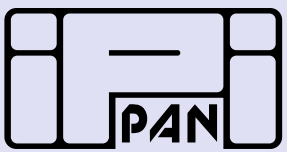
# Automatyczna analiza zależnościowa języka polskiego

Piotr Rybak
Alina Wróblewska




INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 15 kwietnia 2019

# Spis treści

# Parsowanie zależnościowe

- Automatyczna analiza składniowa = parsowanie składniowe.
- Analiza składniowa to ważne zadanie w bardziej zaawansowanych zadaniach NLP.
- Rodzaje analizy składniowej:
  - Parsowanie składnikowe (ang. constituent parsing)
  - **Parsowanie zależnościowe** (ang. dependency parsing)

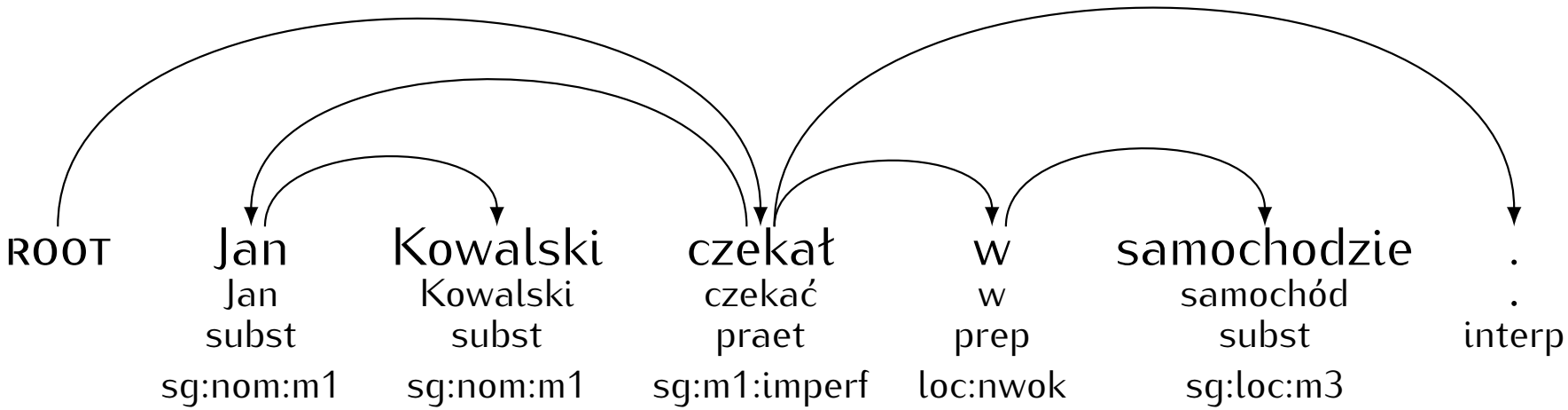
# Struktura zależnościowa

- Struktura zależnościowa to drzewo:
  - każdy wierzchołek ma jedną krawędź wejściową,
  - korzeń ROOT nie ma krawędzi wejściowych i ma jedną krawędź wyjściową.
- Wierzchołki w drzewie odpowiadają segmentom w zdaniu.
- Krawędzie skierowane reprezentują relacje:
  - segment, z którego wychodzi krawędź, jest nadrzędnikiem segmentu, do którego wchodzi dana krawędź,
  - etykieta krawędzi odpowiada funkcji gramatycznej podrzędnika.

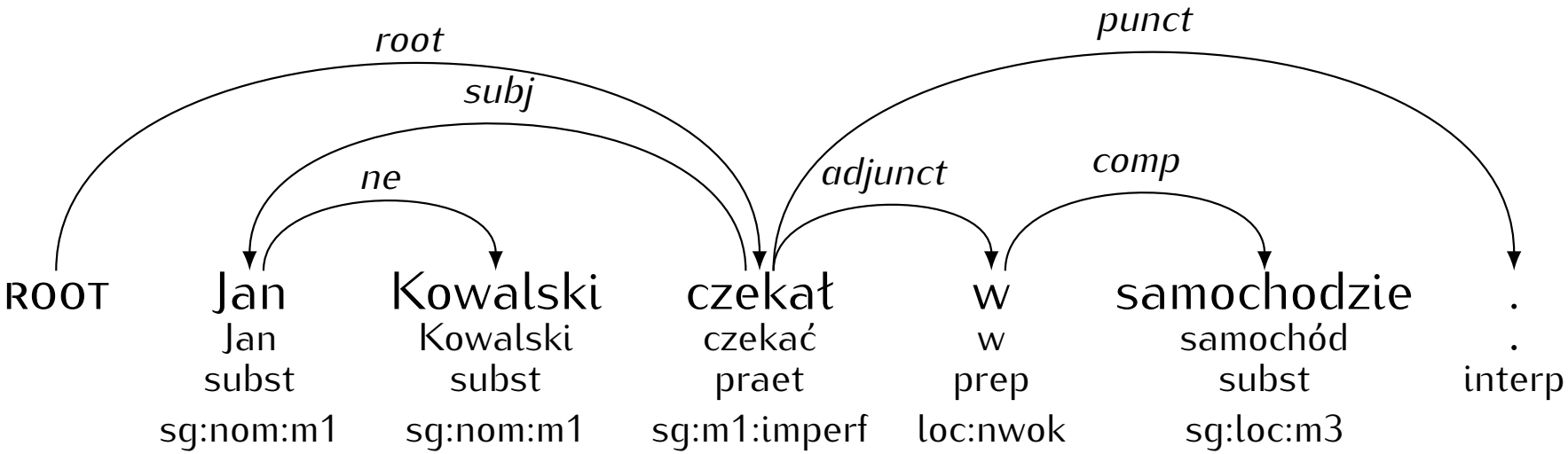
# Struktura zależnościowa – przykład

| | | | | | | |
|---|---|---|---|---|---|---|
| ROOT | Jan | Kowalski | czekał | w | samochodzie | . |
| | Jan | Kowalski | czekać | w | samochód | . |
| | subst | subst | praet | prep | subst | interp |
| | sg:nom:m1 | sg:nom:m1 | sg:m1:imperf | loc:nwok | sg:loc:m3 | |

# Struktura zależnościowa – przykład

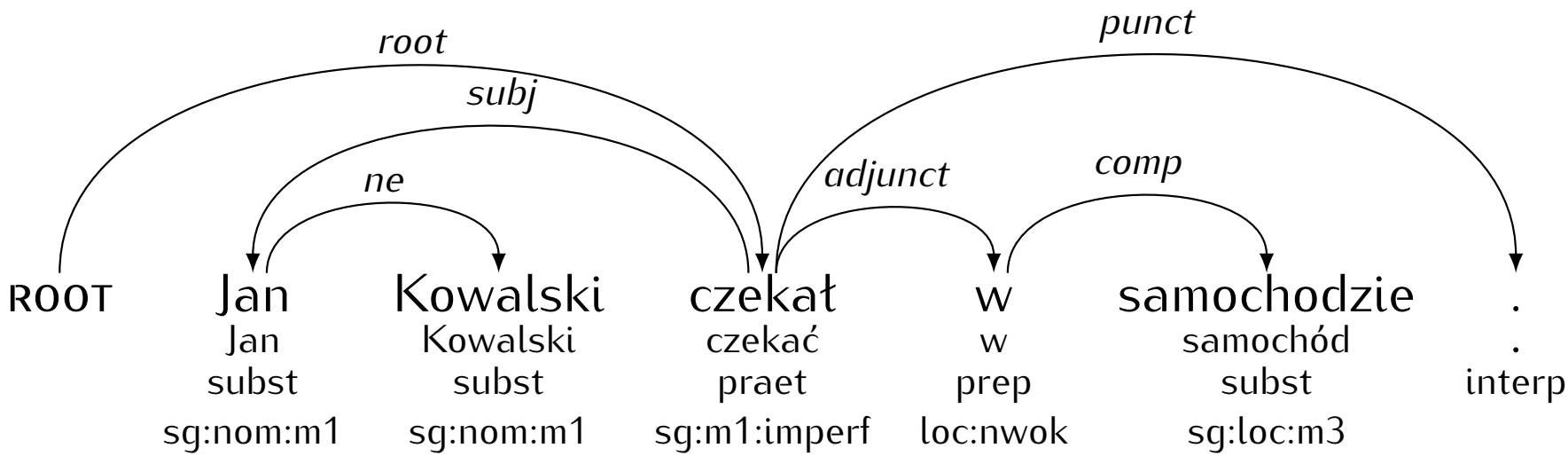


# Struktura zależnościowa – przykład




- *subj* (ang. subject) – podmiot
- *ne* (ang. named entity) – jednostka nazewnicza
- *comp* (ang. complement) – dopełnienie
- *adjunct* – okolicznik lub przydawka
- *punct* (ang. punctuation) – znak interpunkcyjny

# Struktura zależnościowa – przykład




$T = (V, A)$

- $V = \{$ROOT, Jan, Kowalski, czekał, w, samochodzie, .$\}$
- $A = \{$(ROOT, czekał, *root*), (czekał, Jan, *subj*), (Jan, Kowalski, *ne*), (czekał, w, *adjunct*), (w, samochodzie, *comp*), (czekał, ., *punct*)$\}$

# Automatyczna analiza zależnościowa

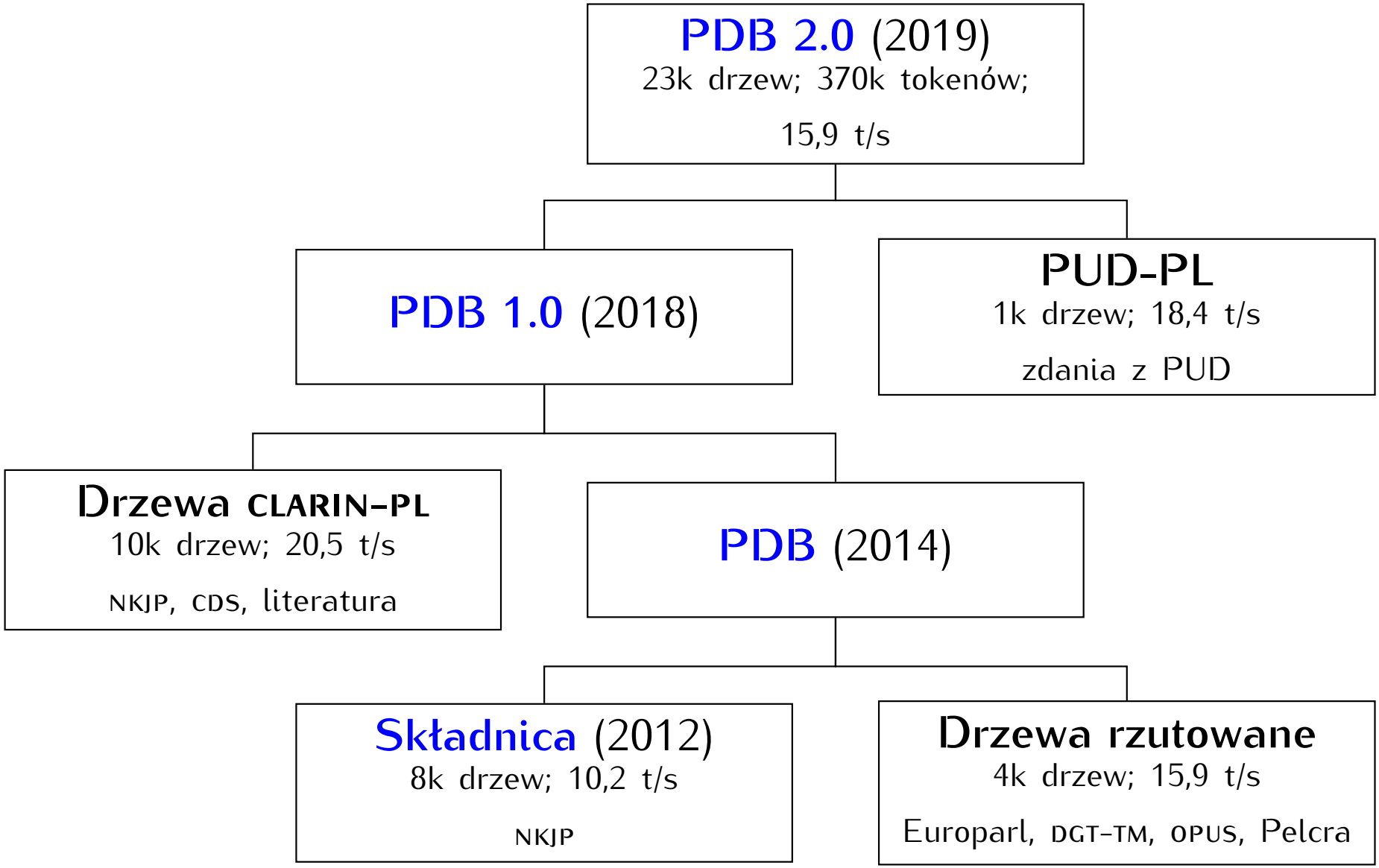
- Parsowanie zależnościowe:
  - analiza składniowo-semantyczna zdania wejściowego,
  - predykcja struktury zależnościowej dla zdania wejściowego.
- Parser zależnościowy (regułowy lub **statystyczny**).
- Dane uczące – duże zbiory zdań zaanotowanych drzewami zależnościowymi.

# Spis treści

# Polish Dependency Bank (PDB)

- Największy bank drzew zależnościowych dla języka polskiego.
- Ma dwie wersje:
  - PDB – standardową, dostosowaną do specyfiki języka polskiego,
  - PDB-UD – uniwersalną, zgodną ze schematem Universal Dependencies (UD).
- Powstawał w IPI PAN w ramach projektów NEKST i CLARIN-PL 2.0, a obecnie jest rozwijany w CLARIN-PL 3.0.

