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. Celem pracy jest sprawdzenie hipotez na ten temat oraz przedstawienie dodatkowych analiz. Ma ona charakter empiryczny i opiera się na danych pochodzących z korpusu Polish Dependency Bank.

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 o różnych długościach w sposób, by zmniejszyć odległość zarówno między nimi samymi, jak i między nimi a innymi elementami zdania, w koordynacjach w języku polskim. Koordynacja to

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

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 vs. 5). Szybko pokazuję, że pierwsza z hipotez zachodzi w większości przypadków, więc następnie omawiam wpływ 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,

[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. ',', '-', '?')

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

Zacznę od przedstawienia pojęcia koordynacji. Koordynacja to Elementami koordynacji mogą być zarówno pojedyncze słowa (2a), frazy (2b), jak i całe zdania (2c):

(2) 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.

Człony koordynacji nazywamy koniunktami, to co je łączy – spójnikiem współrzędnym (w przykładach w tej pracy jest on ilustrowany pogrubionym tekstem), a wyraz nadrzędny względem obu członów – nadrzędnikiem koordynacji (w przykładach ilustrowany kursywą). Jak widać w (2c) nie zawsze istnieje nadrzędnik koordynacji. W podanych wyżej przykładach koniunktami są: (2a) – Ania, Julia; (2b) – Wesoła Marysia, smutny Janek; (2c) – zjadł obiad, poszła spać.

Ze względów semantycznych zwykle wyróżnia się cztery rodzaje koordynacji: koordynacje koniunkcyjne (3a), koordynacje dysjunkcyjne (3b), koordynacje adwersatywne (3c) oraz koordynacje kauzalne (3d) (Haspelmath, 2007). Każde z nich używają różnych zestawów spójników, które łączą koniunkty. W koordynacjach koniunkcyjnych człony łączą m. in. spójniki *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 używamy m. in. spójników *ale*, *lecz*, *zaś*, *natomiast*, *jednak*, a w koordynacjach kauzalnych – *bo*, *bowiem*.

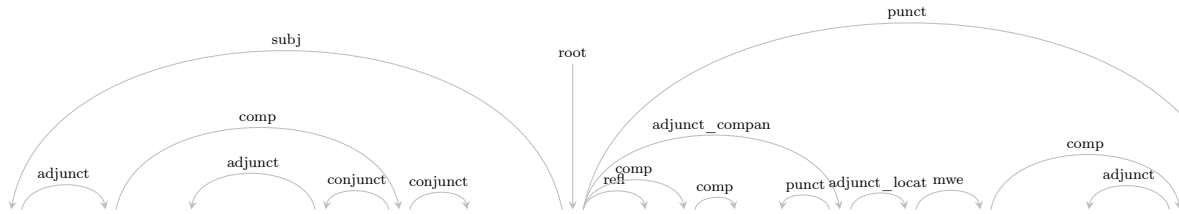
(3) a. Marta *zjadła* [jabłko **i** gruszkę].

b. Ona miała [szesnaście **lub** siedemnaście] *lat*.

c. *Byli* [ładni, **ale** głupi].

d. [Nie zrobiłem pracy domowej, **bo** nie chciałem].

2.2. Zarys teorii zależności składniowej

(4) 

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). W XX wieku teoria ta mocno rozwinęła się zwłaszcza w lingwistyce klasycznej i słowiańskiej (Mel'čuk, 1988). Tesnière (1959) podjął próvę stworzenia kompleksowej teorii gramatyki, w której to wszystko byłoby oparte na zależnościach. Przedstawił on jej potencjał do uchwycenia podobieństw, jak i różnic między językami.

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 czy analiza sentymentu, 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) oraz cały czas powstają nowe, także w innych językach.

2.3. Minimalizacja długości zależności

Minimalizacja długości zależności (DLM) to zasada, według której języki naturalne dążą do zmniejszania odległości między słowami zależnymi syntaktycznie. Zasada ta jest odnotowywana w lingwistyce już od długiego czasu i pozwala nam na bardziej efektywne analizowanie i generowanie języka naturalnego. W ciągu ostatnich 20 lat hipoteza

o nacisku na DLM została wykorzystana do wyjaśnienia wielu z najbardziej uniwersalnych właściwości języków (Futrell et al., 2015). Jak twierdzą Hawkins (1994) i Futrell et al. (2020), rozróżniamy jej występowanie na poziom gramatyczny, jak i codzienne użycie języka. Liu (2008) twierdzi, że długości zależności mogłyby być wskaźnikiem trudności danego języka, sugerując przy tym pewną uniwersalność stosowania DLM w celu łatwiejszego zrozumienia mowy i pisma.

