

Schemat anotacji struktur zależnościowych

Alina Wróblewska

Instytut Podstaw Informatyki Polskiej Akademii Nauk

Warszawa, 17 październik 2011



- 1 Struktury zależnościowe
- 2 Polskie typy relacji zależnościowych
- 3 Anotacja zjawisk składniowych
- 4 Eksperyment

Struktura zależnościowa jest anotowana jako graf skierowany:

- wierzchołki reprezentują tokeny w zdaniu plus korzeń ROOT,
- krawędzie skierowane reprezentują binarne relacje zależnościowe pomiędzy tokenami,
- jeden z tokenów jest *głową* (nadrzędnikiem) relacji zależnościowej, a drugi jest podrzędnikiem,
- etykiety na krawędziach wskazują typ relacji zależnościowej,
- wierzchołki mają przypisany indeks odpowiadający pozycji tokena w zdaniu; ROOT ma zawsze indeks 0.

Przykład: Struktura zależnościowa



- Struktury zależnościowe są zakodowane w formacie CoNLL
- Wybór formatu został poddyktowany przez dostępne systemy parsujące i akceptowane przez nie formaty.
- Zakodowane informacje:
 - 1 ID – indeks tokena,
 - 2 FORM – forma ortograficzna lub znak interpunkcyjny,
 - 3 LEMMA – forma podstawowa,
 - 4 CPOSTAG – ‘ogólna’ część mowy,
 - 5 POSTAG – ‘szczegółowa’ część mowy,
 - 6 FEATS – zbiór cech morfosyntaktycznych,
 - 7 HEAD – indeks nadrzędnika danego tokenu,
 - 8 DEPREL – typ relacji zależnościowej.

Przykład: format kodowania struktur zależnościowych

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL
1	Na	na	prep	prep	acc	4	adj
2	wszelki	wszelki	adj	adj	sglacc m3 pos	3	adj
3	wypadek	wypadek	subst	subst	sglacc m3	1	comp
4	weźmiemy	wziąć	verb	fin	pl pri perf	0	pred
5	cię	ty	subst	ppron12	sglacc m1 sec nakc	4	obj
6	za	za	prep	prep	acc	4	comp
7	wielbłąda	wielbłąd	subst	subst	sglacc m2	6	comp
8	.	.	interp	interp	_	4	punct

- 1 Struktury zależnościowe
- 2 Polskie typy relacji zależnościowych
- 3 Anotacja zjawisk składniowych
- 4 Eksperyment

- Wybór odpowiedniego typu zależności jest istotny, żeby właściwie zaanotować strukturę zależnościową.
- Precyzyjna definicja typów zależnościowych ułatwia wybór.
- Rozróżniamy następujące typy zależności:
 - Dopełnienia (ang. *complements*),
 - Nie-dopełnienia (ang. *non-complements*).

- 1 *comp* – dopełnienie (ang. *complement*)
 - Dopełnienie przymiottnikowe
 - Dopełnienie przysłówkowe
 - Dopełnienie rzeczownikowe
 - Dopełnienie przyimkowe

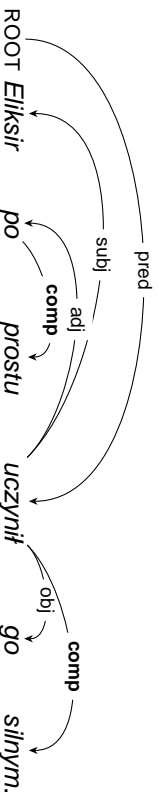
- **1** *comp* – dopełnienie (ang. *complement*)
- **Dopełnienie przymiotnikowe** z nadrzędnikiem w postaci:
 - formy czasownikowej np., *uczynić kogoś silnym*,
 - przyimka np., *z wolna, od dawna, po prostu*.
- **Dopełnienie przysłówkowe**
- **Dopełnienie rzeczownikowe**
- **Dopełnienie przyimkowe**

- **1** *comp* – dopełnienie (ang. *complement*)
 - **Dopełnienie przymiotnikowe**
 - **Dopełnienie przysłówkowe z nadrzędnikiem w postaci:**
 - czasownika,
 - przyimka np., *na pewno, na zewnątrz.*
 - **Dopełnienie rzeczownikowe**
 - **Dopełnienie przyimkowe**

