadrzędnika z prawej strony. Dla sylab istotna statystycznie jest jedna wartość, a mianowicie różnica między koordynacjami z nadrzędnikiem z prawej strony a koordynacjami z nadrzędnikiem z lewej strony. Częściowo potwierdza to hipotezę, która zachodzi również w języku angielskim (Przepiórkowski i Woźniak, 2023). Jeśli chodzi o różnicę między koordynacjami z nadrzędnikiem z prawej strony a koordynacjami

bez nadrzędnika, to nawet patrząc tylko na sylaby nie widać istotności statystycznej. Dla innych miar długości w wynikach również występuje trend skierowany zgodnie z oczekiwaniami, nie jest on jednak na tyle silny by być istotnym statystycznie.

5.2. Interpretacja wyników

Wyniki te, mimo że nie są jednoznaczne, nie wykluczają prawdziwości postawionych wcześniej hipotez. Wręcz przeciwnie, wskazują na istnienie pewnego trendu, zgodnego z oczekiwaniami. Trend ten, poza przypadkiem przyjęcia sylab za miarę długości przy porównywaniu koordynacji z nadrzędnikiem z lewej strony, a tych z nadrzędnikiem z prawej, nie jest jednak istotny statystycznie. Może to wynikać z faktu, że próba jest mniejsza niż przy analogicznym badaniu dla języka angielskiego (Przepiórkowski i Woźniak, 2023), gdzie badano 24446 koordynacji, podczas gdy w tym badaniu było ich prawie dwukrotnie mniej – 13247. Patrząc tylko na liczby koordynacji z nadrzędnikiem z prawej strony, to w tym badaniu było ich 2045, podczas gdy w badaniu dla języka angielskiego było ich 4179, czyli ponad dwukrotnie więcej. Gdyby założyć istotność statystyczną zauważonego trendu, wraz z DLM byłby on argumentem za symetrycznymi reprezentacjami koordynacji (londyńską, praską), jako że to one ilustrują, że suma długości zależności się minimalizuje, gdy nadrzędnik „przyciąga” do siebie krótszy człon.

5.3. Przegląd literatury

Główną pracą, z której brałem inspirację do tego badania, jest praca Przepiórkowskiego i Woźniaka (2023), w której autorzy badali zjawisko to w języku angielskim. Zauważyli oni, że dla różnych typów koordynacji w języku angielskim zostało potwierdzone, że pierwszy człon jest na ogół krótszy niż ostatni człon (Gibson i in., 1996; Temperley, 2005; Lohmann, 2014). Zwrócili uwagę, że nie ma badania, w którym ktoś wziąłby pod uwagę wszystkie koordynacje, nie patrząc tylko na jeden konkretny jej typ, a także uwzględniając różne miary długości członów. W swojej pracy autorzy chcieli sprawdzić tezę, że może to nie lewy człon jest zawsze krótszy, a człon bliższy nadrzędnikowi – jako że koordynacje z nadrzędnikiem z prawej strony występują rzadziej niż te z nadrzędnikiem z lewej strony, toteż zdawałoby się to być dalej zgodne z wynikami zacytowanych badań. Zbadali więc ją na przykładzie korpusu Penn Treebank (Marcus, 1993). Dokładniej, użyli oni jednej z wersji tego korpusu, a mianowicie wersji, którą sami nazwali PTB_&, która została udostępniona przez Ficlera i Goldberga (2016) i która miała w sobie pewne poprawki względem oryginału oraz była wzbogacona o informacje o koordynacjach. W swoim badaniu autorzy zbadali 24446 koordynacji, z których 4179 miało nadrzędnik z prawej strony. Ich wyniki wskazują na to, że w języku angielskim lewy człon koordynacji jest zazwyczaj krótszy niż prawy człon. Ponadto, proporcja

koordynacji, w których lewy człon jest krótszy, zwiększa się wraz ze wzrostem modułu z różnicy długości członów. Nie dzieje się tak jednak w przypadku koordynacji z nadrzędnikiem z prawej strony, a między współczynnikiem tej korelacji dla tego typu koordynacji a zarówno koordynacjami bez nadrzędnika, jak i tymi z nadrzędnikiem z lewej strony występuje istotna statystycznie różnica. Autorzy wskazali, że ich wyniki, wraz z DLM, mogą być argumentami za symetrycznymi reprezentacjami koordynacji.

