# Automatyczne metody ujednoznaczniania drzew rozbioru wypowiedzeń w języku polskim jako ostatnia faza przetwarzania parsera Świgra

Dominika Rogozińska, Marcin Woliński

Instytut Podstaw Informatyki PAN

5 grudnia 2016

# Kontekst

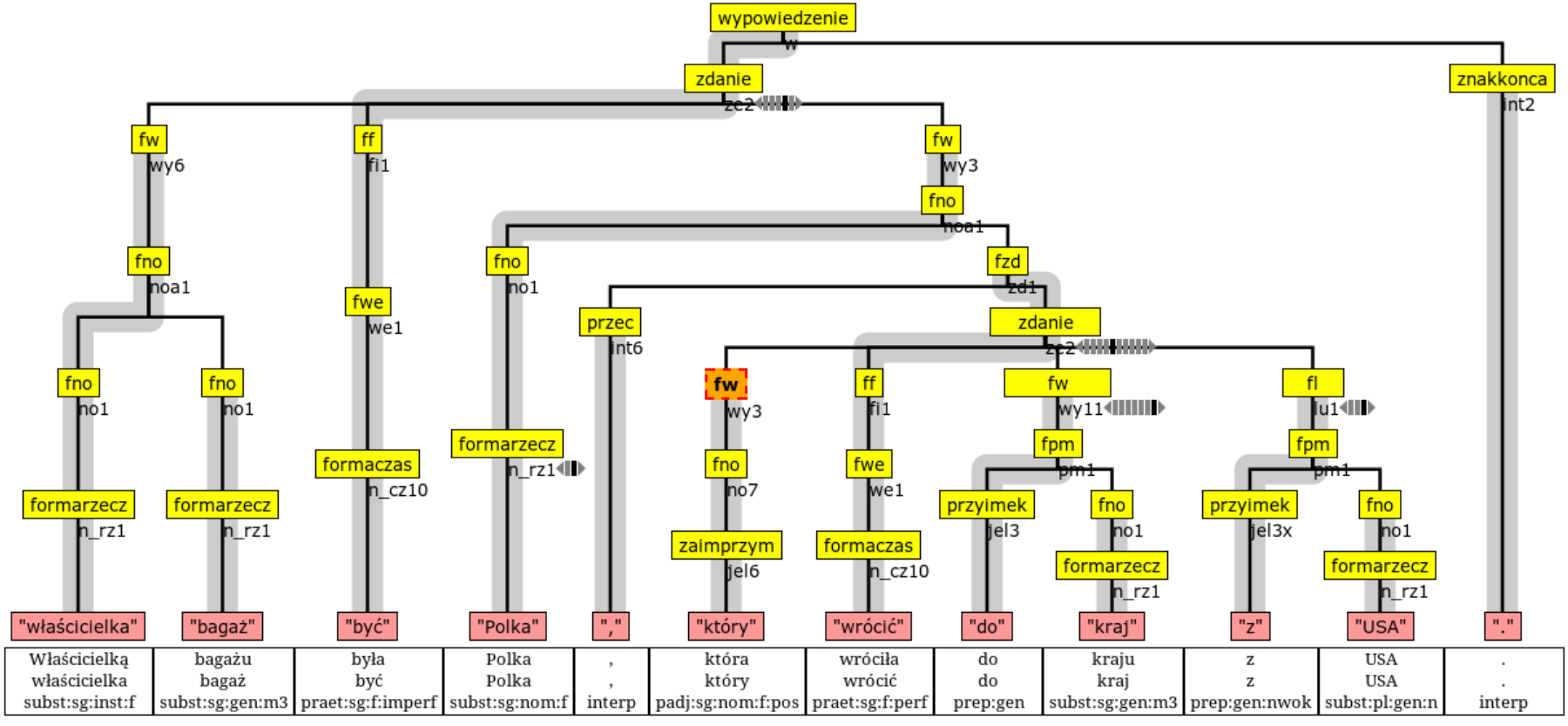
- Parser *Świgra 2* – implementacja *Gramatyki formalnej języka polskiego* Marka Świdzińskiego (1992), z późniejszymi znaczącymi zmianami
- *Składnica frazowa* – bank drzew składnikowych generowanych za pomocą *Świgry 2*, ujednoznacznianych i weryfikowanych ręcznie