# Rys historyczny



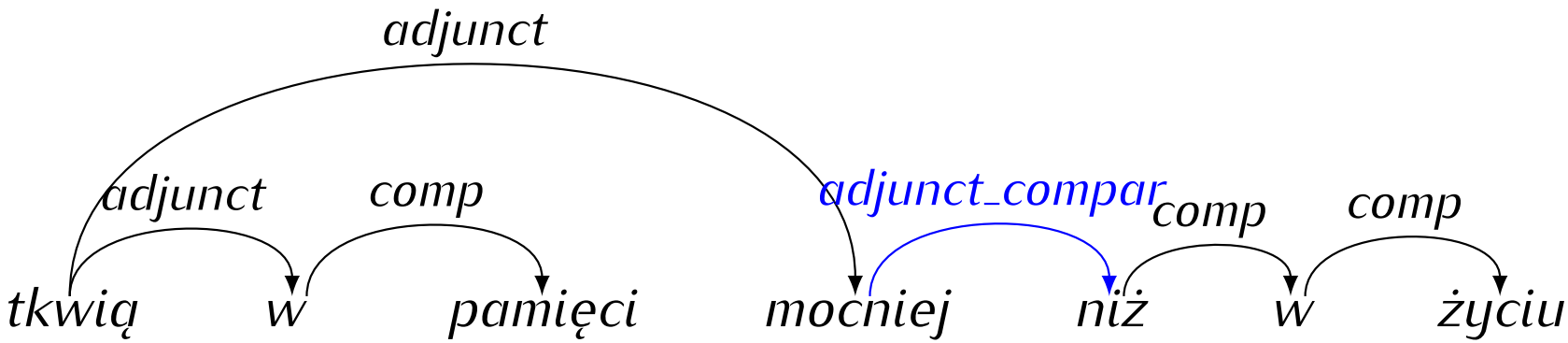

t/s – średnia liczba tokenów w zdaniu

# PDB 2.0

- Wszystkie drzewa w PDB są anotowane ręcznie.
- Procedura anotacji:
  - zdanie zostaje przetworzone przez dwa różne parsery,
  - wygenerowane drzewa zostają poprawione przez dwóch niezależnych lingwistów,
  - superanotator tworzy ostateczne drzewo.
- Nowości w PDB:
  - konstrukcje porównawcze,
  - konstrukcje utożsamiające z JAKO,
  - frazy względne,
  - mowa zależna,
  - wtrącenia i komentarze,
  - role tematyczne.

# Konstrukcje porównawcze w PDB

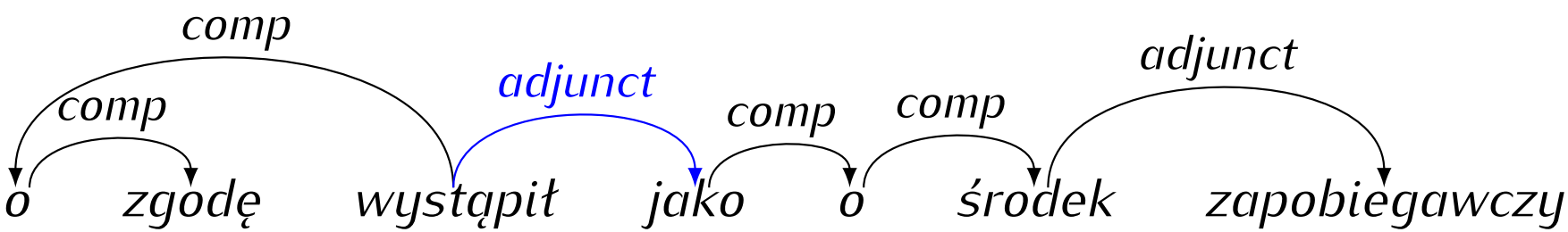
- Dwa typy konstrukcji porównawczych (Bondaruk, 1998):
  - Porównania równościowe, np. *tak ... jak*
  - Porównania nierównościowe z NIŻ
- Porównania są wprowadzane przez spójniki podrzędne, np. JAK, NIŻ, JAKBY, NICZYM.




A. Bondaruk. 1998. Comparison in English and Polish Adjectives: A Syntactic Study, PASE Studies and Monographs (tom 6). Folium, Lublin.

# Konstrukcje utożsamiające z JAKO w PDB

- Konstrukcje z JAKO pełnią funkcję utożsamiającą (Wróblewska i Wieczorek, 2018):
  *zgoda ≈ środek zapobiegawczy.*
- JAKO to predykatywny spójnik podrzędny: parafraza
  *wystąpił o zgodę będącą środkiem zapobiegawczym.*




A. Wróblewska and A. Wieczorek. 2018. Status składniowy *jako* we współczesnej polszczyźnie. Język Polski, XCVIII(3):16–30.

# Role tematyczne

- 28 etykiet odpowiadających tzw. *frame elements* z FrameNet-u (Fillmore i Baker, 2009), np. Theme, Result, Place, Manner, Attitude.
- Role rozszerzają znaczenie rzeczownikowych dopełnień dalszych (*obj_th*, thematically restricted object) oraz okoliczników i niektórych przydawek (*adjunct*).
- Etykiety semantyczne w 11 kolumnie w formacie CoNLL i CoNLL-U!

Ch. J. Fillmore and C. Baker. 2009. A Frames Approach to Semantic Analysis. [w:] B. Heine i H. Narrog (red.), The Oxford Handbook of Linguistic Analysis, s. 313–340. Oxford University Press.

# Universal Dependencies (UD)

- Inicjatywa Universal Dependencies[1] (Nivre i in., 2016):
  - opracowanie schematu anotacji, który jest uniwersalny dla większości języków,
  - stworzenie dużej wielojęzycznej kolekcji banków drzew zależnościowych zaanotowanych zgodnie z uniwersalnym schematem.
- Schemat UD jest obecnie standardem anotowania drzew zależnościowych.
- UD v2.3 ma 129 banków drzew dla 76 języków.

J. Nivre, M-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, Ch. D. Manning, R. T. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty i D. Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. [w:] Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, s. 1659–1666.

[1] `http://universaldependencies.org`

# PDB-UD

- Automatyczna konwersja PDB→PDB-UD
  - Regułowa konwersja części mowy, znaczników morfologicznych i drzew zależnościowych.
  - Poprawianie błędów i niespójności w oryginalnych drzewach PDB.
- `http://git.nlp.ipipan.waw.pl/alina/PDBUD`
- Rozszerzona i poprawiona wersja polskiego banku UD-SZ.
- Plan: PDB-UD zastąpi bank UD-SZ w kolekcji UD.

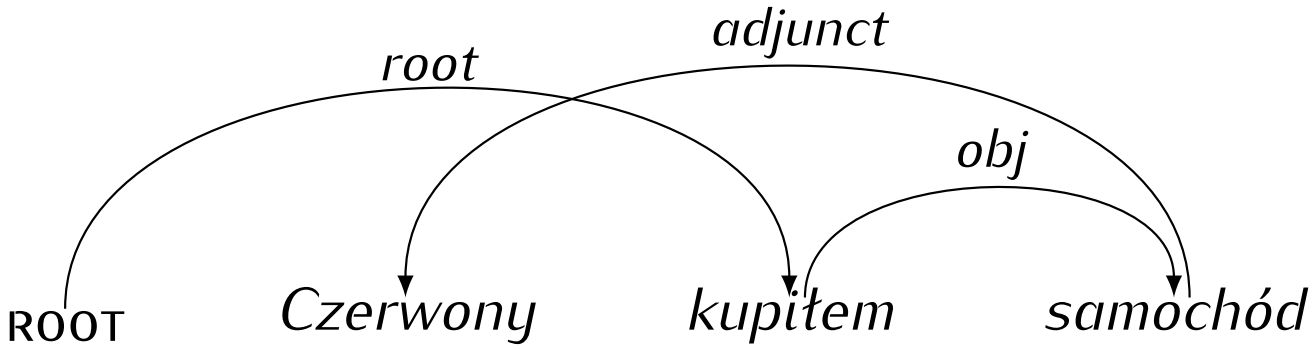
# Porównanie PDB i PDB-UD

| | PDB (CoNLL) | PDB-UD (CoNLL-U) |
|---|---|---|
| # zdań | 23 208 | 23 208 |
| # segmentów | 369 795 | 369 795 |
| # t/s | 15,93 | 15,93 |
| # typy zależności | 28 | 31 (58 z podtypami) |

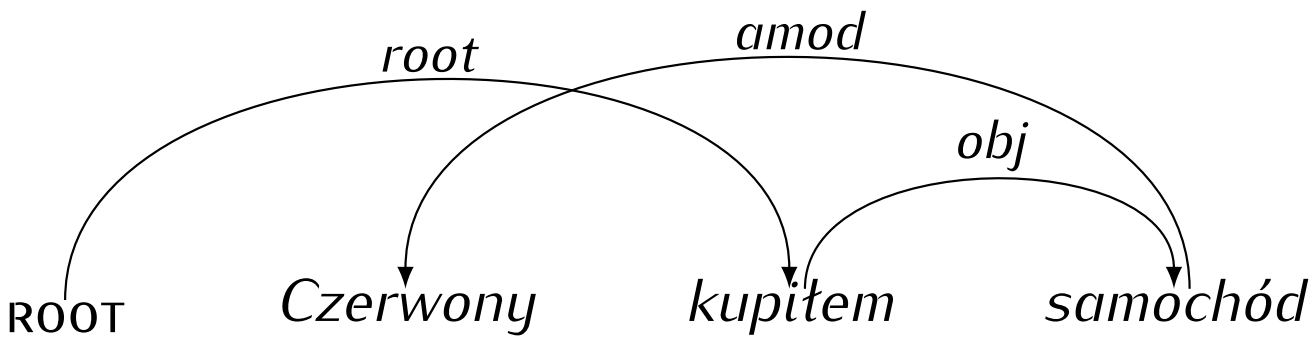
# Porównanie PDB i PDB-UD

| | PDB (CoNLL) | | PDB-UD (CoNLL-U) | |
|---|---|---|---|---|
| # zdań | 23 208 | | 23 208 | |
| # segmentów | 369 795 | | 369 795 | |
| # t/s | 15,93 | | 15,93 | |
| # typy zależności | 28 | | 31 (58 z podtypami) | |
| | liczba | % | liczba | % |
| nieprojekcyjne krawędzie | 6340 | 1,71% | 5352 | 1,45% |
| drzewa nieprojekcyjne | 1959 | 8,61% | 1466 | 6,32% |
| dodatkowe krawędzie | n/a | | 15 872 | 4,29% |
| grafy rozszerzone | n/a | | 9625 | 41,5% |

