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

Praca licencjacka na temat „Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie Polish Dependency Bank" jest poświęcona zjawisku minimalizacji długości zależności w 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. Szczególną uwagę poświęcam koordynacji, czyli jednej ze struktur występujących w języku polskim i przedstawiam różne jej reprezentacje proponowane przez lingwistów. Opisuję też sam korpus, wyciągnięte z niego dane oraz ich preprocessing, a także przedstawiam analizy statystyczne badające wpływ pozycji nadrzędnika koordynacji na rozkład długości jej członów oraz interpretuję je, porównując z istniejącą wcześniej literaturą.

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

Rozdział 1

Wstęp

W tym rozdziale przedstawiam motywację i cel niniejszej pracy licencjackiej, 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 umieszczania 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 ich zależności to odpowiednio 2, 1 oraz 3 – mierzone licząc odległości (w słowach) między słowem podrzędnym, a nadrzędnym. 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. Zjawisko to jest istotne dla teorii składniowej i reprezentacji językowych, ponieważ dotyczy zarówno formy, jak i znaczenia zdań. Przykładem koordynacji jest (1), gdzie jej nadrzędnikiem jest słowo *widziałem*, a członami *Asię* oraz *jej śmiesznego, młodszego brata*. Oba człony złączone są spójnikiem *i* oraz razem tworzą większą strukturę, zależną od jej nadrzędnika.

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

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, słowa oraz tokeny[1]. W przykładzie (1) odpowiednie wartości wynosiłyby (4 vs. 31, 2 vs. 9, 1 vs. 4,

[1]tokeny – zalicza się do nich całe słowa (np. 'być', 'kolor'), części słów (m. in. wyrazy po oderwaniu końcówek fleksyjnych oraz same końcówki, (np. 'zrobił', 'em')), a także interpunkcję (np. ', ', '-', '? ')

1 vs. 5). Szybko pokazuję, że pierwsza z hipotez zachodzi w większości przypadków, więc następnie przechodzę do omówienia wpływu obecności i pozycji nadrzędnika oraz długości różnicy między analizowanymi członami na proporcje danych, w których hipoteza ta jest prawdziwa. 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 & Woźniak, 2023).

1.2. Zakres i struktura pracy

Praca składa się z sześciu rozdziałów. W rozdziale drugim omawiam teoretyczne podstawy pracy, tj. przedstawiam czym są koordynacje – na przykładzie języka polskiego, 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 prezentuję hipotezy badawcze, ich testowanie wraz z analizami statystycznymi w języku R. W rozdziale piątym omawiam wyniki badań i ich interpretację w kontekście istniejącej wcześniej literatury naukowej. W rozdziale szóstym podsumowuję pracę, wyciągam z niej wnioski oraz proponuję perspektywy dalszych badań.

Rozdział 2

Podstawy teoretyczne

W tym rozdziale omawiam teoretyczne podstawy pracy, tj. opisuję czym są koordynacje, przedstawiam zarys teorii zależności składniowej, prezentuję minimalizację teorii zależności oraz wskazuję 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 elementów językowych w większe całości. Jest ono również znane pod nazwą struktura współrzędnie złożona. Według definicji Oxford Bibliographies[2] koordynacja to zjawisko, w którym dwa lub więcej elementów, nazywanych w tej pracy członami, są ze sobą połączone 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. Dodatkowo, wydobycie (*ang.* extraction), czyli przemieszczenie jakiejś frazy na lewy kraniec zdania, może występować we wszystkich członach jednocześnie, ale nie może występować tylko w jednym z nich. Koordynacja jednak zachowuje się czasem niesymetrycznie[3] 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

[3] np. wyjątkiem od reguły posiadania tej samej funkcji składniowej jest *Kto i kogo kopnął?* , gdzie wyrazy *kto* oraz *kogo* mają je różne.

(2) a. Łucja gra na pianinie, a Łukasz na gitarze.

b. Laura idzie, a Kuba biegnie do parku.

Według Oxford Bibliographies, jednymi z głównych strategii przyjmowanych w lingwistyce co do koordynacji były: założenie, że koordynacja jest wariantem relacji podrzędnej, gdzie pierwszy człon jest głową, a drugi jest względem niego podrzędny; stwierdzenie, że struktura koordynacyjna wywodzi swoje właściwości nie od spójnika, lecz od wspólnych cech jej członów; przyjęcie analizy trójwymiarowej lub wielodominacyjnej (ang. *multidominance*), która nie może być reprezentowana poprzez tradycyjną strukturę fraz. Koordynacja 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 skoordynowana[3].