# Kontekst

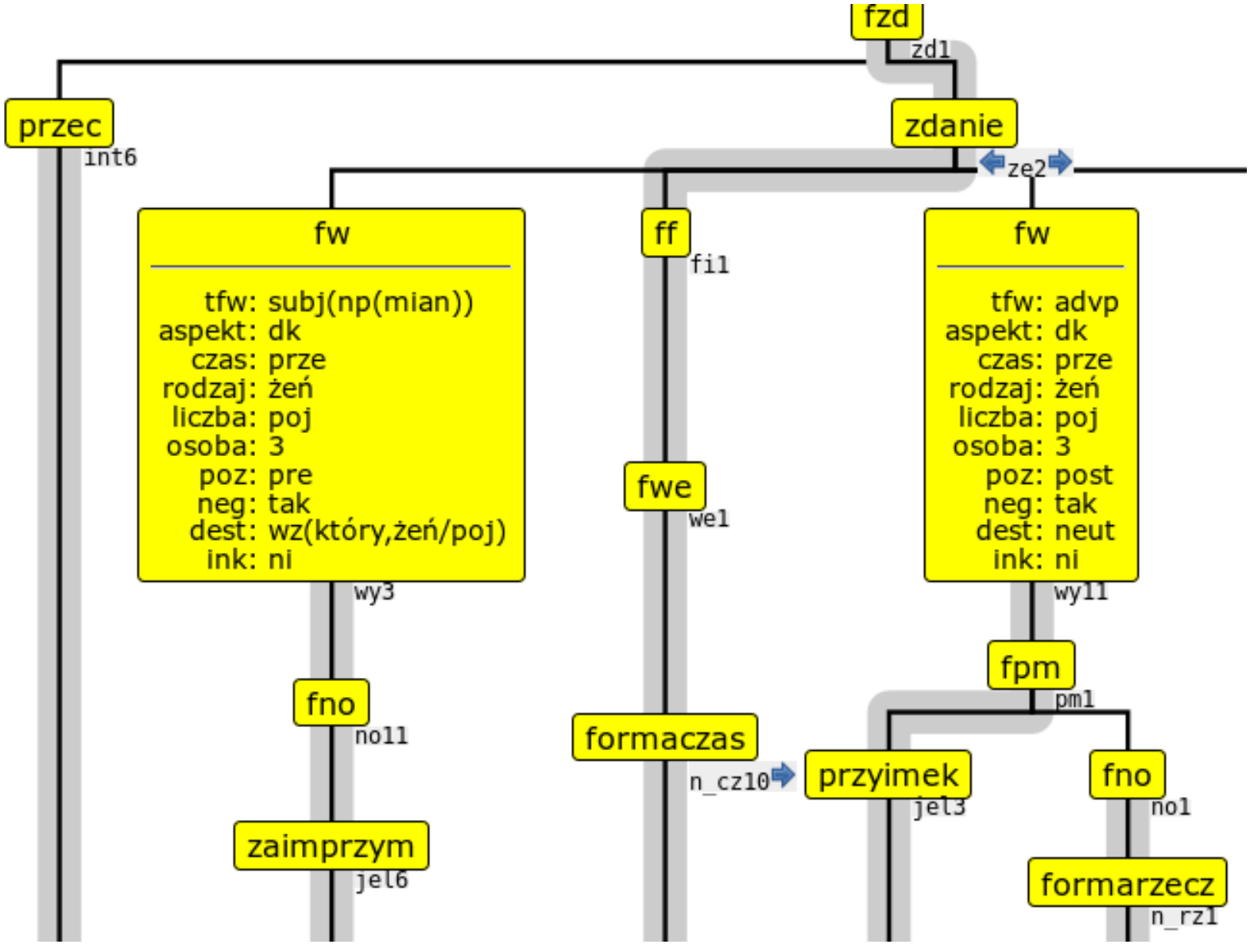
- Parser *Świgra 2* – implementacja *Gramatyki formalnej języka polskiego* Marka Świdzińskiego (1992), z późniejszymi znaczącymi zmianami
- *Składnica frazowa* – bank drzew składnikowych generowanych za pomocą *Świgry 2*, ujednoznacznianych i weryfikowanych ręcznie

- *Świgra*, jako parser regułowy, generuje upakowane lasy wszystkich drzew składniowych możliwych dla danego zdania