# Nieprojekcyjność

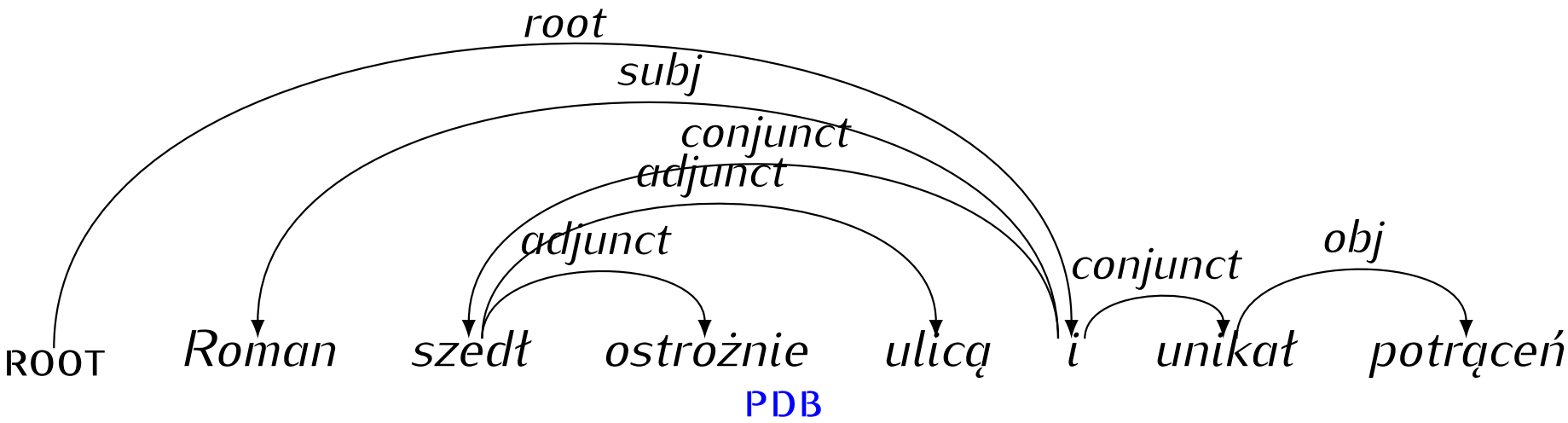



PDB

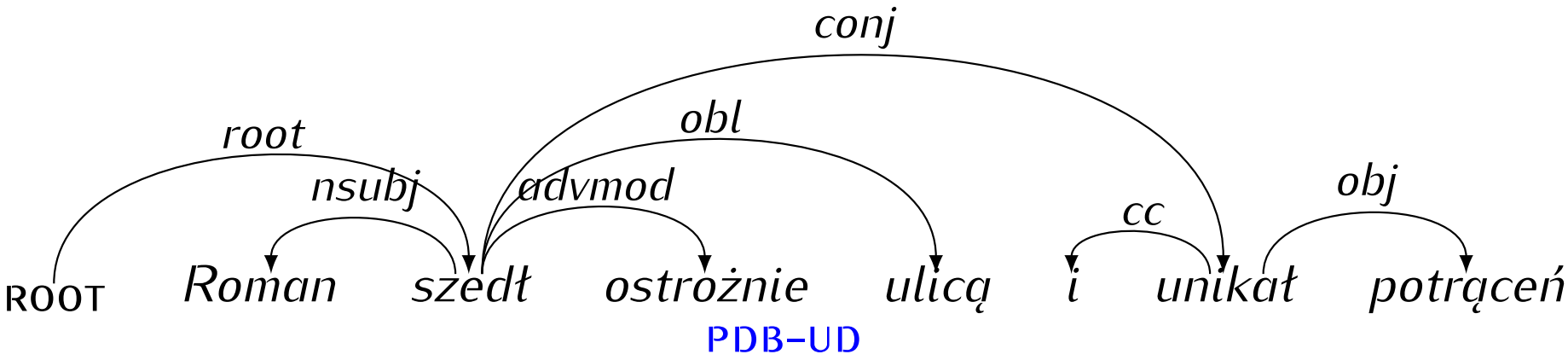



PDB–UD

# Grafy rozszerzone PDB-UD

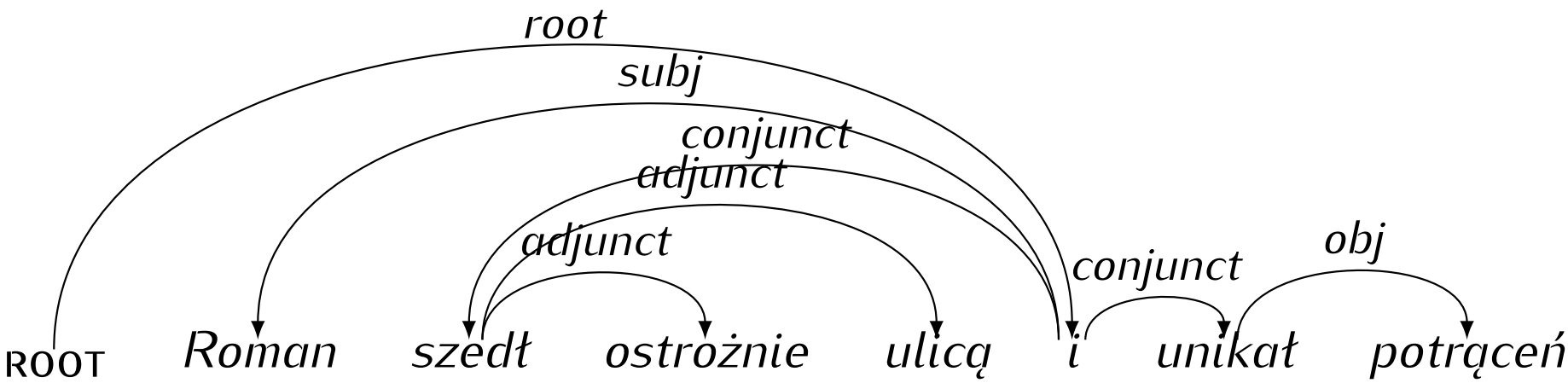



PDB

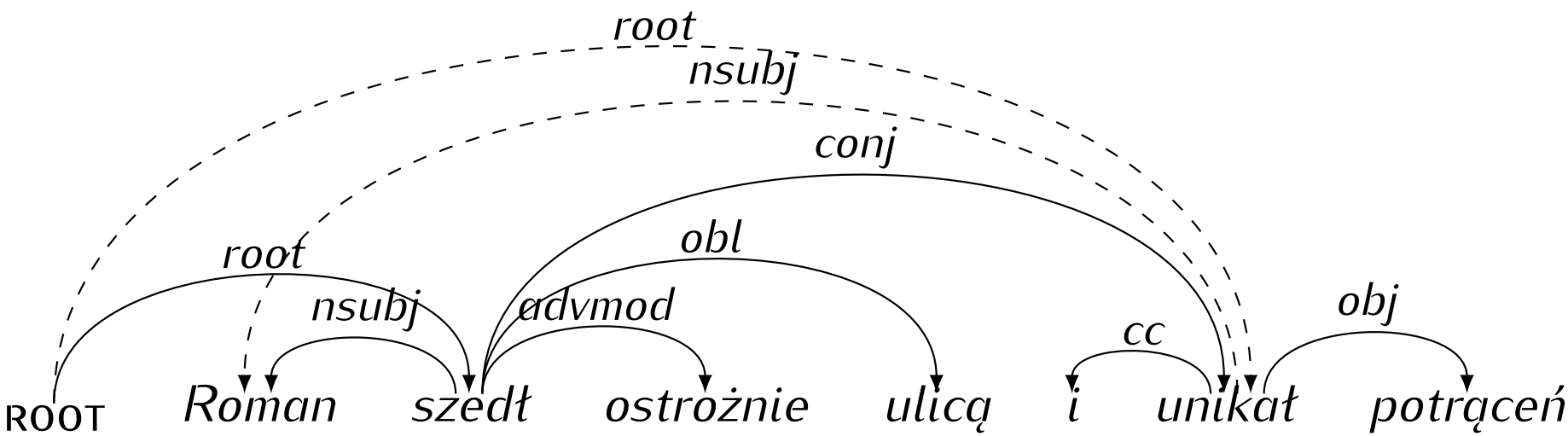



PDB–UD

# Grafy rozszerzone PDB-UD




PDB



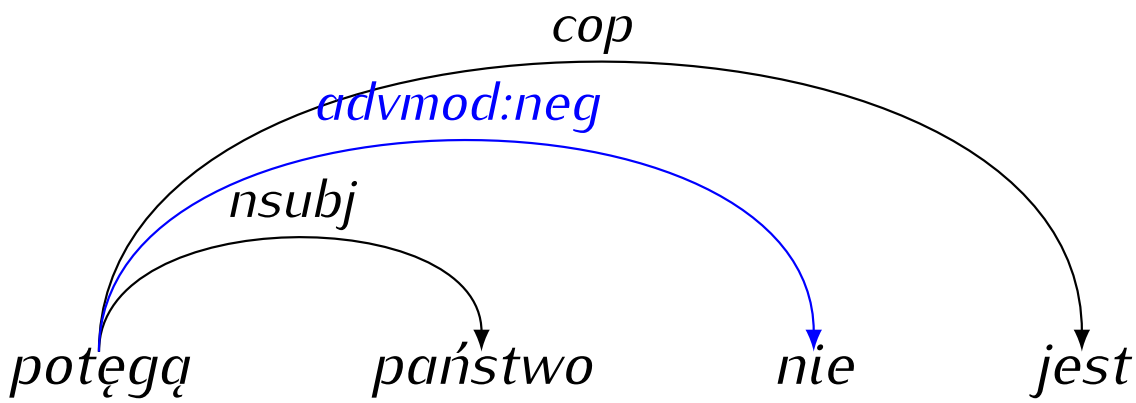

PDB–UD

# PDB-UD vs. UD-Polish-LFG i UD-English-EWT

| | UD-LFG | PDB-UD | UD-EWT |
|---|---|---|---|
| # zdań | 17 246 | 23 208 | 16 622 |
| # segmentów | 130 967 | 369 795 | 254 829 |
| # t/s | 7,6 | 15,9 | 15,3 |
| # typy zależności | 27 (39) | 31 (58) | 36 (49) |
| nieprojekcyjne krawędzie | 0,3% | 1,4% | 0,9% |
| drzewa nieprojekcyjne | 0,7% | 6,3% | 4,8% |
| dodatkowe krawędzie | 1,3% | 4,3% | 3,7% |
| grafy rozszerzone | 8,1% | 41,5% | 29,6% |

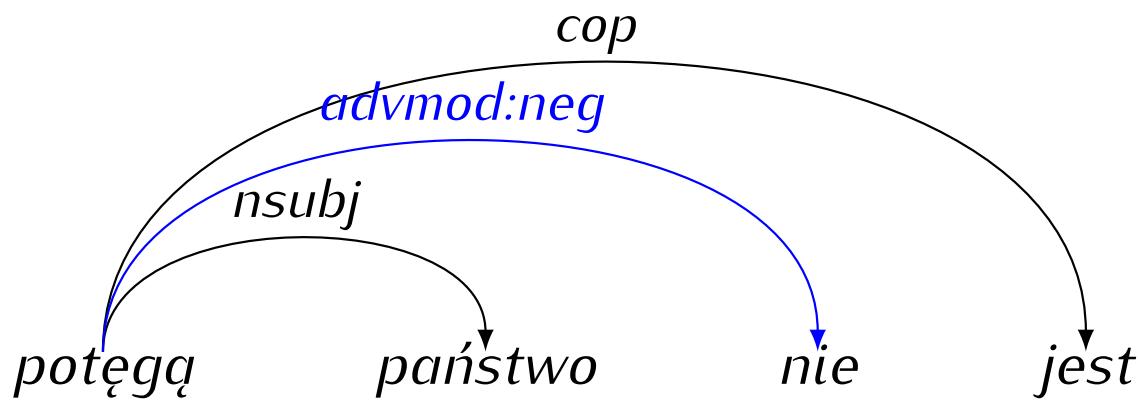
UD-EWT: N. Silveira, T. Dozat, M-C. de Marneffe, S. Bowman, S. Connor, J. Bauer i Ch. Manning. 2014. *A Gold Standard Dependency Corpus for English*. [w:] Proceedings of the Ninth International Conference on Language Resources and Evaluation.
UD-LFG: A. Patejuk i A. Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. IPI PAN, Warszawa.

# Konstrukcje predykatywne: UD-SZ vs. PDB-UD

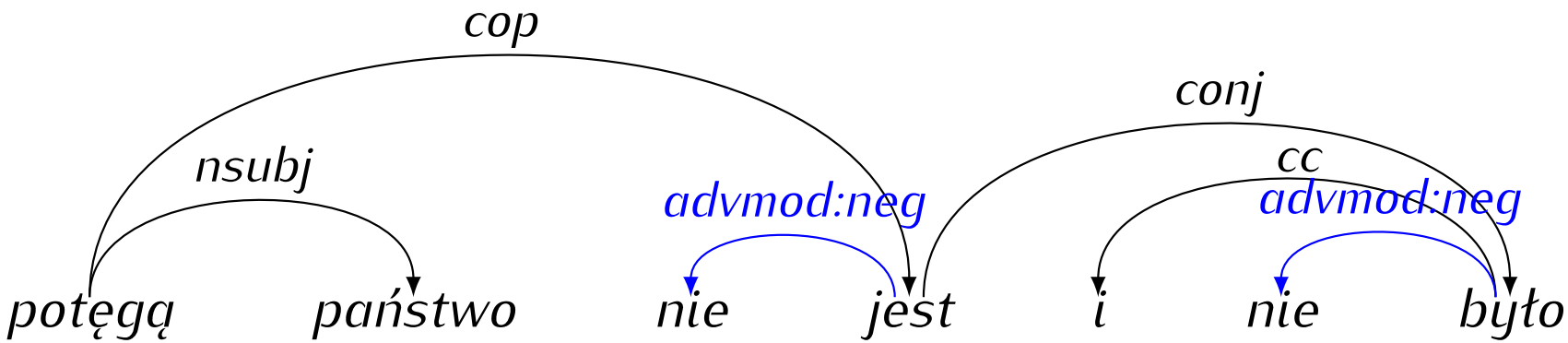



UD-SZ (i UD-LFG)

# Konstrukcje predykatywne: UD-SZ vs. PDB-UD

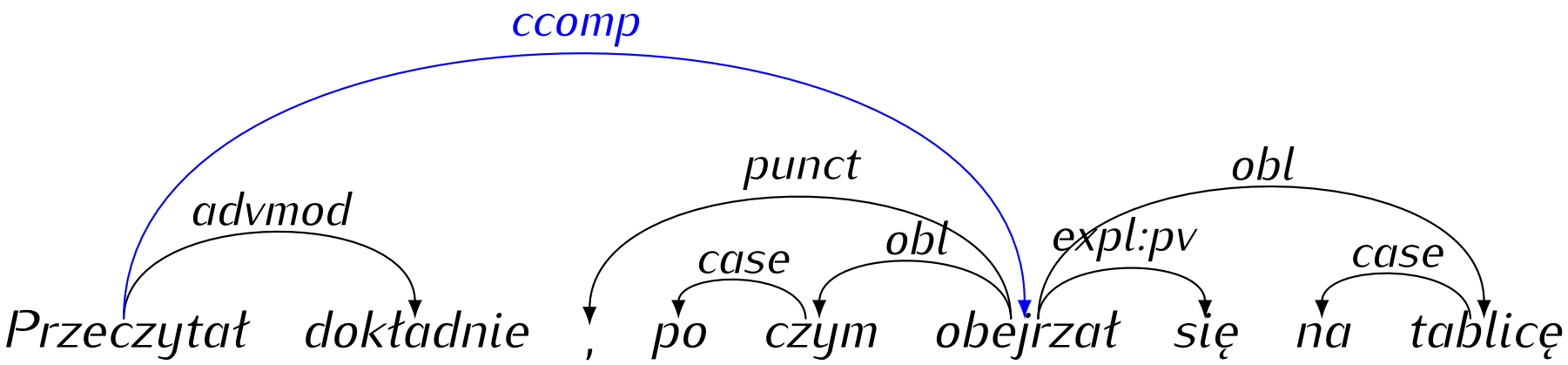



UD–SZ (i UD–LFG)