Według DLM, jeśli oba ustawienia członów koordynacji binarnych (jeden człon raz z lewej strony, raz z prawej) są gramatycznie poprawne, to bliżej głowy znajdzie się krótszy człon. W przykładach (5a–b) oba ustawienia wydają się brzmieć naturalnie, jednak gdy wydłużymy jeden z członów – w (5c–d), zauważymy, że bardziej naturalne będzie ustawienie krótszego członu z lewej strony.

(5) a. *Nie ma* [Kamila **i** Julii].

b. *Nie ma* [Julii **i** Kamila].

c. *Nie ma* [wiecznie spóźnionej Julii **i** Kamila].

d. *Nie ma* [Kamila **i** wiecznie spóźnionej Julii].

(6) a. I gave <a book> <to John> .
Dałem[1] <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 przeczytałem[1] od lat> .

za: Przepiórkowski & Woźniak (2023)

Jednym ze sposobów na badanie DLM jest tworzenie sztucznych języków losowych (w których kolejność słów, czy relacje zależności są losowo dobierane) oraz porównywanie długości zależności w tych językach z długościami zależności w językach naturalnych. Badania wykazały, że języki naturalne mają istotnie krótsze długości zależności niż sztucznie wygenerowane wartości dla przykładowych losowych języków (Futrell et al., 2015), co sugeruje, że istnieje uniwersalna tendencja u ludzi do wykorzystywania DLM.

DLM jest również powiązana z innymi właściwościami języków naturalnych, między innymi z pozycyjnością głowy. Oznacza ona kierunek występowania nadrzędnika frazy względem jej dopełnienia. Badania wykazały, że istnieje związek między pozycyjnością głowy a długością zależności, przy czym języki o pozycyjności głowy na końcu

[1]Wyrażenia *I gave* oraz *I've read* można przetłumaczyć również jako odpowiednio *dałam* i *przeczytałam*, nie zawierają one informacji o rodzaju; dla uproszczenia wszystkie przykłady tłumaczę używając rodzaju męskiego

frazy(zdania?) mają krótszą długość zależności niż języki o pozycyjności głowy na początku frazy(zdania?) (Futrell et al., 2015).

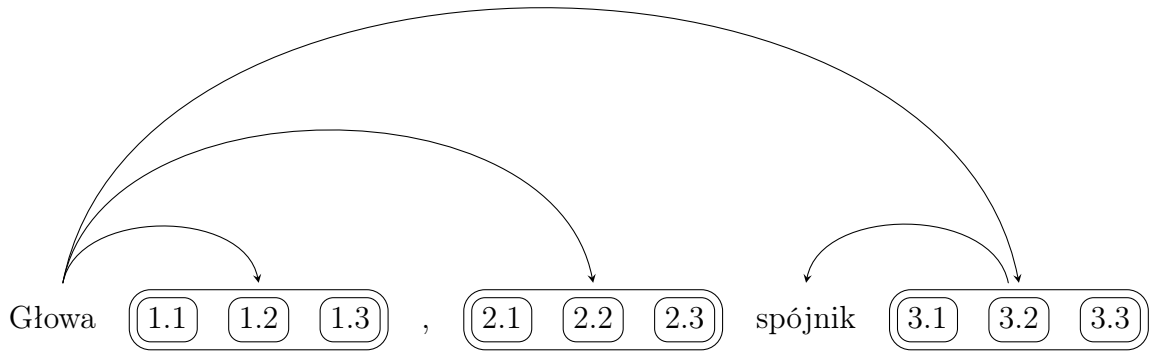
DLM nie jest jednak jedynym czynnikiem kształtującym strukturę syntaktyczną języków naturalnych. Istnieją również inne ograniczenia i preferencje, takie jak harmonia języka (Jing et al., 2022), czy preferencje semantyczne, które mogą wpływać na kolejność słów i długość zależności. Niektóre z tych czynników mogą być sprzeczne lub komplementarne względem DLM. Dlatego DLM należy rozumieć jako jeden z wielu czynników wpływających na organizację języka naturalnego.