# Przykład drzewa składnikowego



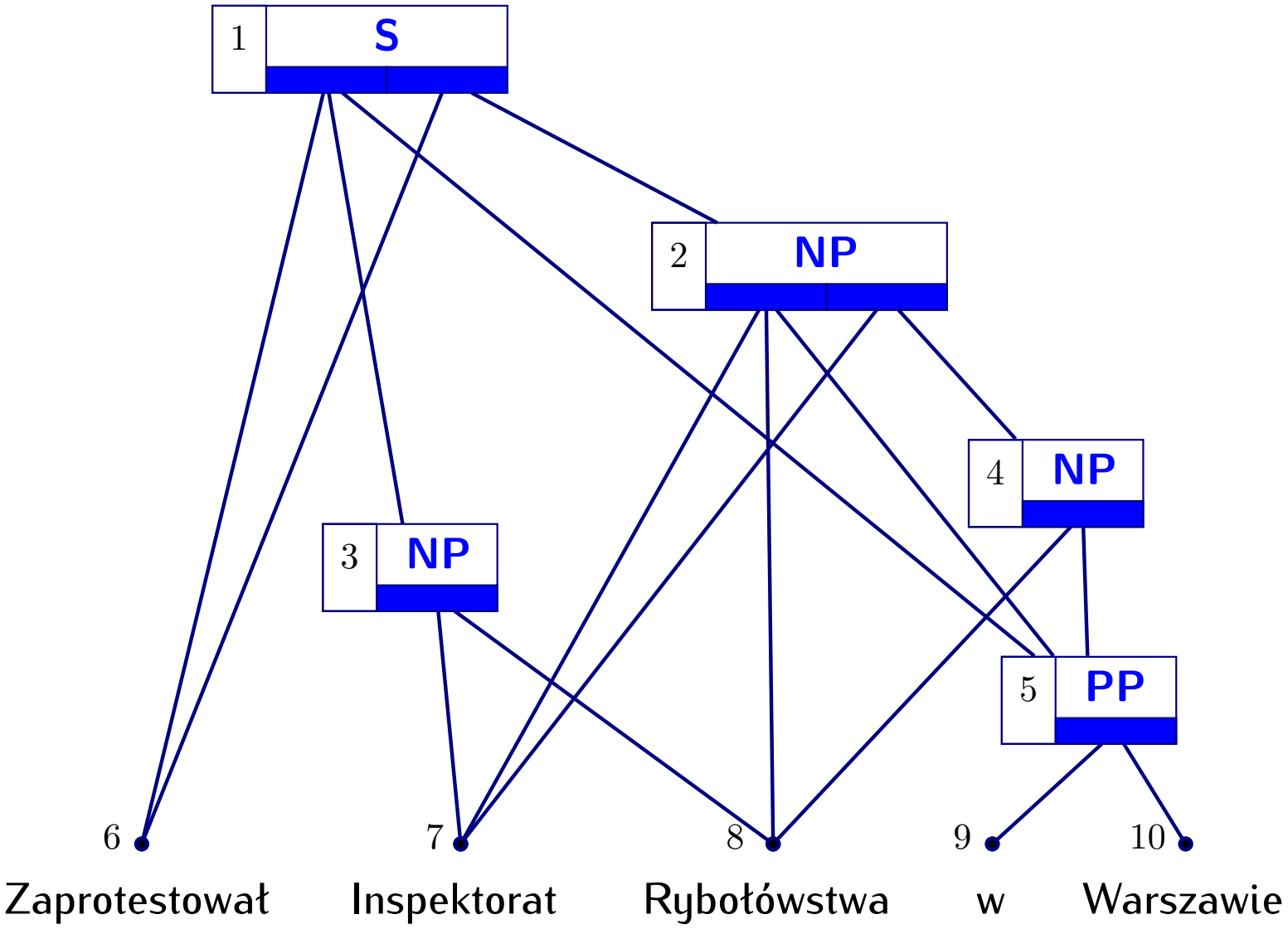
# Atrybuty jednostek nieterminalnych



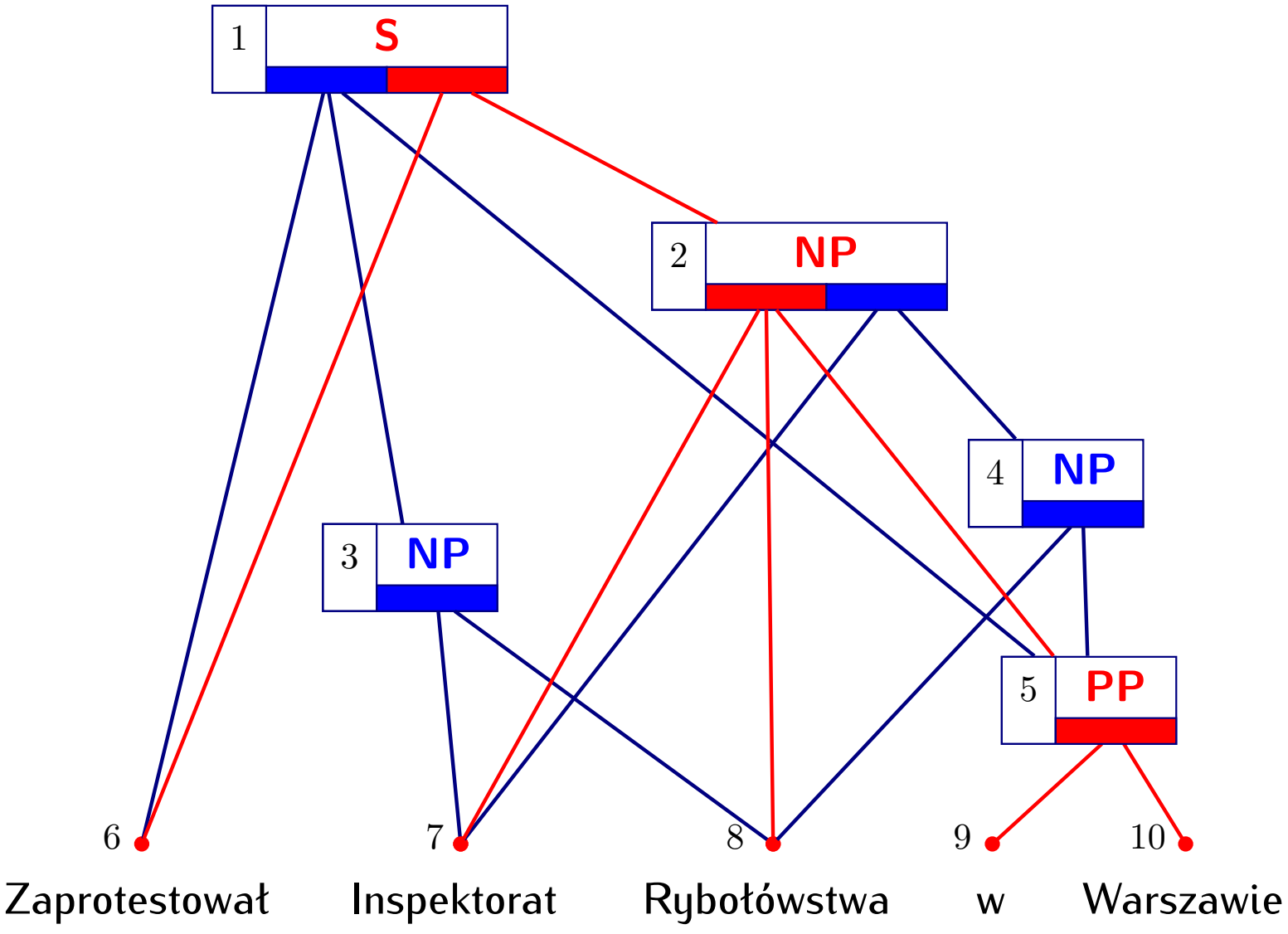
# Definicja zadania

- Celem jest opracowanie algorytmu ujednoznaczniającego składnikowe drzewa składniowe
- Automatyzacja zadania wykonywanego przez anotatorów Składnicy