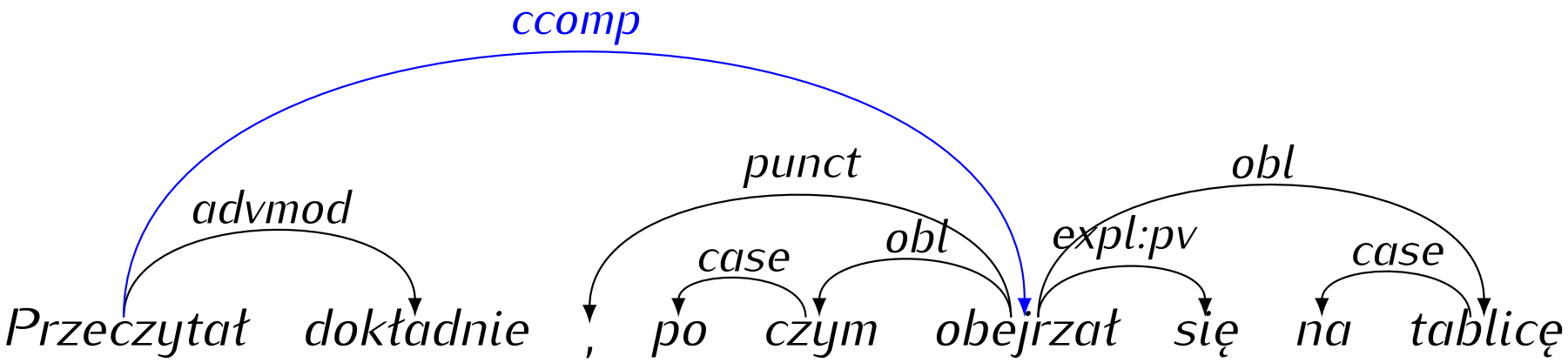

PDB–UD

# Zdania względne: UD-SZ vs. PDB-UD

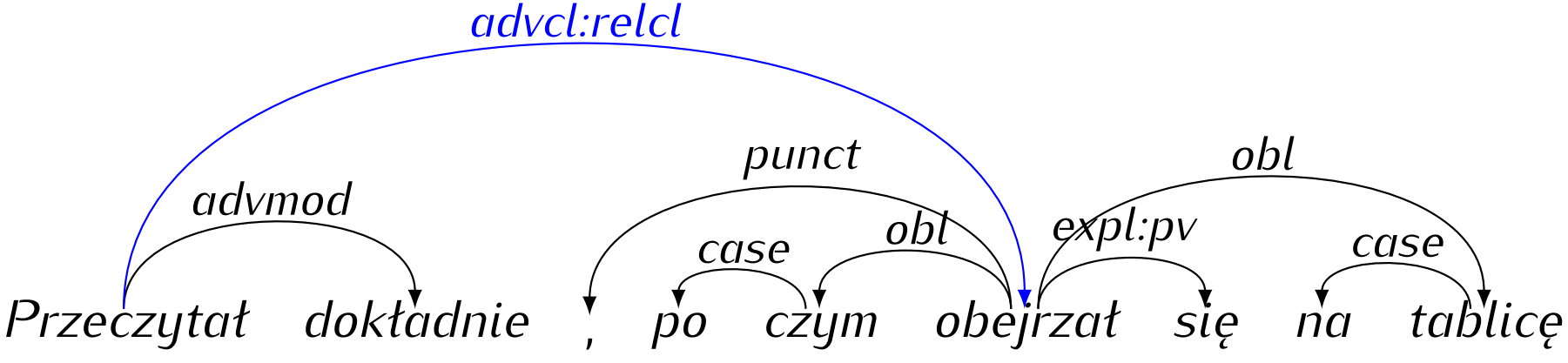



UD-SZ (i UD-LFG)

# Zdania względne: UD-SZ vs. PDB-UD




UD-SZ (i UD-LFG)




PDB-UD

# Zdania względne: UD-SZ vs. PDB-UD

*Jak więc hobbici mogli używać narzędzi i posługiwać się ogniem, czego ślady odkryto na wyspie Flores?*

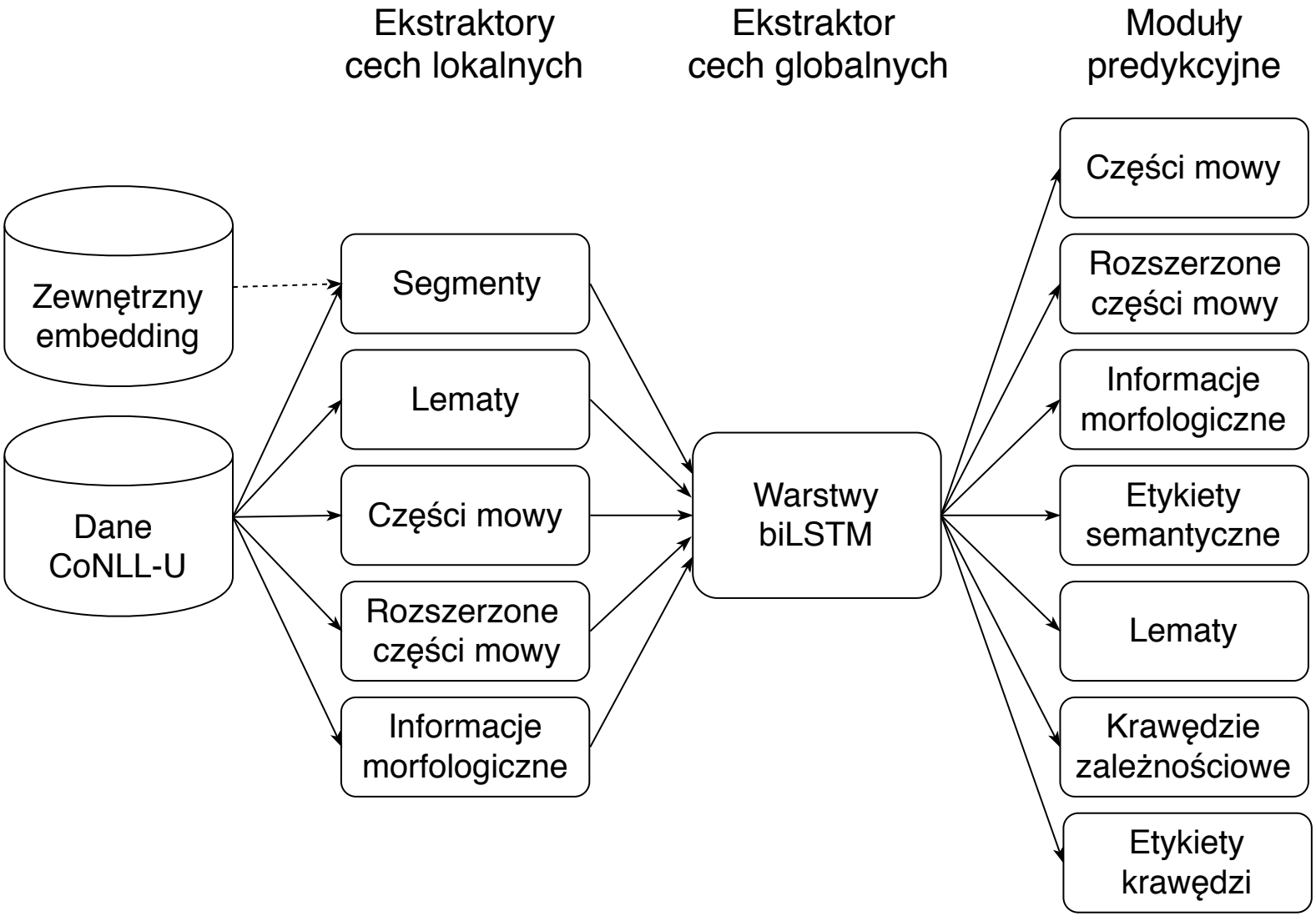
- Ani używać ani posługiwać się nie dopuszczają dopełnień zdaniowych (ccomp).
- To nie jest koordynacja, bo inne znaczenie w *Jak więc hobbici mogli używać narzędzi i posługiwać się ogniem i czego ślady odkryto na wyspie Flores?*.
- To jest rodzaj zdania względnego.

# Spis treści

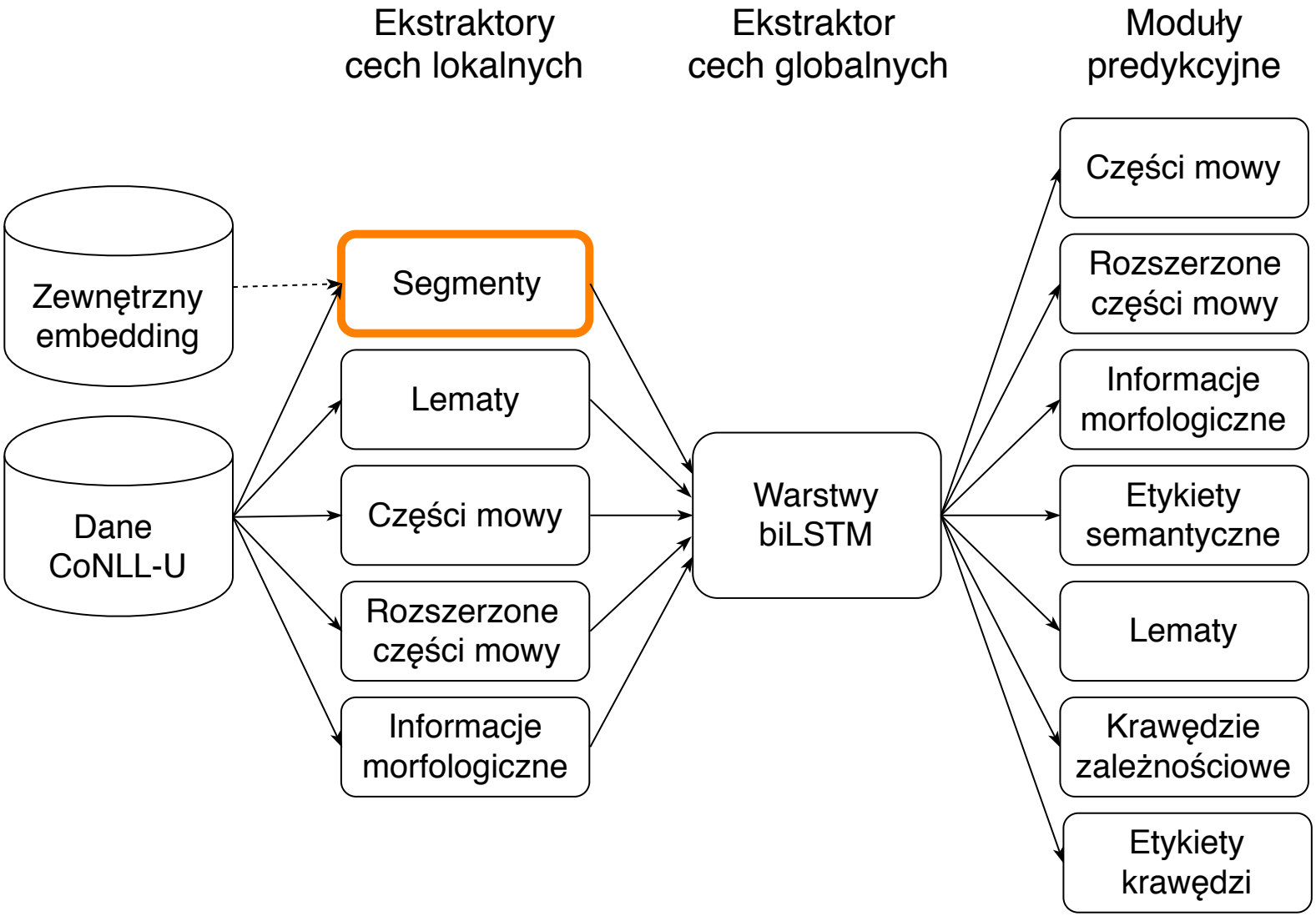
# Czym jest COMBO?

- System oparty o sieci neuronowe służący do:
  - Analizy morfoskładniowej,
  - Lematyzacji,
  - Parsowania zależnościowego,
- Zaimplementowany w Pythonie i bibliotece Keras,
- Kod: https://github.com/360er0/COMBO