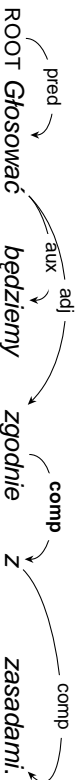
- *comp* – dopełnienie (ang. *complement*)
- **Dopełnienie przymiotnikowe**
- **Dopełnienie przysłówkowe**
- **Dopełnienie rzeczownikowe z nadrzędnikiem w postaci:**
 - przymiotnika, np. *pełny mleka, przeciwny globalizacji,*
 - przyimka,
 - liczebnika,
 - formy czasownikowej
 - nie podlega zmianie do funkcji podmiotu podczas pasytywizacji,
 - przyjmuje role Location, Instrument, Goal (ale nie Recipient, Experiencer, które są zarezerwowane dla *obj_th*).
- **Dopełnienie przyimkowe**

- *comp* – dopełnienie (ang. *complement*)
 - **Dopełnienie przymiotnikowe**
 - **Dopełnienie przysłówkowe**
 - **Dopełnienie rzeczownikowe**
 - **Dopełnienie przymiolkowe z nadrzędnikiem w postaci:**
 - formy czasownikowej,
 - przymiotnika np., *zdolny do*,
 - przysłówka np., *dopiero w, własnie przez, odpowiednio do*.
- Wiele kombinacji przysłówka z frazą przymiolkową tworzy przymyki wtórne (Milewska, 2003) np., *daleko od, wraz z, zgodnie z*.

- Dopełnienie przyimiotnikowe przyimka i czasownika

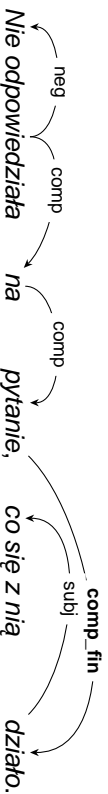


- Dopełnienie przyimkowe przyimiotnika



2 *comp_fin* – dopełnienie zdaniowe (oznajmujące, pytające, wykrzyknikowe) z nadrzędnikiem w postaci:

- formy czasownikowej,
- spójnika podrzędnego,
- rzeczownika.



- 3 *comp_inf* – dopełnienie bezokolicznikowe z nadrzędnikiem w postaci:

- spójnika podrzędnego,
- przymiotnika np., *gotowy coś zrobić*,
- rzeczownika np., (*mieć*) *prawo*, (*mieć*) *szansę coś zrobić*,
- czasownika np., *chcieć*, *kazać*, *zabronić coś zrobić*,
- quasi-czasownika np., *można*, *trzeba coś zrobić*.



4 *complm* – ang. *complementizer*

- jest realizowany jako spójnik podrzędny np., że, iż, żeby, aby, by,
- jest nadrzędnikiem finitywnej frazy zdaniowej (*comp_fin*) lub bezkolicznikowej frazy zdaniowej (*comp_inf*),
- jest podrzędnikiem:
 - formy czasownikowej,
 - przyimiotnika,
 - rzeczownika.



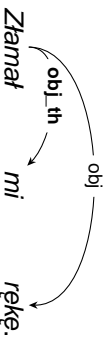
5 *obj* – dopełnienie ‘bliższe’:

- jest realizowane jako fraza rzeczownikowa głównie w bierniku, ale również w dopełniaczu, celowniku lub narzędniku,
- jest podrzędnikiem formy czasownikowej,
- staje się podmiotem mianownikowym w konstrukcjach strony biernej.



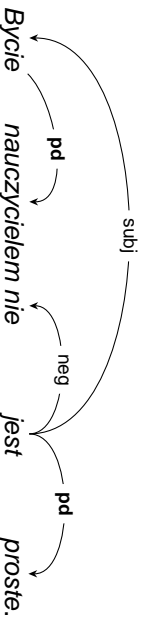
6 *obj_th* – dopełnienie celownikowe:

- jest realizowane jako fraza rzeczownikowa w celowniku,
- jest podrzędnikiem formy czasownikowej,
- wypełnia role semantyczne Recipient, Experienter, Beneficiary,
- nie może zostać zmienione w dopełnienie bliższe w wyniku alteracji struktury argumentów ('dative shift', Kibort, 2008).