(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 powyższym przykładzie koniunktami są: (3a) – Ania, Julia; (3b) – Wesoła Marysia, smutny Janek; (3c) – zjadł obiad, 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). Każde z nich, do łączenia koniunktów, używają różnych zestawów spójników. W koordynacjach koniunkcyjnych człony łączy się m. in. 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 m. in. spójniki *ale*, *lecz*, *zaś*, *natomiast*, *jednak*, a w koordynacjach kauzalnych – *bo*, *bowiem*. W tej pracy skupię się na pierwszych trzech rodzajach koordynacji, jako że to one są oznaczone w PDB.

(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].

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

2.2. Zarys teorii zależności składniowej

De Marneffe i Nivre (2019) oraz Pedersen et al. (2004) zwracają uwagę na to, że teoria zależności składniowej ma długą i bogatą historię, która sięga aż starożytności. 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. Przedstawiał on jej potencjał do uchwycenia podobieństw, jak i różnic między językami. Wróblewska (2014) opisuje, że podstawowymi założeniami teorii Tèsniere'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. w 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, syntaktycznych[3], jednak w tej pracy skupiam się tylko na reprezentacji uwzględniającej czynniki morfoskładniowe oraz wymagania członu głównego 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, a jego krawędź jest zawsze zależnością *root*. W zdaniu nie może być więcej niż jeden korzeń, a z korzenia da się przejść

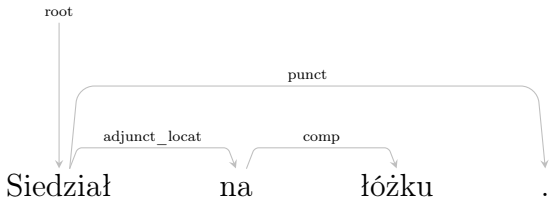
[3] `https://en.wikipedia.org/wiki/Dependency_grammar`, dostęp z dn. 07.04.2023

po strzałkach do każdej innej częsci zdania. Strzałki krawędzi są skierowane zawsze od wyrazu nadrzednego do wyrazu podrzędnego.

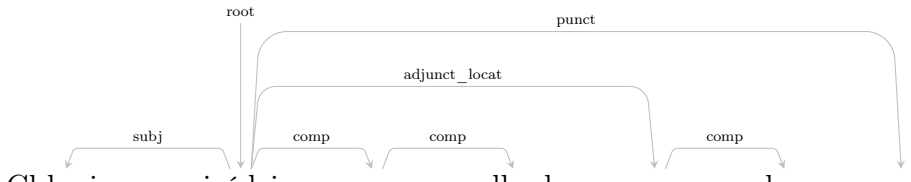
Aby odróżnić od siebie różne zależności, krawędzie mogą być etykietowane, często funkcjami gramatycznymi, jak w przykładach (5a–b). Oto objaśnienia użytych etykiet:

- *root* – korzeń zdania
- *comp* – dopełnienie zdaniowe
- *adjunct_locat* – modyfikator miejsca (modyfikator jest pojęciem trochę różniącym się od tradycyjnego *okolicznika*)
- *punct* – znak interpunkcyjny
- *subj* – podmiot

(5) a.



b.



Przykładowe drzewa zależnościowe z korpusu PDB

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 & Patejuk, 2020; Wróblewska, 2020) oraz cały czas powstają nowe, także w innych językach.

2.3. Minimalizacja długości zależności

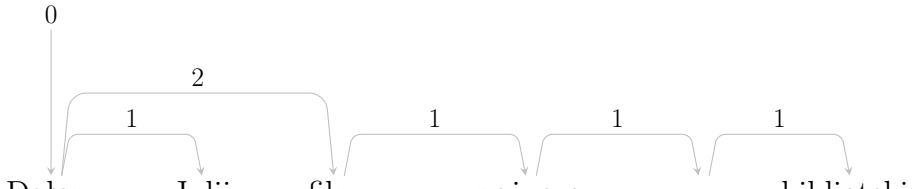
Minimalizacja długości zależności (DLM – ang. *Dependency Length Minimization*) to zasada, według której języki naturalne dążą do zmniejszania odległości między słowami, które są od siebie zależne syntaktycznie. Ułatwia to przetwarzanie informacji i

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.