5.4. Perspektywy dalszych badań

Z racji, że w badaniu tym nie udało się jednoznacznie potwierdzić hipotez, które postawiono, warto byłoby je powtórzyć na większej próbie danych, tak aby móc uzyskać bardziej jednoznaczne wyniki. Warto byłoby również powtórzyć badanie dla języka angielskiego, na bardziej zróżnicowanej próbie danych, tak aby móc porównać wyniki dla obu języków. Mogłyby one być wtedy bardziej zbieżne lub wręcz przeciwnie – bardziej rozbieżne. W pierwszym przypadku, ponadto, mogłyby one potwierdzić postawione hipotezy, a co za tym idzie, potwierdzić wyższość symetrycznych reprezentacji zależnościowych nad asymetrycznymi, lub też odwrotnie – wskazać, że trend zauważony w obu badaniach jest jedynie następstwem braku zróżnicowania danych i nie jest związany z faktycznymi preferencjami językowymi, toteż przy większym zróżnicowaniu danych, mogłoby go nie być. Można by również powtórzyć badanie dla innych języków, tak aby móc porównać wyniki dla większej ich liczby, a także móc wyciągnąć jakieś wnioski na temat całych rodzin językowych. Aby jednak móc to wszystko osiągnąć, potrzeba by mieć dostęp do korpusów z większą liczbą danych, w których równocześnie koordynacje byłyby anotowane w sposób pozwalający na jednoznaczne wskazanie nadrzędnika koordynacji (lub jego braku) oraz na jednoznaczne wskazanie, które słowa/tokensy wchodziły w skład którego członu koordynacji. Jak wspomnieli Przepiórkowski i Woźniak (2023), wydaje się, że najlepiej by było przeprowadzić takie badanie na UD (Universal Dependencies, <https://universaldependencies.org/>), czyli korpusie zawierającym ponad 200 zbiorów drzew zależnościowych dla ponad 100 języków i posiadającym anotacje zależnościowe w jednym, uniwersalnym formacie. Format ten jednak nie spełnia jednego z powyższych założeń, a mianowicie, nie wskazuje jednoznacznie, które części zdania należą do którego członu koordynacji. Nowsze wersje UD, w formacie nazywanym EUD (Enhanced Universal Dependencies) spełniają oba kryteria, jednak nie są one jeszcze tak powszechnie dostępne, jak wersje podstawowe. Możliwe, że w przyszłości, będzie się dało wykonać takie badanie właśnie na zbiorach EUD.

Bibliografia

- Covington, M.A. (1984). *Syntactic Theory in the High Middle Ages: Modistic Models of Sentence Structure* (Cambridge Studies in Linguistics). Cambridge University Press.
- Ficler, J. i Goldberg, Y. (2016). Coordination annotation extension in the Penn Tree Bank. W *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 834–842.
- Futrell, R., Mahowald, K., i Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. W *Proceedings of the National Academy of Sciences* 112(33), 10336–10341. <https://doi.org/10.1073/pnas.1502134112>
- Futrell, R., Levy R. P., i Gibson, E. (2020). Dependency locality as an explanatory principle for word order. W *Language* 96(2), 371–412.
- Gibson, E., Schütze, C. T. i Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. W *Journal of Psycholinguistic Research*, 25(1), 59–92. https://tedlab.mit.edu/tedlab_website/researchpapers/Gibson_et_al_1996_JPR.pdf
- Hajič, J., Panevová, J., Hajičová, E., Petr Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M. i Urešová, Z. (2006). Prague Dependency Treebank 2.0 (PDT 2.0). <https://hdl.handle.net/11858/00-097C-0000-0001-B098-5>
- Haspelmath, M. (2007). Coordination. W *Language Typology and Syntactic Description, Volume II: Complex constructions*, 1–51. Cambridge University Press.
- Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.
- Hudson, R. (1984). *Word Grammar*. Blackwell.
- Hudson, R. (1990). *English Word Grammar*. Blackwell.
- Hudson, R. (2010). *An Introduction to Word Grammar*. Cambridge University Press.

- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. W *Proceedings of the 10th Machine Translation Summit Conference*, 79–86. <https://aclanthology.org/2005.mtsummit-papers.11.pdf>
- Kruijff, G.-J. M. (2002). Formal and computational aspects of dependency grammar: History and development of DG. W *Technical report*, ESSLI2002.
- Lohmann, A. (2014). *English Coordinate Constructions: A Processing Perspective on Constituent Order*. Cambridge University Press.
- Marcus, M. P., Santorini, B. i Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. W *Computational Linguistics* 19, 313–330. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1246&context=cis_reports
- de Marneffe, M.-C., MacCartney, B. i Manning, C. D. (2006). Generating typed dependency parsers from phrase structure parses. W *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 449–454. https://www.lrec-conf.org/proceedings/lrec2006/pdf/440_pdf.pdf
- de Marneffe, M.-C. i Nivre, J. (2019). Dependency Grammar. W *Annual Review of Linguistics* 5, 197–218. <https://doi.org/10.1146/annurev-linguistics-011718-011842>
- Mel’čuk, I.A. (1974). *Opyt teorii linvističeskix modelej «Smysl ⇔ Tekst»*. Nauka.
- Mel’čuk, I.A. (1988). *Dependency Syntax: Theory and Practice*. SUNY Press.
- Mel’čuk, I.A. (2009). Dependency in natural language. W *Dependency in Linguistic Description*, 1–110. John Benjamins.
- Pedersen, M., Eades, D. Amin, S. K. i Prakash, L. (2004). Relative clauses in Hindi and Arabic: A Paninian dependency grammar analysis. W *Proceedings of the Workshop on Recent Advances in Dependency Grammar*, 9–16. COLING.
- Pęzik, P., Ogrodniczuk, M. i Przepiórkowski, A. (2011). Parallel and spoken corpora in an open repository of Polish language resources. W *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 511–515. <https://nlp.ipipan.waw.pl/Bib/pez:ogr:prz:11.pdf>
- Przepiórkowski, A., Bańko, M., Górski, R. i Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN. https://nkjp.pl/settings/papers/NKJP_ksiazka.pdf

- Przepiórkowski, A. (2017). *Argumenty i modyfikatory w gramatyce i w słowniku*. Wydawnictwa Uniwersytetu Warszawskiego. https://wuw.pl/data/include/cms/Argumenty_modyfikatory_Przepiorkowski_Adam_2017.pdf
- Przepiórkowski, A. i Patejuk, A. (2020). From Lexical Functional Grammar to Enhanced Universal Dependencies. W *Language Resources & Evaluation* 54, 185–221. <https://doi.org/10.1007/s10579-018-9433-z>
- Przepiórkowski, A. i Woźniak, M. (2023). Conjunct lengths in English, Dependency Length Minimization, and dependency structure of coordination. [Manuskrypt przyjęty do publikacji w ACL 2023]. <https://ling.auf.net/lingbuzz/007283>
- Popel, M., Mareček, D., Štěpánek, J., Zeman, D. i Žabokrtský, Z. (2013). Coordination structures in dependency treebanks. W *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 517–527. <https://aclanthology.org/P13-1051.pdf>
- Steinberger, R., Eisele, A., Kłoczek, S., Pilos, S. i Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. W *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 454–459. https://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf
- Temperley, D. (2005). The dependency structure of coordinate phrases: A corpus approach. W *Journal of Psycholinguistics Research* 34(6), 577–601. <http://davidtemperley.com/wp-content/uploads/2015/11/temperley-jpr05.pdf>
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. C. Klincksieck.
- Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. W *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2214–2218. https://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
- Wróblewska, A. (2014). *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank* [Rozprawa Doktorska, Instytut Podstaw Informatyki Polskiej Akademii Nauk]. <https://nlp.ipipan.waw.pl/Bib/wro:14.pdf>
- Wróblewska, A. i Krasnowska-Kieraś, K. (2017). Polish evaluation dataset for compositional distributional semantic models. W *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 784–792. Association for Computational Linguistics. <https://aclanthology.org/P17-1073.pdf>

Wróblewska, A. (2020). Towards the conversion of National Corpus of Polish to Universal Dependencies. W *Proceedings of the 12th Language Resources and Evaluation Conference*, 5308—5315. European Language Resources Association. <https://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.653.pdf>

Załączniki

- A – link do plików z preprocessingiem danych: <https://github.com/kvmilos/PracaLicencjacka/tree/master/preprocessing>
- B – link do pliku z analizą danych: <https://github.com/kvmilos/PracaLicencjacka/blob/master/analizy/r.R>
- C – link do tabeli danych po preprocessingu w formacie „csv”: <https://github.com/kvmilos/PracaLicencjacka/blob/master/tabela.csv>
- D – link do pliku z analizą danych: <https://github.com/kvmilos/PracaLicencjacka/blob/master/analizy/r.R>