2.4. Różne reprezentacje koordynacji

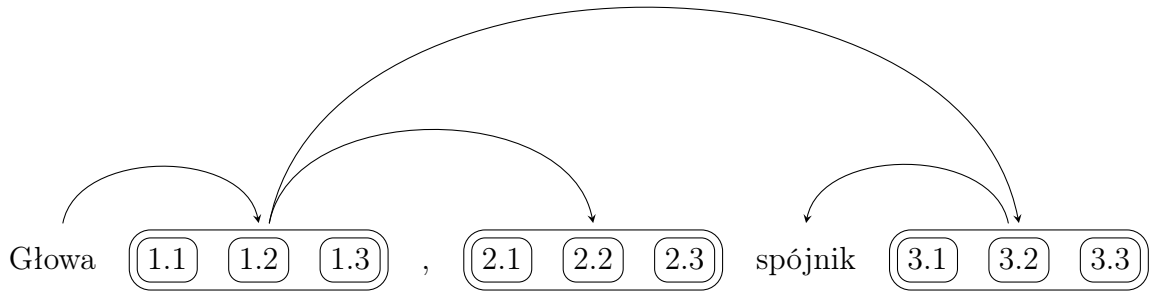
Jeśli chodzi o reprezentacje koordynacji w postaci drzew zależnościowych, to możemy wyróżnić 4 podstawowe podejścia (Przepiórkowski & Woźniak, 2023; Popel et al., 2013):

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

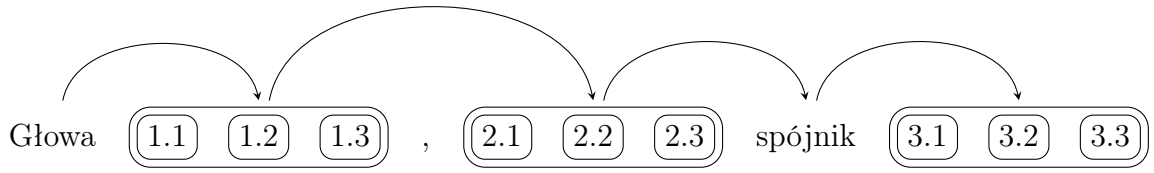
(7)



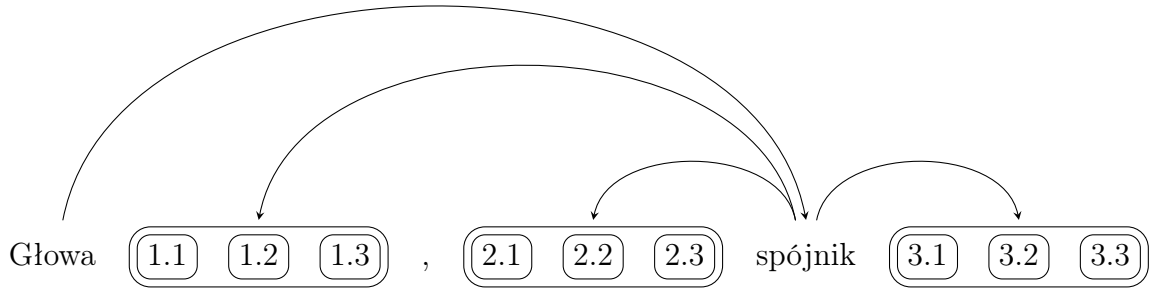
Podejście Stanfordzkie – w angielskiej nomenklaturze określane także mianem *bouquet*. Zakłada ono, że bezpośrednim nadrzędnikiem pierwszego z członów koordynacji jest jej głowa, pierwszy człon równocześnie jest nadrzędnikiem pozostałych członów koordynacji, a ostatni z członów jest nadrzędnikiem spójnika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

Różnice między tymi podejściami możemy badać wraz z DLM, ponieważ każde z nich może mieć inne długości zależności dla tego samego zdania. Podejścia zilustrowane przykładami (8, 9) sugerują, że niezależnie od tego, czy głowa koordynacji jest z lewej, czy z prawej strony, zgodnie z zasadą DLM pierwszy człon powinien być krótszy – skróciłoby to sumę długości wszystkich zależności w zdaniu. Przy podejściu (10) pozycja głowy koordynacji nie wpływa na długość zależności, zatem nie ma żadnych powodów, aby prawy człon stawał się krótszy niż lewy. Zakładając podejście (7) możemy

zauważyć, że pozycja głowy koordynacji z prawej strony, zgodnie z DLM może skrócić długości zależności, ustawiając krótszy człon po spójniku koordynacji, co byłoby nielogiczne dla pozostałych podejść.

2.5. Hipotezy

Rozdział 3

Dane

Tekst rozdziału

3.1. Polish Dependency Bank

Tekst sekcji

3.2. Preprocessing danych

Tekst sekcji

3.3. Dane po preprocessingu

Tekst sekcji

Rozdział 4

Analiza statystyczna

Tekst rozdziału

4.1. Analiza statystycznaw

Tekst sekcji

4.2. Testowanie hipotez

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. In T. Shopen (Ed.), *Language typology and syntactic description, Volume II: Complex constructions* (pp. 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.

Kruijff, G.-J. M. (2002). Formal and computational aspects of dependency grammar: History and development of dg. *Technical report*, ESSLI2002.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2), 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

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.

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)*, pp. 517–527. Sofia, Bulgaria

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

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