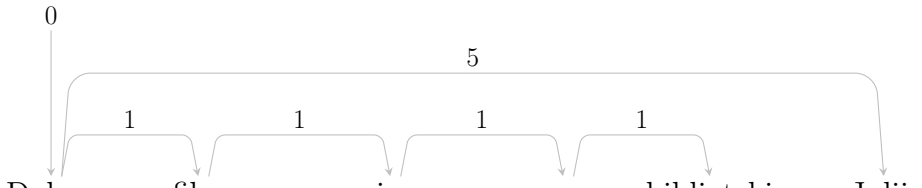
Jednym ze sposobów na badanie DLM jest tworzenie sztucznych języków losowych oraz porównywanie długości zależności w tych językach z długościami zależności w językach naturalnych. Futrell et al. (2015) przedstawili wyniki badań na dużym korpusie tekstów z 37 języków, w których zmierzyli średnią długość 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).

(6) a.



b.



Futrell et al. (2020) oraz Hawkins (1994) twierdzą, że występowanie DLM można rozróżnić na poziom gramatyczny (*grammar*), jak i codzienne użycie języka (*usage*).

Przepiórkowski & Woźniak (2023) wskazują, że na poziomie gramatyki, pewne skonwencjonalizowane kolejności słów okazują się minimalizować średnią długość zależności. Jako przykład podają sytuację w języku angielskim, gdy NP oraz PP są zależne od V[4]. Wtedy kolejność V-NP-PP miałaby średnio krótszą długość zależności, niż V-PP-NP, jako że frazy rzeczownikowe są w języku angielskim średnio krótsze niż frazy przyimkowe.

[4]V – czasownik (ang. *verb*), NP – fraza rzeczownikowa (ang. *nomimal phrase*), PP – fraza przyimkowa (ang. *prepositional phrase*)

Jak dodają Przepiórkowski & Woźniak (2023), Hawkins (1994) argumentuje, że tendencja ta jest skonwencjonalizowana – występuje w gramatyce, ale nie w użyciu codziennym. Jako powód wskazuje, że w języku angielskim kolejność V-NP-PP występuje częściej niż V-PP-NP nie tylko gdy NP jest krótsze od PP, ale i wtedy gdy są podobnej długości. Gdy jednak wydłużymy NP, to kolejność V-PP-NP staje się bardziej naturalna, co znów jest zgodne z hipotezą DLM.

DLM jest również powiązana z innymi właściwościami języków naturalnych, między innymi z pozycyjnością głowy. Pozycyjność głowy jest jednym z kryteriów klasyfikacji języków naturalnych i ma wpływ na ich strukturę syntaktyczną i semantyczną. Oznacza ona kierunek występowania głowy frazy względem jej dopełnienia. Głowa frazy to jej główny element frazy, który decyduje o jej kategorii gramatycznej i znaczeniu. Dopełnienie to element zależny od głowy, który uzupełnia jej znaczenie. Na przykład we frazie *jeść jabłko* czasownik *jeść* jest głową, a rzeczownik *jabłko* – dopełnieniem. W zależności od pozycyjności głowy, języki można podzielić na te o pozycyjności głowy na początku frazy (*head-initial*) oraz na te o pozycyjności głowy na końcu frazy (*head-final*). Na przykład, w języku angielskim, który jest head-initial, czasownik zazwyczaj znajduje się przed rzeczownikiem (*eat an apple*), natomiast w języku japońskim, który jest head-final, ta sama fraza zapisana byłaby jako *jabłko[acc] jeść[npast]* (*ringo-o taberu*)[5]. Pozycyjność głowy ma wpływ na kierunek rozgałęziania się struktury zdaniowej: struktury head-initial są prawostronnie rozgałęzione, a struktury head-final są lewostronnie rozgałęzione.

Badania wykazały, że istnieje związek między pozycyjnością głowy a długością zależności, przy czym języki o pozycyjności głowy na końcu frazy mają średnio krótszą długość zależności niż języki o pozycyjności głowy na początku frazy (Futrell et al., 2015).