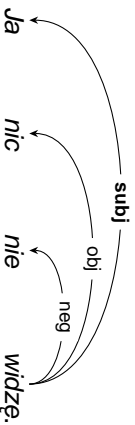
7 *pd* – dopełnienie predykatywne (orzecznik):

- jest realizowane najczęściej jako fraza nominalna w mianowniku lub narzędniku, przymiottnik,
- jest podrzędnikiem czasownika *być* ang. 'copula verb' lub *stać się*, *nazywać się*, itp.



8 *subj* – podmiot:

- jest realizowany jako fraza nominalna, fraza zdaniowa, przymiotnik, liczebnik, itp. lub zaimek 'pro-drop', który nie jest zaanotowany w strukturze zależnościowej,
- jest podrzędnikiem orzeczenia.



● *adj* – okolicznik

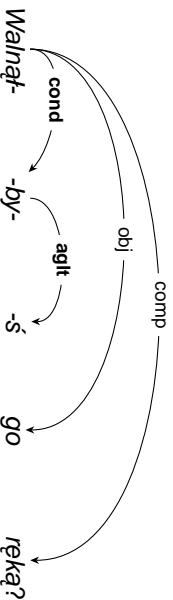
- przymiotnikowy podrzędnik rzeczownika lub liczebnika,
- przysłówkowy podrzędnik czasownika, przysłówka, przymiotnika,
- rzeczownikowy atrybut z rzeczownikowym nadrzędnikiem,
- fraza przyimkowa,
- podrzędna fraza zdaniowa z rzeczownikowym, liczebnikowym, czasownikowym nadrzędnikiem,
- warunkowa podrzędna fraza zdaniowa z nadrzędnikiem realizowanym przez orzeczenie głównej frazy zdaniowej,
- etc.

10 *aglt* – enklityka:

- jest realizowana jako ‘mobilny’ przyrostek, np. -em, -m, -eś, -ś, -śmy, -ście,
- jej nadrzędnikiem jest czasownik (nawet jeśli jest doczeptiona do spójnika podrzędnego) lub warunkowa klityka *by*.

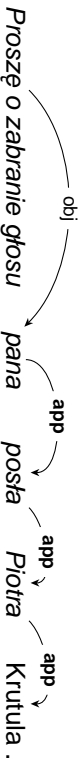
11 *cond* – klityka warunkowa *by*:

- jej nadrzędnikiem jest forma czasownikowa,
- może być przyłączona do czasownika lub może pojawiać się w ‘dowolnym’ miejscu w zdaniu.



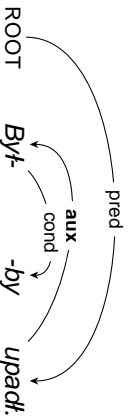
12 *app* – apozycja:

- apozycja i jej nadrzędnik muszą się odnosić do tego samego bytu,
- apozycja (najczęściej NP) jest podrzędnikiem bezpośrednio poprzedzającym ją rzeczownika,
- apozycja (drugi rzeczownik w dwurzeczownikowym złożeniu) jest podrzędnikiem pierwszego rzeczownika w tym złożeniu,
- nazwisko jest anotowane jako podrzędnik (apozycja) imienia.

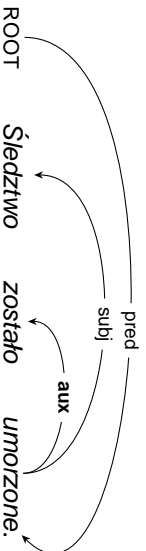


13 *aux* – czasownik posikowy:

- *być* występuje w analitycznym czasie przyszłym i w analitycznym czasie przeszłym konstrukcji warunkowych,
- jego nadrzędnikiem jest głównie forma czasownika (imiesłów przyimiotnikowy bierny, bezokolicznik).



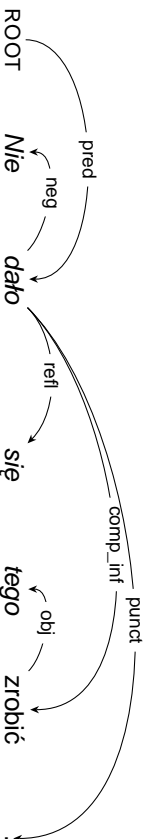
- *być* lub *zostać* występuje w konstrukcjach strony biernej,
- jego nadrzędnikiem jest imiesłów przyimiotnikowy bierny.