# Schemat systemu



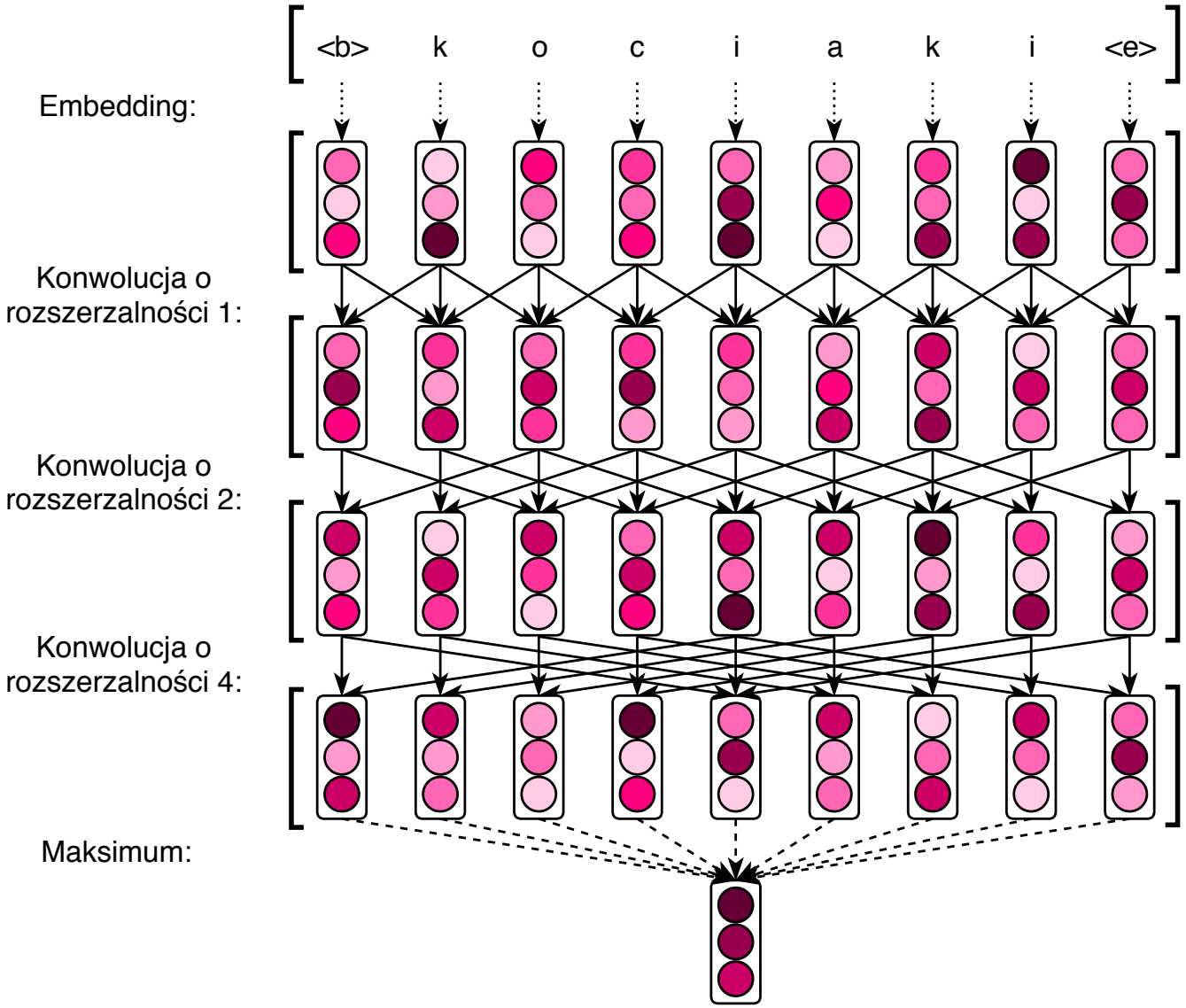
# Ekstrakcja cech segmentów



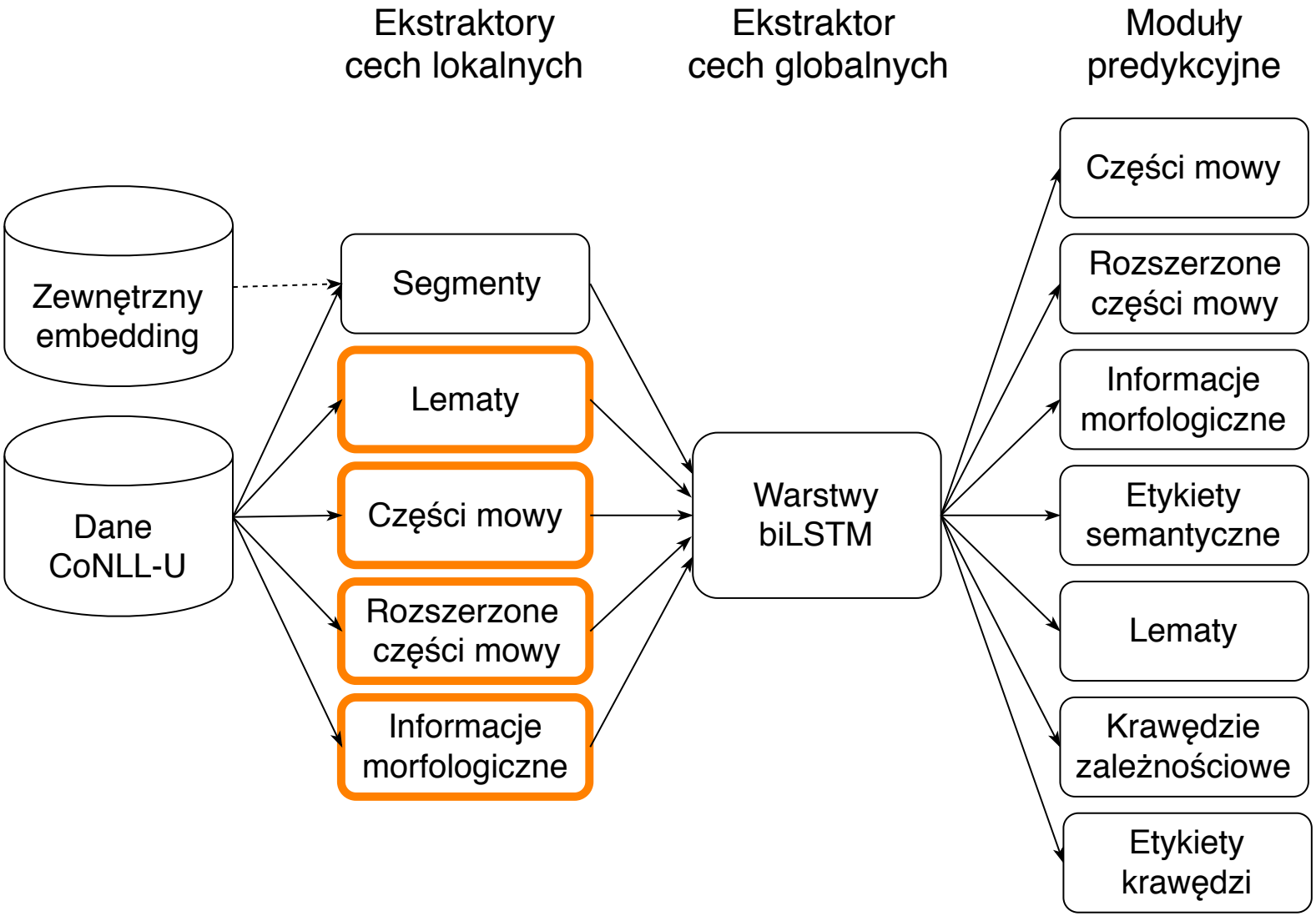
# Ekstrakcja cech segmentów

- Zewnętrzny embedding (np. word2vec, fasttext, itp.),
  - Nie są modyfikowane podczas treningu systemu,
  - Transformacja pojedynczą warstwą gęstą,
- Embedding po znakach przy użyciu sieci konwolucyjnej.

# Ekstrakcja cech segmentów

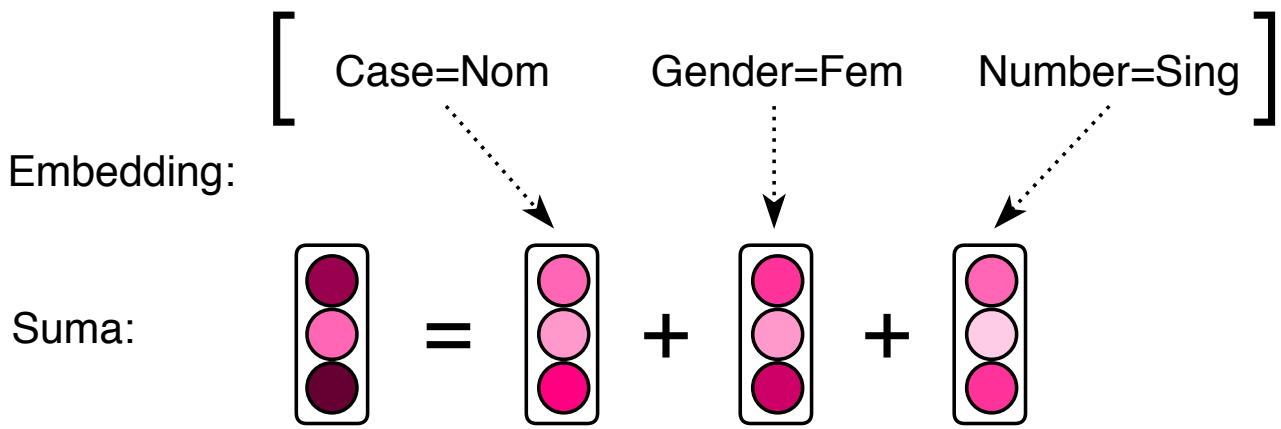


# Ekstrakcja pozostałych cech

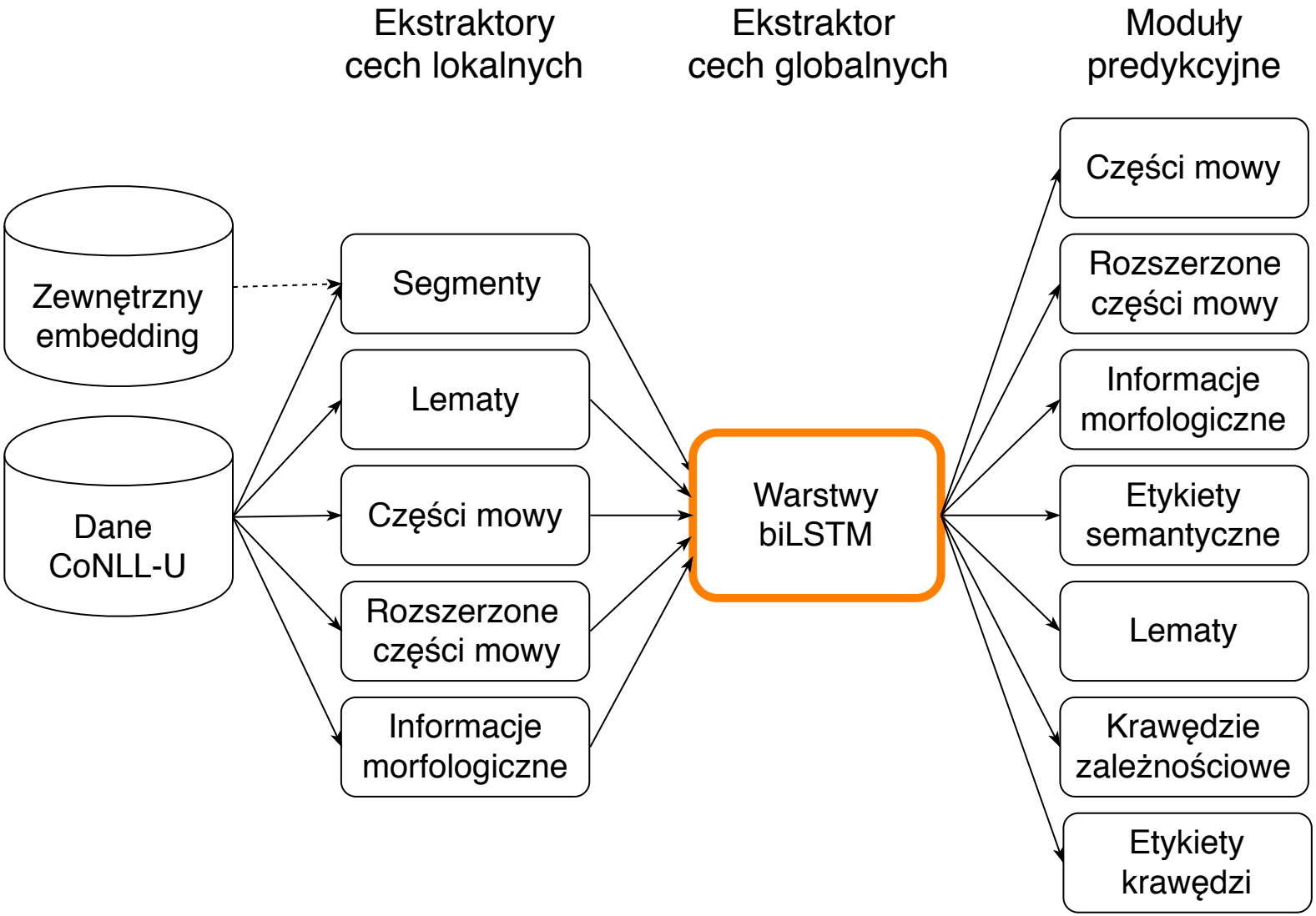


# Ekstrakcja pozostałych cech

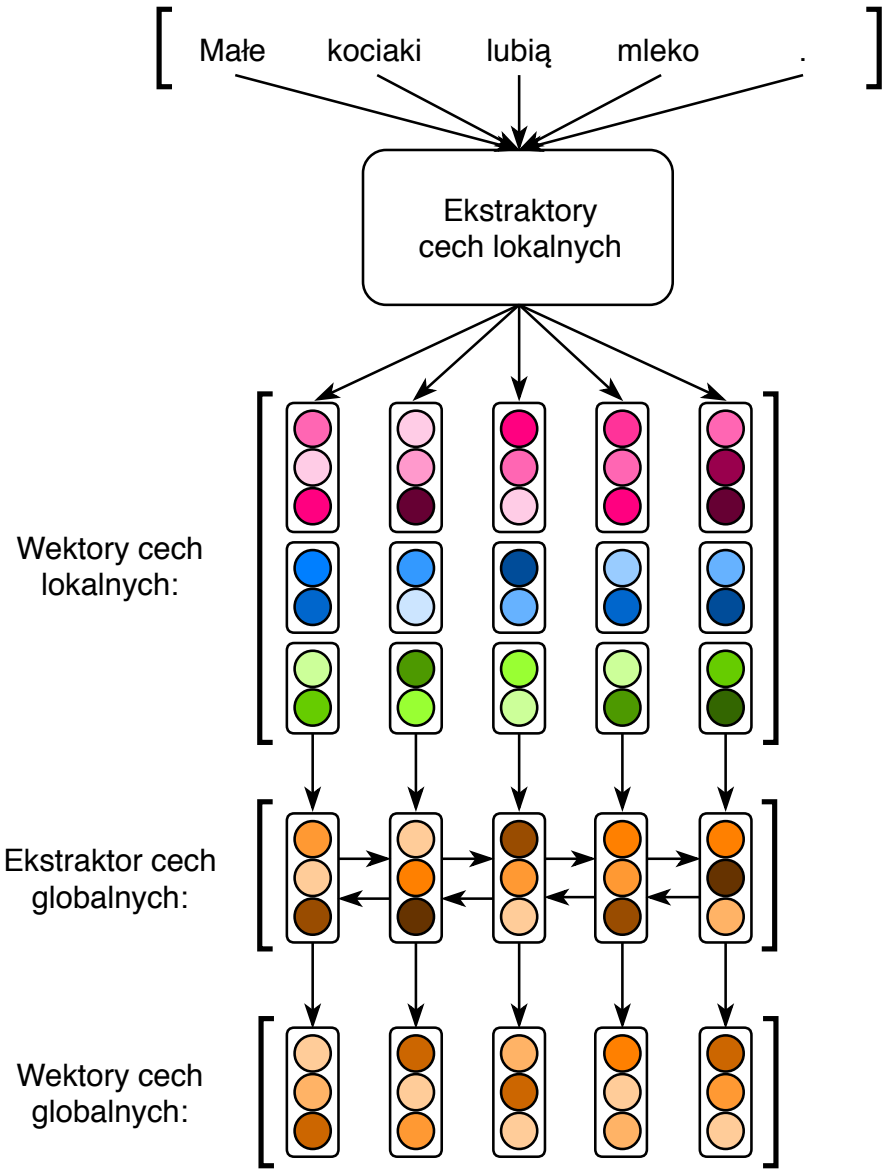
- **Lematy**:
  - Embedding po znakach (analogicznie jak dla segmentów).
- **Części mowy**:
  - Embedding wyznaczony podczas treningu systemu.
- **Rozszerzone części mowy**:
  - Embedding wyznaczony podczas treningu systemu.
- **Informacje morfologiczne**:
  - Suma embeddingów wyznaczonych podczas treningu.