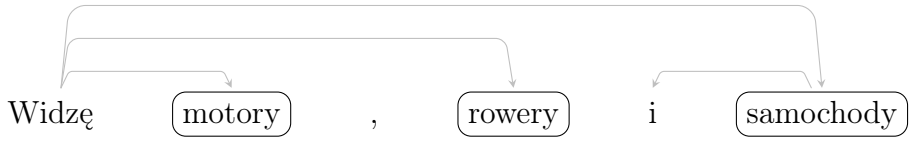
DLM nie jest jedynym czynnikiem kształtującym strukturę syntaktyczną języków naturalnych. Istnieją również inne ograniczenia i preferencje, które mogą wpływać na kolejność słów i długość zależności, m. in. wskazana pozycyjność głowy, harmonia języka (Jing et al., 2022), czy preferencje semantyczne. Niektóre 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

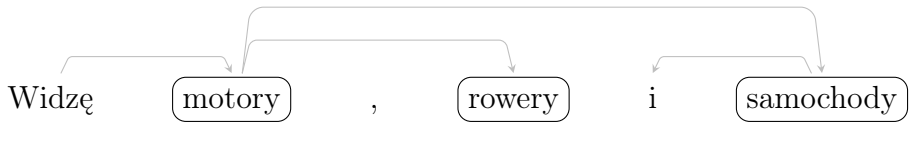
Jeśli chodzi o przedstawienie zależności składniowych w koordynacji (w postaci drzew zależnościowych), to możemy wyróżnić 4 podstawowe podejścia (Przepiórkowski & Woźniak, 2023; Popel et al., 2013), zilustrowane przykładami (7)–(10):

[5] `https://en.wikipedia.org/wiki/Head-directionality_parameter`, dostęp z dn. 08.04.2023

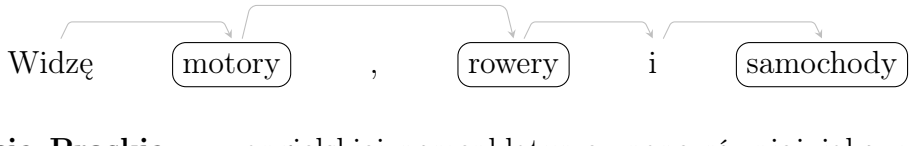
Podejście Londyńskie – jak wskazują Przepiórkowski & Woźniak (2023), podejście to nazwać możemy londyńskim, w duchu nazywania podejść od nazw miast, w których zostały one opublikowane. W angielskiej nomenklaturze możemy znaleźć je również pod nazwą *multi-headed*. Zakłada ono, że nadrzędnik koordynacji jest bezpośrednim nadrzędnikiem każdego z jej członów, a nadrzędnikiem spójnika koordynacji jest jej ostatni człon.

(7) 

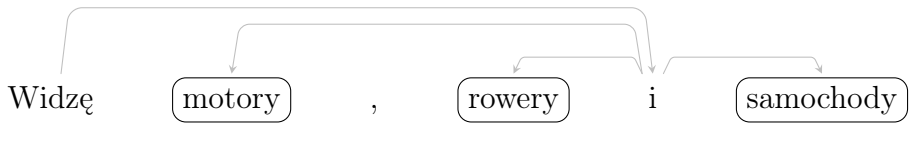
Podejście Stanfordzkie – w angielskiej nomenklaturze określane także mianem *bouquet*. Zakłada ono, że bezpośrednim nadrzędnikiem pierwszego z członów koordynacji jest jej głowa, pierwszy człon równocześnie jest nadrzędnikiem pozostałych członów koordynacji, a ostatni z członów jest nadrzędnikiem spójnika koordynacji.

(8) 

Podejście Moskiewskie – w angielskiej nomenklaturze spotkać się możemy również z określeniem *chain*. Zakłada ono, że każda struktura (tj. człony, głowa oraz spójnik) wewnątrz koordynacji jest bezpośrednim nadrzędnikiem następnej struktury wewnątrz koordynacji.

(9) 

Podejście Praskie – w angielskiej nomenklaturze znane również jako *conjunction-headed*. Zakłada, że głowa koordynacji jest nadrzędnikiem spójnika koordynacji, który to jest nadrzędnikiem każdego z jej członów. To właśnie to podejście wykorzystywane jest w PDB.

(10) 

2.5. Hipotezy

.