- 14 *neg* – partykuła przecząca ‘nie’
Jej nadrzędnikiem jest forma czasownikowa.
- 15 *pred* – orzeczenie zdaniowe lub główny rzeczownik we frazie rzeczownikowej
- jest realizowany jako forma czasownikowa (czasownik finitywny, forma bezosobowa ‘-no/-to’, bezokolicznik) lub główny rzeczownik w niezależnie zaanotowanej frazie rzeczownikowej,
 - jego nadrzędnikiem jest korzeń ROOT.
- 16 *refl* – znacznik zwrotności ‘się’
Jego nadrzędnikiem jest forma czasownikowa.

17 *punct* – znak interpunkcyjny

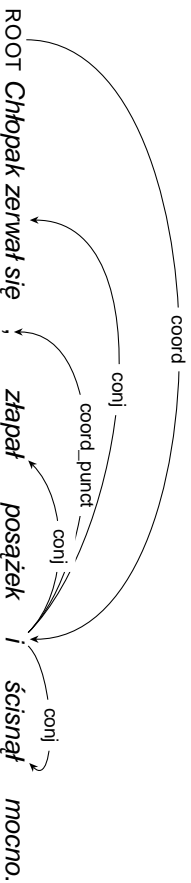
Jego nadrzędnikiem jest element, którego granice są wyznaczone przez znak interpunkcyjny.



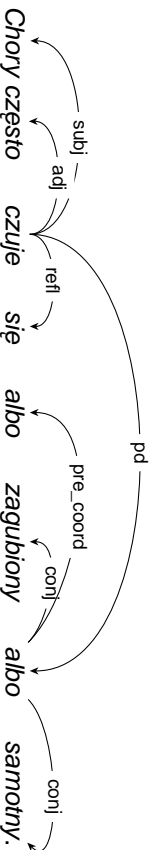
- 1 Struktury zależnościowe
- 2 Polskie typy relacji zależnościowych
- 3 Anotacja zjawisk składniowych
- 4 Eksperyment

Koordinacja: *conjunct*, *coord*, *coord_punct*

- 18 *conjunct* – element skoordynowany, którego nadrzędnikiem jest spójnik współrzędny.
- 19 *coord* – spójnik współrzędny skoordynujący dwie frazy zdaniowe, którego nadrzędnikiem jest ROOT.
- 20 *coord_punct* – interpunkcyjny spójnik współrzędny, np. przecinek, dwukroppek, którego nadrzędnikiem jest ROOT lub spójnik współrzędny.



- 21 *pre_coord* – pierwszy człon złożonego spójnika współrzędnego,
 np. *albo ... albo...*, *ani ... ani...*, który jest podrzędnikiem drugiego
 członu spójnika.



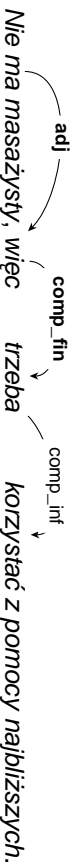
Spójnik koordynujący inne frazy np. rzeczownikowe jest anotowany jako odpowiedni typ zależnościowy, zob. *pd*.

- Orzeczenie zdania głównego jest nadrzędnikiem spójnika podrzędnego – relacja *adj.*
- Spójnik podrzędny jest nadrzędnikiem frazy zdaniowej – relacja *comp_fin* lub *comp_inf*.

Przykład: Okolicznikowe zdanie podrzędne (1)

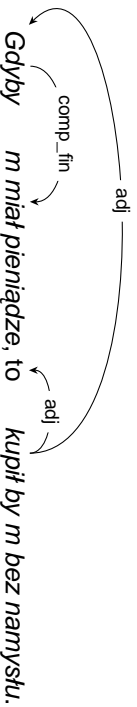


- Spójniki podrzędne: *albowiem, bo, gdyż, przeto, toteż, więc, zatem.*



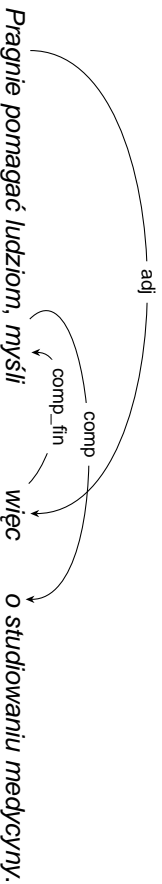
Przykład: Okolicznikowe zdanie podrzędne (2)

- Spójniki podrzędne: *jeśli, gdyby*.
- Opcjonalna parttykuła 'to' wprowadzająca zdanie główne.