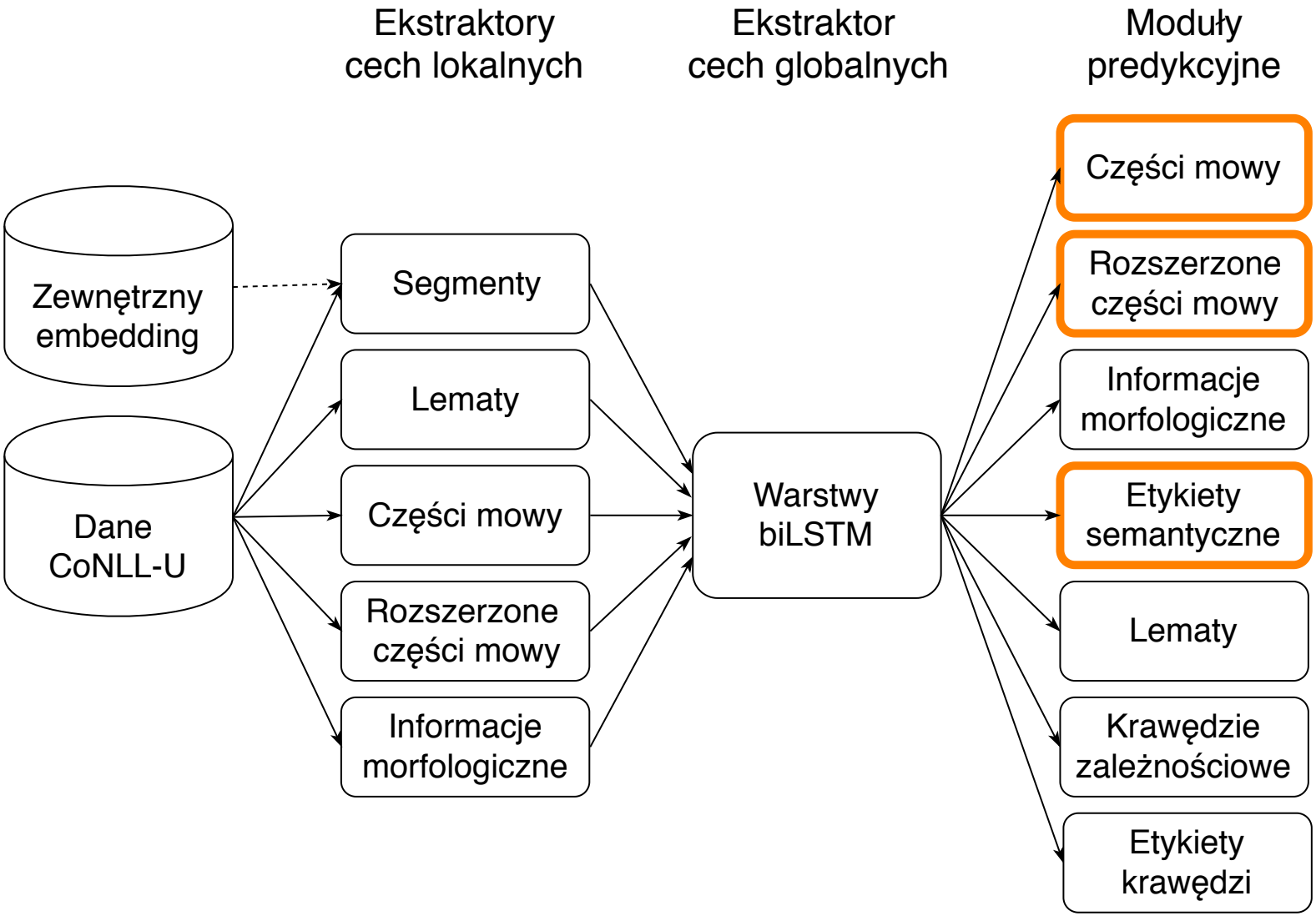
# Ekstraktor cech globalnych



# Ekstraktor cech globalnych



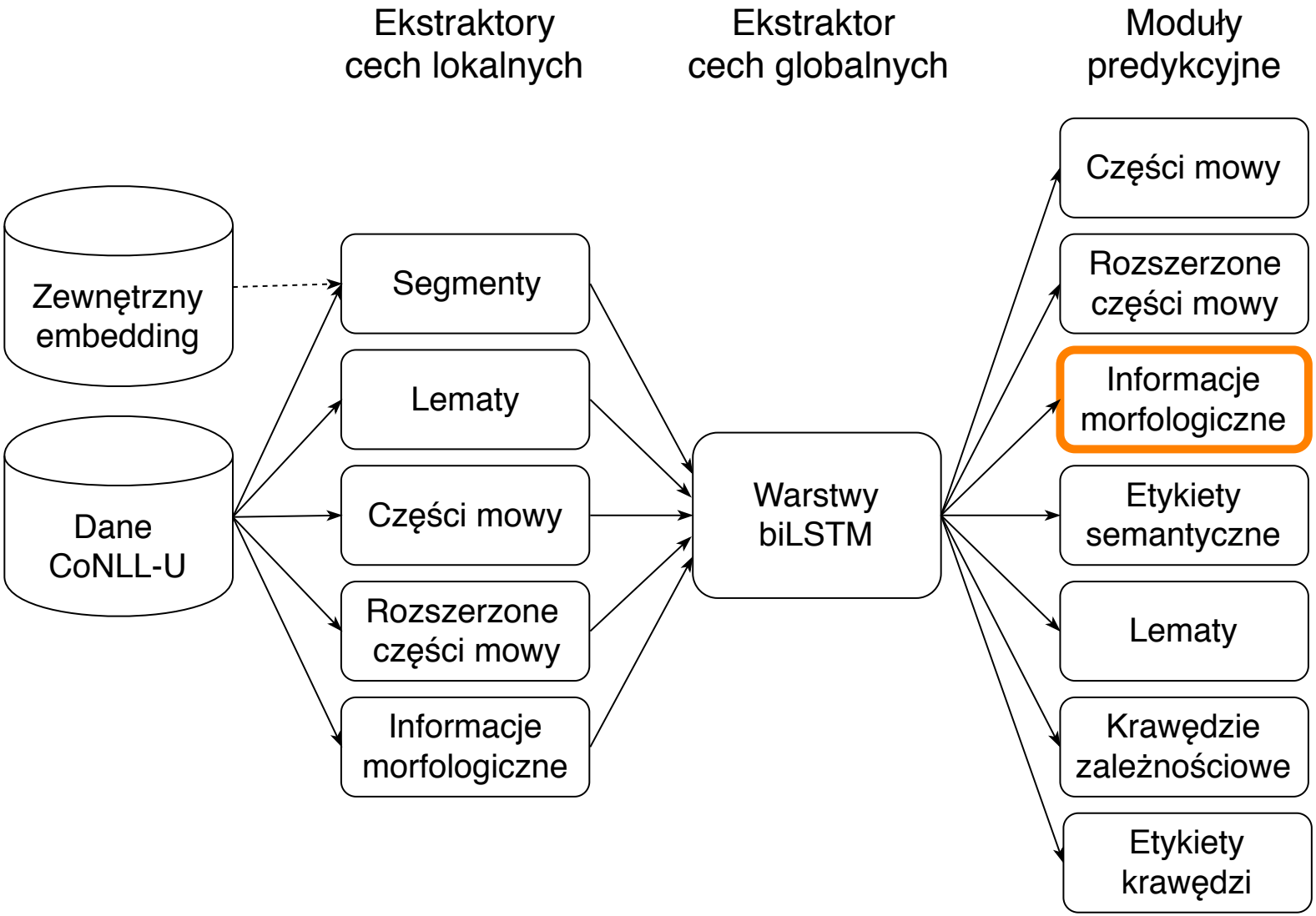
# Predykcja części mowy i etykiet semantycznych



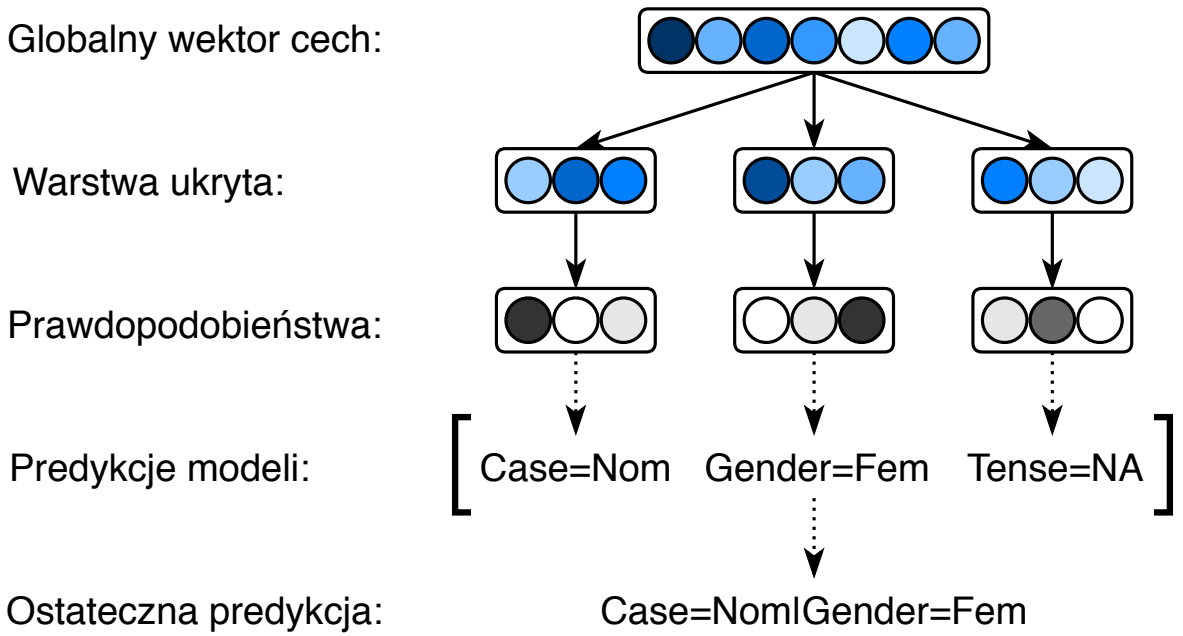
# Predykcja części mowy i etykiet semantycznych

- **Części mowy**:
  - Jednowarstwowa sieć neuronowa.
- **Rozszerzone części mowy**:
  - Jednowarstwowa sieć neuronowa.
- **Etykiety semantyczne**:
  - Jednowarstwowa sieć neuronowa.