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) NKJP1M[6] (14 tysięcy drzew, 217 tysięcy tokenów), (2) równoległe korpusy polsko-angielskie: *Europarl* (Koehn, 2005), *Pelcra Parallel Corpus* (Pęzik et al., 2011), *DGT-Translation Memory* (Steinberger et al., 2012), *OPUS* (Tiedemann, 2012), (3) *CDSCorpus* (Wróblewska & 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). Średnio, zdanie z tego korpusu posiada 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*[7]. Każda z tych części znajduje się w 3 oddzielnych plikach – jeden z nich to zbiór wszystkich zdań w danej części korpusu w pliku z rozszerzeniem '`.txt`', drugi to zbiór tych samych zdań, ale już podzielonych na tokeny oraz z zaznaczeniem zależności, jest on o formacie '`.conll`' i na jego podstawie można wyświetlić zdania te jako drzewa zależnościowe, a trzeci plik to zbiór metadanych o tych zdaniach, zawierający m. in. informacje skąd one pochodzą i jest on o formacie '`.json`'.

[6]NKJP – Narodowy Korpus Języka Polskiego (zob. (Przepiórkowski et al., 2012)). Część tego korpusu, którą znakowano ręcznie, nazywa się NKJP1M

[7]*train*, *dev*, *test* – zwyczajowe nazwy na trzy zbiory danych w przetwarzaniu języka naturalmego, w których część *train* służy do uczenia modelu, część *dev* do jego ewaluacji i *test* do ostatecznej oceny.

3.2. Preprocessing danych

Preprocessing danych robię w języku Python i podzieliłem go na cztery osobne pliki, które znajdują się w Załączniku A. Najpierw wczytuję opisane wyżej dane 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. Następnie dla każdej koordynacji tworzę wiersz w tabeli, w którym umieszczam kolejno: dla koordynacji, które mają nadrzędnik – pozycję nadrzędnika[i], słowo będące nadrzędnikiem[ii], pełny tag[iii][8] nadrzędnika, skrócony tag nadrzędnika[iv], informacje morfosyntaktyczne o nadrzędniku[v], 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 etykietę koordynacji[vi], spójnik współrzędny[vii], tag spójnika[viii], liczbę koniunktów[ix], oraz następujące informację o pierwszym i ostatnim członie koordynacji; pełny człon[x; xxi], człon podzielony na sylaby[xi; xxii][9], głowa tego członu[xii; xxiii], pełny[xiii; xxiv] oraz skrócony[xiv; xxv] tag głowy członu, informacje morfosyntaktyczne o głowie członu[xv; xxvi], liczbę słow danego członu[xvi; xxvii], liczbę jego tokenów[xvii; xxviii], liczbę jego sylab[xviii; xxix], liczbę jego znaków (wliczając spacje)[xix; xxx] oraz informację o tym, czy jest on ciągły, tj. czy wszystkie jego tokeny występują kolejno po sobie, czy między nimi znajduję się jakiś token niebędący częścią tego członu[xx; xxxi]. Na sam koniec dodaję do tabeli całe zdanie, w którym występuje koordynacja[xxxii], jego identyfikator[xxxiii] oraz informację o tym, czy jest ono w zbiorze treningowym, walidacyjnym czy testowym[xxxiv].

3.3. Dane po preprocessingu

Dane po preprocessingu zawarte są w Załączniku B i mają następujący format:

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

governor.position[i]	**governor.word**[ii]	**governor.tag**[iii]	**governor.pos**[iv]
R 0	ptak	subst:sg:nom:m2	subst

[8] tag – oznaczenie danej części mowy, tutaj wraz z jej odmianą

[9] Aby 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.

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 pochyla głowę w stronę leżącego obok okruszka.
Boguś mieszka tu i pracuje.

sent.id[xxxiii]	sent.file[xxxiv]
CDScorpus_6721_B#1673	test
200-2-000000212_morph_9.61-s#6421	test

Zdania te, zapisane zgodnie z poprzednimi przykładami, wyglądałyby następująco:

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

b. Boguś [mieszka] tu [**i** pracuje][10].

[10]Jak widać w tym przykładzie, między członami, poza spójnikiem, występuje także wyraz spoza koordynacji – *tu*. Według PDB jest on podrzędny względem spójnika *i*, ale relacją *adjunct_ locat*, a nie *conj*, zatem nie jest on częścią koordynacji, dlatego rozbiłem nawias kwadratowy na dwie części.