Przykład: Okolicznikowe zdanie podrzędne (4)

- Spójniki podrzędne *bowiem*, *przeto*, *więc*, *zatem* nie pojawiają się na początku zdania podrzędnego.
- Powstają 'non-projective' struktury zależnościowe.



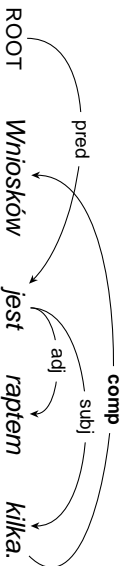
Przykład: Okolicznikowe zdanie podrzędne (5)



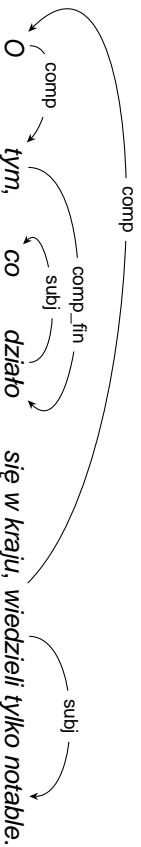
- Dwuczłonowe spójniki podrzędne, np. *mimo że*, *podczas gdy*, drugi człon stanowi nadrzędnik pierwszego, a relacja zależnościowa jest anotowana jako *adj*.

Tego nie wiemy, mimo że zdanie stało się znakiem rozpoznawczym.

- Relacje zależnościowe są anotowane zgodnie ze schematem, nawet jeśli w wyniku tego powstają struktury zależnościowe określane jako 'non-projective'.



- Korelat – zaimek lub zaimek we frazie przyminkowej, który koreluje z podrzędną frazą zdaniową.
- Nie chcemy dodatkowej funkcji zależnościowej, dlatego:
 - zaimek/fraza przyminkowa jest podrzędnikiem orzeczenia,
 - podrzędna fraza zdaniowa jest podrzędnikiem zaimka.



- 1 Struktury zależnościowe
- 2 Polskie typy relacji zależnościowych
- 3 Anotacja zjawisk składniowych
- 4 **Eksperyment**

- Struktury zależnościowe przekonwertowane z drzew składnikowych.
- Informacje wykorzystywane do skonstruowania reguł konwersji:
 - formy ortograficzne, lemmaty, części mowy, cechy morfologiczne,
 - kategorie fraz,
 - typy reguł frazowych stosowanych do tworzenia drzew,
 - zakodowane typy relacji zależnościowych \Rightarrow podmiot,
 - rozróżnienie pomiędzy ‘frazami luźnymi’ i ‘frazami wymaganymi’.
- 7500 struktur zależnościowych (średnio 9,8 tokenów w zdaniu).

- MaltParser (Nivre et al., 2006) – publicznie dostępny system generujący parsery zależnościowe.
- Parser buduje strukturę zależnościową dla zdania, bazując na przejściach (akcjach shift-reduce) przewidywanych przez klasyfikator.
- Algorytm parsowania: stackeager (Nivre, 2009).
- Model cech: FORM, CPOS, POS, FEATS, LEMMA i DEPREL.
- Klasyfikator: trenowany z biblioteką LIBSVM (Chang i Lin, 2001).

- Korpus treningowy (6832 struktur) i korpus testowy (759 zdań).
- Zbiór 17 ręcznie zaanotowanych zdań (16.6 tokenów/zdanie).
- Wyniki ewaluacji:
 - 88,6% LAS¹/91,6% UAS² (część banku zależnościowego) ,
 - 75,2% LAS/78,4% UAS (zbiór 17 zdań z gazet).

¹ LAS (labelled attachment score) – liczba tokenów, którym został przypisany poprawny nadrzędnik i etykieta.

² UAS (unlabelled attachment score) – liczba tokenów, którym został przypisany poprawny nadrzędnik.

- Został zdefiniowany schemat anotacji struktur zależnościowych.
- Można przekonwertować drzewa składnikowe na struktury zależnościowe zaanotowane zgodnie ze schematem.
- Bank struktur zależnościowych stanowi materiał do wytrenowania polskiego parsera zależnościowego.

Dziękuję za uwagę!

- Wynik parsowania zdań z gazety 75,2% LAS.
- Parser nie radzi sobie ze zdaniami z przestawionymi argumentami.
- Struktury takie są niobecne w banku.
- Należy włączyć podobne struktury do banku, żeby zoptymalizować jakość parsowania.