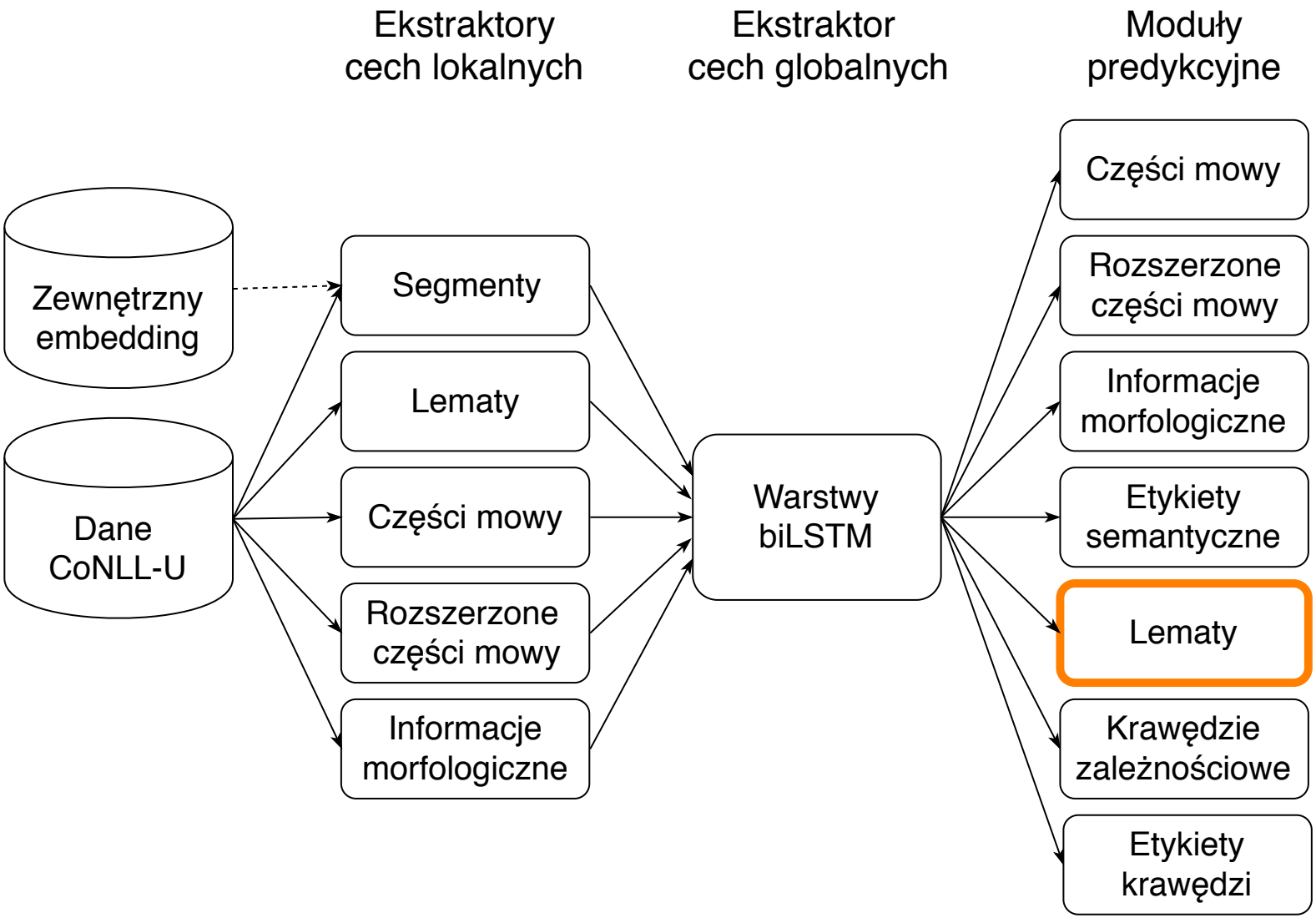
# Predykcja informacji morfologicznych



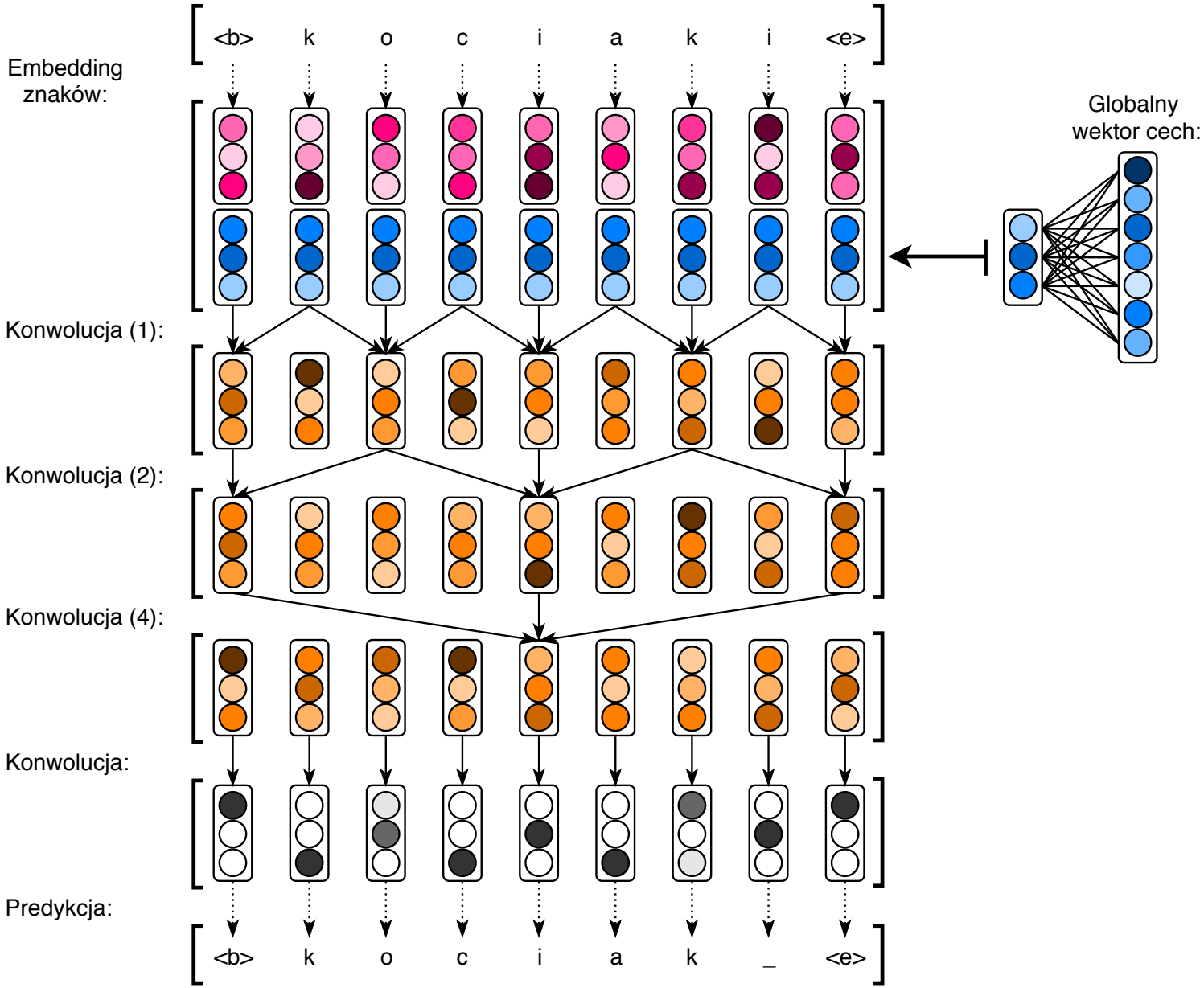
# Predykcja informacji morfologicznych



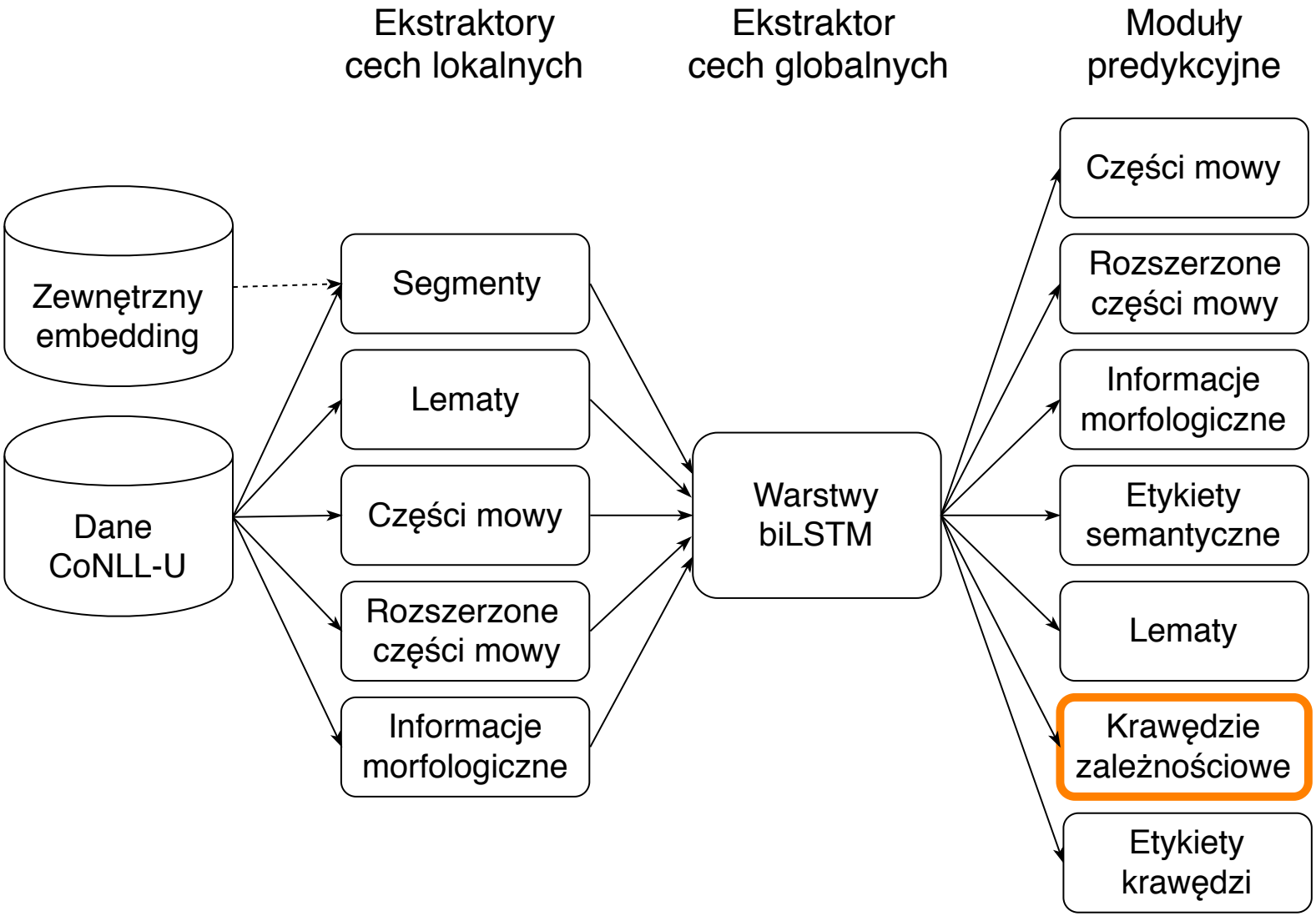
# Predykcja lematów



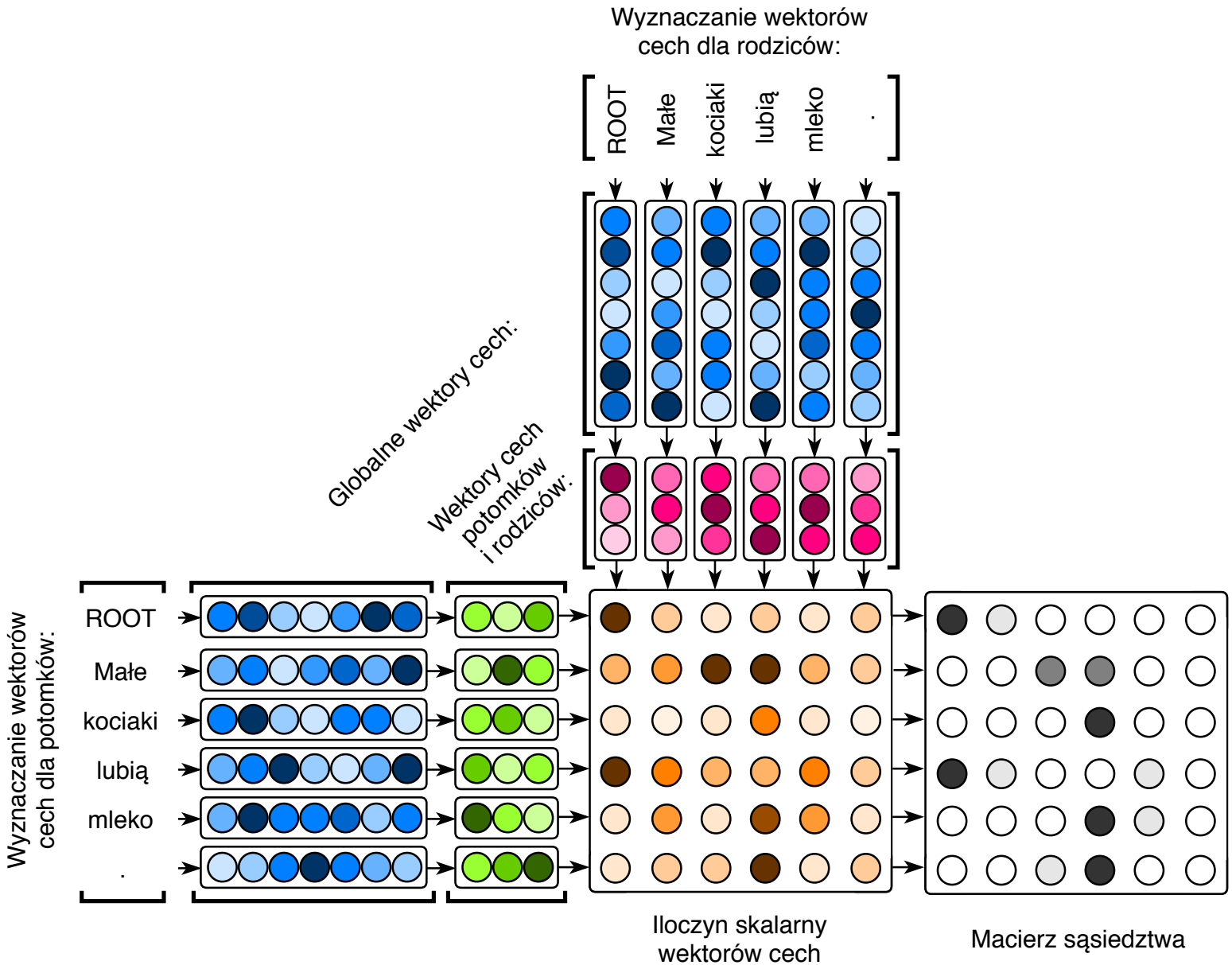
# Predykcja lematów



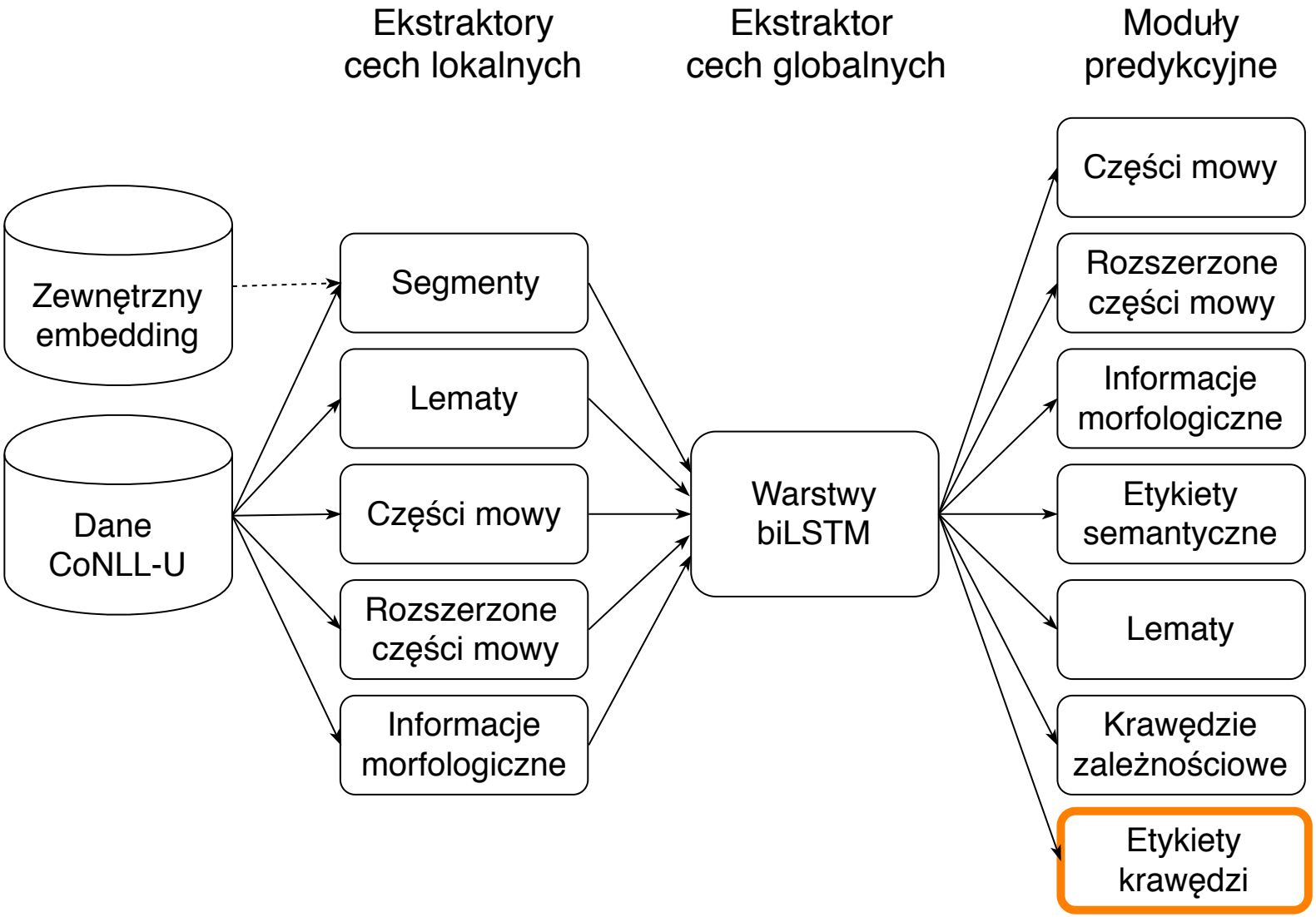
# Predykcja krawędzi zależnościowych



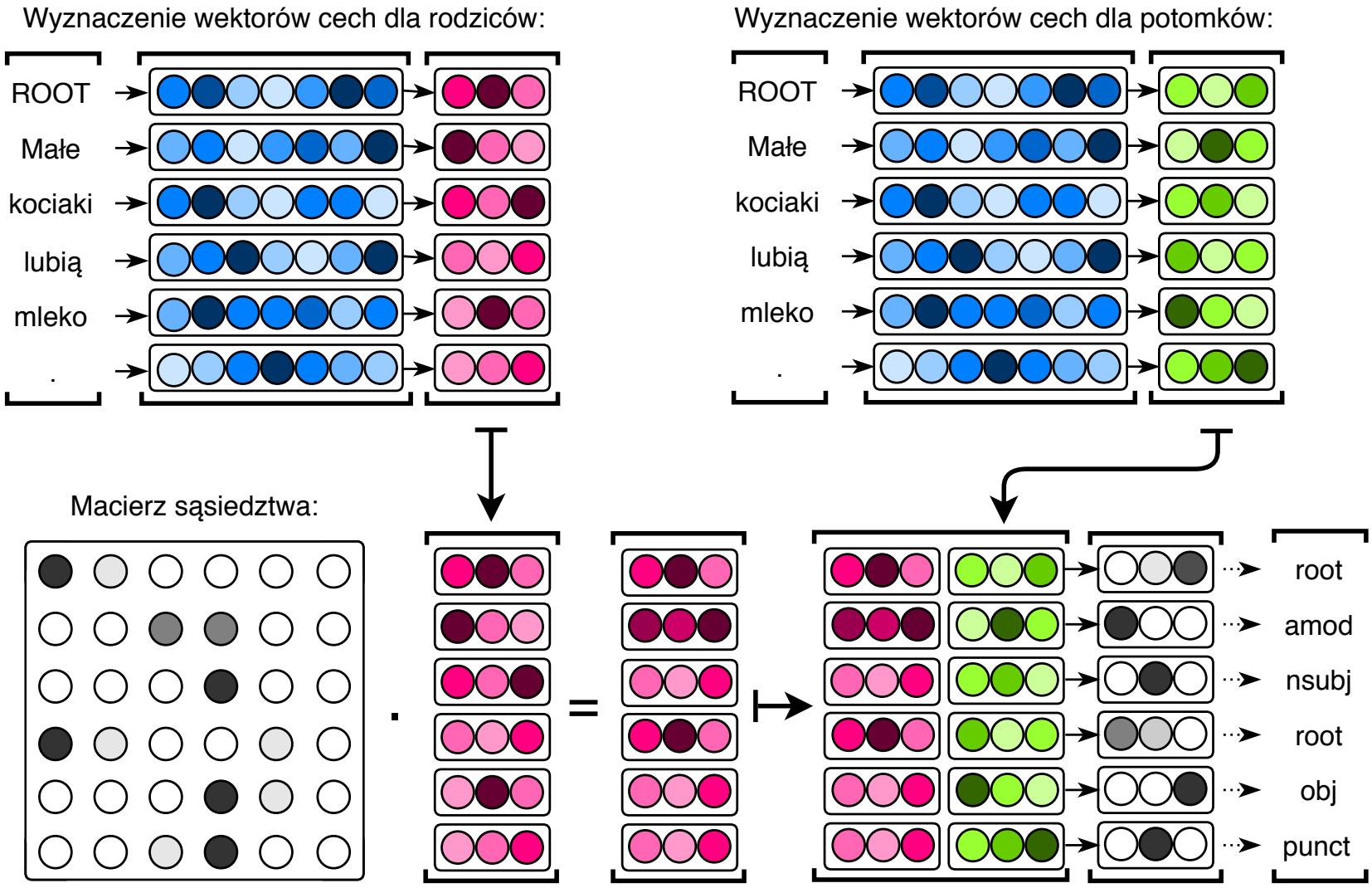
# Predykcja krawędzi zależnościowych



# Predykcja etykiet zależnościowych



# Predykcja etykiet zależnościowych



# Analiza modelu

- Dodatkowa funkcja straty,
- Samo-trening,
- Modyfikacja zewnętrznego embeddingu,
- Wyniki z CoNLL 2018.

# Miary ewaluacyjne

- Miary z konkursu CoNLL 2018.
- Każda miara jest średnią harmoniczną ($F_1$) dokładności (P, precision) i kompletności (R, recall):

$$F_1 = \frac{2PR}{P + R}$$

- **UAS** (unlabelled attachment score)
- **LAS** (labelled attachment score)
- **CLAS** (content-word labelled attachment score)
- **MLAS** (morphology-aware labelled attachment score)
- **BLEX** (bi-lexical dependency score)

# Dodatkowa funkcja straty

- **Motywacja**: mniejsza liczba cykli w drzewach zależnościowych,
- Dodatkowa funkcja straty:

$$\text{loss}(A) = \sum_{k=1}^{K} \text{tr}(A^k)$$

- $A$ - przewidziana macierz sąsiedztwa.

| K | UAS | % Cykli |
|---|---|---|
| K=0 | 86,76 | 5,70 |
| K=3 | 86,75 | 4,84 |
| Różnica | -0,01 | -0,86 |

Tabela: Porównanie wyników dla 15 wybranych języków z UD.

# Samo-trening systemu

- Trening modelu na danych treningowych,
- Predykcja na 10M zdań z Common Crawl,
- Trening modelu na powyższych danych,
- Dotrenowanie modelu na danych treningowych,

| **Rodzaj treningu** | **LAS** | **MLAS** | **BLEX** |
|---|---|---|---|
| Standardowy | 79,69 | 64,46 | 71,71 |
| Samo-trening | 80,86 | 67,33 | 73,43 |
| Różnica | 1,17 | 2,87 | 1,72 |

Tabela: Porównanie wyników dla 20 wybranych języków z UD.

# Modyfikacja zewnętrznego embeddingu

| Model | LAS | MLAS | BLEX |
|---|---|---|---|
| Niemodyfikowany | 82,13 | 68,04 | 73,84 |
| Modyfikowany | 82,27 | 68,21 | 74,00 |
| Różnica | 0,14 | 0,17 | 0,16 |

Tabela: Porównanie wyników dla 12 wybranych języków z UD.

# Wyniki z CoNLL 2018

- Klasyfikacja końcowa:
  - LAS: 73,02 (3 miejsce),
  - MLAS: 60,25 (4 miejsce),
  - BLEX: 64,44 (3 miejsce),
- Wyniki w podziale na wielkość zbioru danych:

| Kategoria | LAS | MLAS | BLEX |
|---|---|---|---|
| Wszystkie | 73,02 | 60,25 | 64,44 |
| Duże | 81,72 | 70,30 | 74,42 |
| PUD | 72,18 | 58,07 | 60,97 |
| Małe | 66,90 | **49,24** | **54,89** |
| Minimalne | 19,26 | 1,89 | 6,17 |

# Spis treści

# Parametry eksperymentów

- Dane: PDB i PDB-UD podzielone na train (80%), test (10%) i dev (10%)
- Systemy:

| system | architektura | klasyfikator | parse | tag | lemat |
|---|---|---|---|---|---|
| MaltParser | trans | RL | tak | nie | nie |
| BIST parser | trans/graph | biLSTM | tak | nie | nie |
| MATE parser | graph | perceptron | tak | tak | tak |
| UDPipe | trans | 1-layer NN | tak | tak | tak |
| Stanford | graph | biLSTM | tak | tak | tak |
| COMBO | graph | biLSTM | tak | tak | tak |

trans – transition-based parser,
graph – graph-based parser,
RL – klasyfikator liniowy oparty na regresji liniowej,
1-layer NN – klasyfikator nieliniowy oparty na 1-warstwowej sieci neuronowej,
biLSTM – Bidirectional Long-Short Term Memory network.

# Eksperymenty

1. Testowanie jakości parsowania zależnościowego.
2. Testowanie jakości morfoskładniowej predykcji drzew zależnościowych.