W tabeli po preprocessingu znajduje się łącznie 13247 koordynacji, w tym w 3828 nie występuje nadrzędnik, w 7730 występuje on po lewej stronie, w 44 pomiędzy członami, a w 2045 po prawej stronie. Koordynacje zagnieżdżone 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.

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*.

Rozdział 4

Analiza statystyczna

Tekst rozdziału

4.1. Hipoteza, metody

Tekst sekcji

4.2. Wyniki analizy statystycznej

Tekst sekcji

Rozdział 5

Dyskusja wyników

Tekst rozdziału

5.1. Podsumowanie wyników badań

Tekst sekcji

5.2. Interpretacja wyników

Tekst sekcji

5.3. Przegląd literatury

Tekst sekcji

Rozdział 6

Zakończenie

Tekst rozdziału

6.1. Podsumowanie pracy i wnioski

Tekst sekcji

6.2. Perspektywy dalszych badań

Tekst sekcji

Bibliografia

Covington, M.A. (1984). *Syntactic theory in the high Middle Ages: Modistic models of sentence structure* (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Futrell, R., Levy R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language 96*(2), 371–412.

Haspelmath, M. (2007). Coordination. *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.*

Jing, Y., Blasi, D., & Bickel, B. (2022). Dependency Length Minimization and its limits: A possible fole for a probabilistic version of the Final-Over-Final condition. *Language 98*(3). https://doi.org/10.1353/lan.2022.0013.

Koehn, P. (2005). Europarl: A parralel corpus for statistical machine translation. *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. *Technical report*, ESSLI2002.

de Marneffe, M.-C. & Nivre, J. (2019). Dependency Grammar. *Annual Review of Linguistics 5*, 197–218. https://doi.org/10.1146/annurev-linguistics-011718-011842

Mel'čuk, I.A. (1988). *Dependency syntax: theory and practice.* SUNY Press.

Pedersen, M., Eades, D. Amin, S. K. & Prakash, L. (2004). Relative Clauses in Hindi and Arabic: A Paninian Dependency Grammar Analysis. *Proceedings of the Workshop on Recent Advances in Dependency Grammar*, 9–16. Geneva, COLING

Pęzik, P., Ogrodniczuk, M. & Przepiórkowski, A. (2011). Parralel and spoken corpora in an open repository of Polish language resources. *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 511–515. http://nlp.ipipan.waw.pl/Bib/pez:ogr:prz:11.pdf

Przepiórkowski, A., Bańko, M., Górski, R. & Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa. https://bcpw.bg.pw.edu.pl/dlibra/publication/4691/edition/4582/content

Przepiórkowski, A. (2017). *Argumenty i modyfikatory w gramatyce i w słowniku*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa. https://wuw.pl/data/include/cms/Argumenty_modyfikatory_Przepiorkowski_Adam_2017.pdf

Przepiórkowski, A. & Patejuk, A. (2020). From Lexical Functional Grammar to enhanced Universal Dependencies. *Lang Resources & Evaluation 54*, 185–221. https://doi.org/10.1007/s10579-018-9433-z

Przepiórkowski, A. & Woźniak, M. (2023). Conjunct lengths in English, Dependency Length Minimization, and dependency structure of coordination, [Manuskrypt zgłoszony do publikacji]

Popel, M., Mareček, D., Štěpánek, J., Zeman, D. & Žabokrtský, Z. (2013). Coordination structures in dependency treebanks. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 517–527. Sofia, Bułgaria. https://aclanthology.org/P13-1051.pdf

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 Languages. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 454–459. http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf

Tesnière, L. (1959). *Éléments de syntaxe structurale*. C. Klincksieck.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2214–2218. http://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]. Warszawa. http://nlp.ipipan.waw.pl/Bib/wro:14.pdf

Wróblewska, A. & Krasnowska-Kieraś, K. (2017). Polish evaluation dataset for compositional distributional semantic models. *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics 1, 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. *Proceedings of the 12th Language Resources and Evaluation Conference*, 5308–5315, Marsylia, Francja. European Language Resources Assosiation. http://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 tabeli danych po preprocessingu w formacie „.csv”: https://github.com/kvmilos/PracaLicencjacka/blob/master/tabela.csv

C – link do pliku z analizą danych: https://github.com/kvmilos/PracaLicencjacka/blob/master/analizy/r.R