3. Porównanie jakości parserów trenowanych na drzewach PDB i grafach PDB-UD.

| System | UAS | LAS |
|---|---|---|
| MaltParser | 83.39 | 80.39 |
| UDPipe | 83.41 | 80.13 |
| BIST transition-based | 86.77 | 83.35 |
| BIST graph-based | 87.01 | 83.67 |
| MATE parser | 89.51 | 87.12 |
| COMBO | 91.36 | 88.92 |
| Stanford parser | **92.78** | **90.61** |

Wniosek: Parsery grafowe są lepsze dla języka polskiego.

| System | UPOS | XPOS | FEATS | AllTags | LEMMA |
|---|---|---|---|---|---|
| MATE | 96.83 | 84.83 | 87.33 | 81.14 | 92.32 |
| UDPipe | 96.80 | 86.22 | 88.14 | 85.77 | 95.51 |
| COMBO | 97.51 | 90.30 | 91.83 | 88.67 | 96.35 |
| COMBO+embedd | 97.96 | 91.60 | 93.22 | 90.09 | **96.92** |
| Stanford | 98.03 | **93.16** | 93.93 | 91.87 | 82.05 |
| Stanford+embedd | **98.07** | 93.09 | **94.00** | **91.93** | 82.05 |

UPOS – uniwersalna część mowy
XPOS – tag specyficzny dla języka
FEATS – lista (uniwersalnych) cech morfologicznych
AllTags – UPOS + XPOS + FEATS
LEMMA – lemat

+embedd – estymacja modelu wspierana przez embedding zewnętrzny

| System | UAS | LAS | CLAS | MLAS | BLEX |
|---|---|---|---|---|---|
| UDPipe | 80.95 | 76.11 | 71.87 | 61.34 | 67.82 |
| UDPipe+embedd | 82.89 | 78.57 | 74.81 | 63.88 | 70.49 |
| MATE | 85.56 | 81.34 | 77.91 | 65.31 | 69.64 |
| COMBO | 90.20 | 86.64 | 83.65 | 73.90 | 79.94 |
| COMBO+embedd | 91.17 | 87.89 | 85.27 | 76.63 | **81.86** |
| Stanford+embedd | 91.65 | 88.70 | 86.25 | 78.26 | 62.67 |
| Stanford | **91.70** | **88.88** | **86.46** | **78.36** | 62.68 |

# 3. Predykcja drzew PDB i grafów PDB-UD

| | PDB | | | PDB-UD | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| COMBO | 88.81 | 88.81 | 88.81 | 87.63 | 82.56 | 85.02 |
| COMBO+embedd | **89.56** | **89.56** | **89.56** | 88.09 | 83.15 | 85.55 |

- Drzewa PDB i grafy PDB-UD kodują te same informacje (tj. współdzielone podrzędniki i nadrzędniki w strukturach koordynacyjnych).
- Liczymy dokładność, kompletność i miarę $F_1$ poszczególnych krawędzi.
- Predykcja drzew jest mniej podatna na błędy niż predykcja grafów.

# Publikacje


Alina Wróblewska i Piotr Rybak.
Dependency parsing of Polish.
*Poznan Studies in Contemporary Linguistics*, de Gruyter, 2019, (przyjęte do druku).

Piotr Rybak i Alina Wróblewska.
Semi-Supervised Neural System for Tagging, Parsing and Lematization.
In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, s. 45–54. Association for Computational Linguistics.

Alina Wróblewska.
Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format.
In *Proceedings of the Second Workshop on Universal Dependencies* (UDW 2018), 2018, s. 173–182. Association for Computational Linguistics.

# Publikacje


Alina Wróblewska i Piotr Rybak.
Dependency parsing of Polish.
*Poznan Studies in Contemporary Linguistics*, de Gruyter, 2019, (przyjęte do druku).

Piotr Rybak i Alina Wróblewska.
Semi-Supervised Neural System for Tagging, Parsing and Lematization.
In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, s. 45–54. Association for Computational Linguistics.

Alina Wróblewska.
Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format.
In *Proceedings of the Second Workshop on Universal Dependencies* (UDW 2018), 2018, s. 173–182. Association for Computational Linguistics.


Dziękujemy za uwagę!

# Podziękowania

Przedstawione badania były finansowane przez Narodowe Centrum Nauki (grant SONATA 8 nr 2014/15/D/HS2/03486) oraz przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach inwestycji w infrastrukturę badawczą CLARIN-PL.