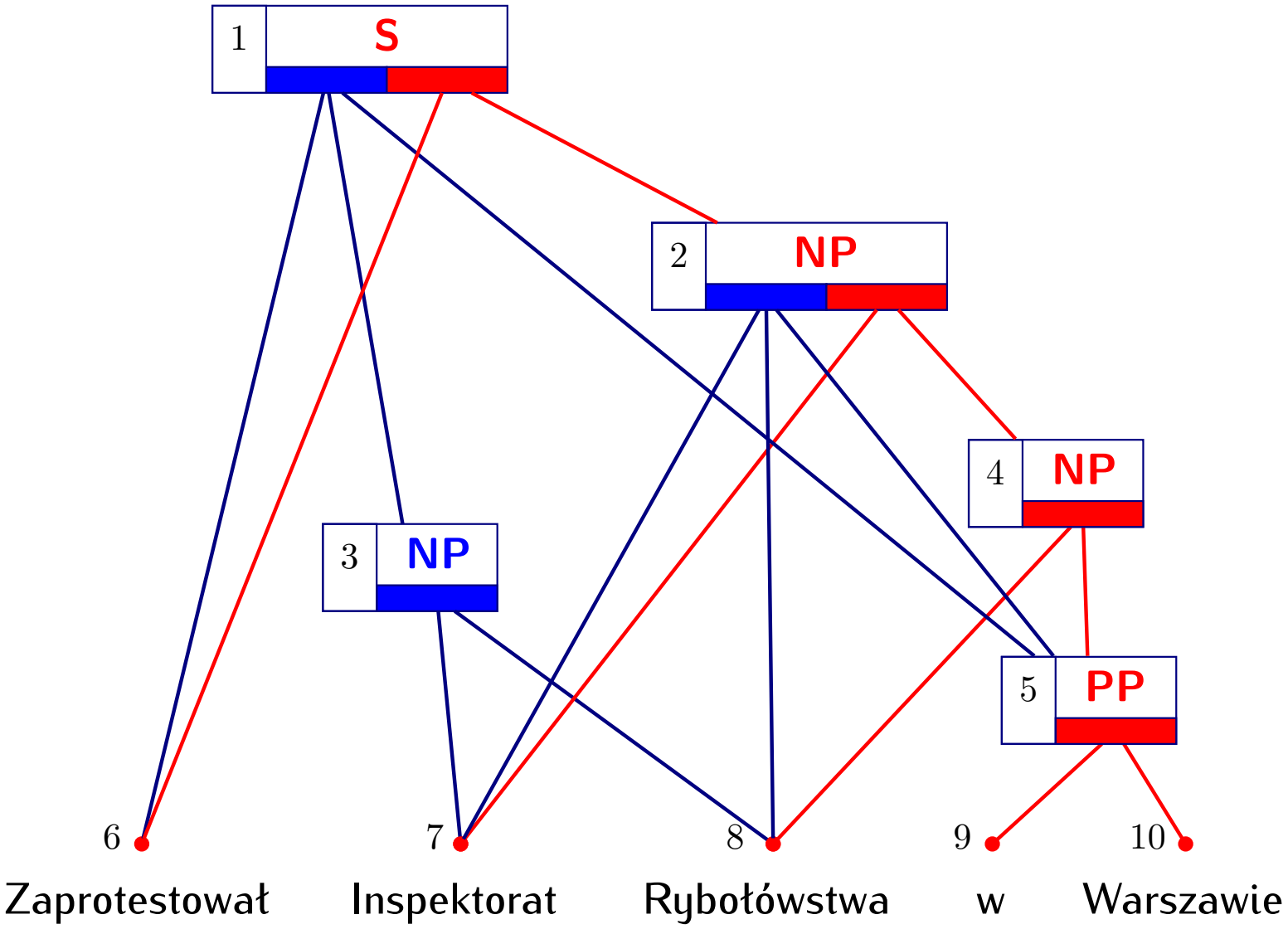
# Reprezentacja lasów składniowych



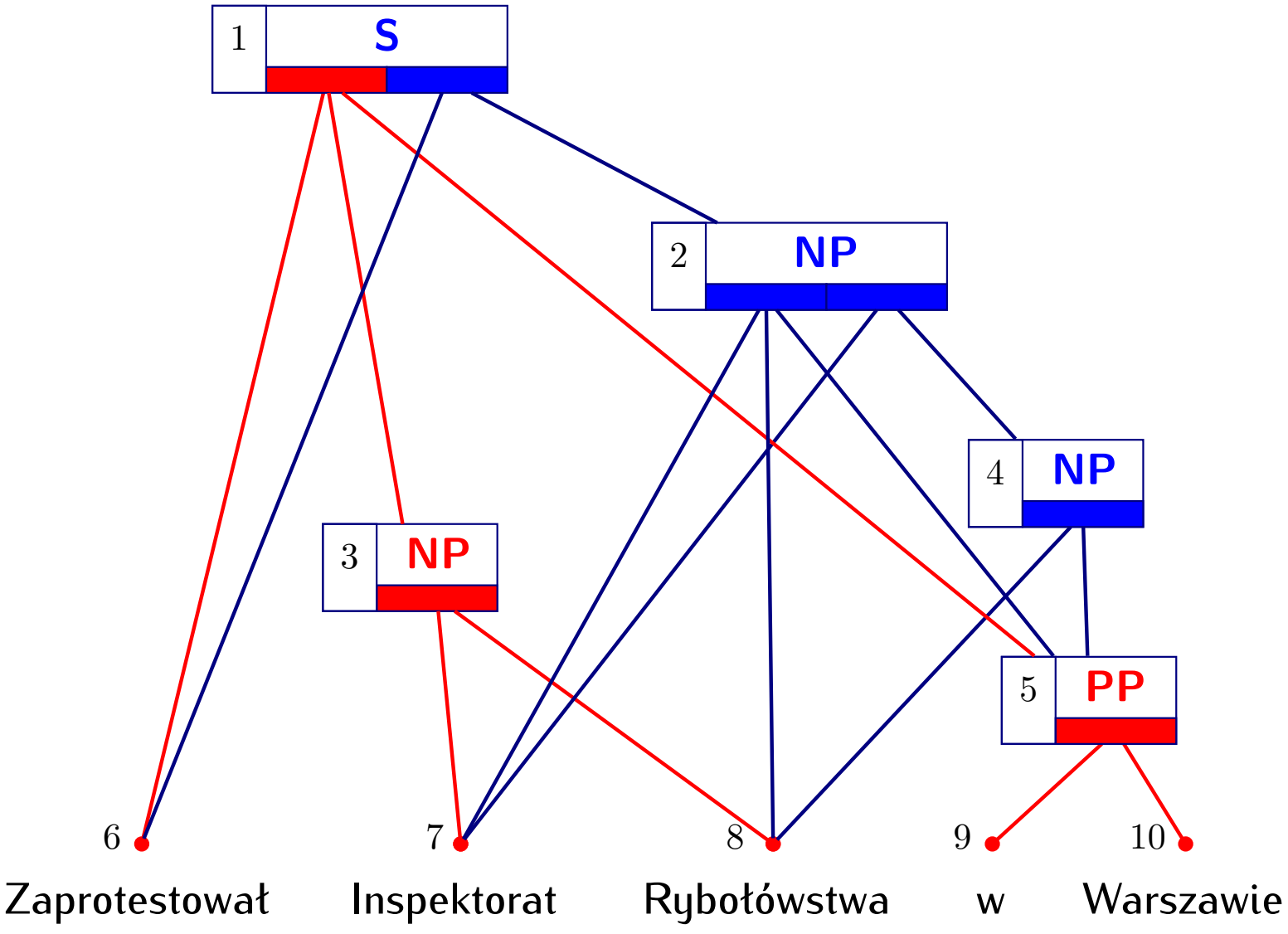
# Reprezentacja lasów składniowych



# Reprezentacja lasów składniowych



# Reprezentacja lasów składniowych



# Bibliografia


- Marcin Woliński, Dominika Rogozińska, *First experiments in PCFG-like disambiguation of constituency parse forests for Polish*. [W:] Zygmunt Vetulani (red.), Proceedings LTC 2013, s. 343–347, 2013.
- Dominika Rogozińska, *Automatyczne metody ujednoznaczniania drzew rozbioru wypowiedzeń w języku polskim jako ostatnia faza przetwarzania parsera Świgra*, praca magisterska, MIM UW, 2016.

# Ewaluacja

- dobrze wybrany wierzchołek: zgadza się zasięg i jednostka nieterminalna

$$\text{precyzja} = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków wybranych przez algorytm}}$$

$$\text{pełność} = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków w danych treningowych}}$$

# Ewaluacja

- dobrze wybrany wierzchołek: zgadza się zasięg i jednostka nieterminalna

$$\text{precyzja} = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków wybranych przez algorytm}}$$

$$\text{pełność} = \frac{\text{liczba dobrze wybranych wierzchołków}}{\text{liczba wierzchołków w danych treningowych}}$$

- w tym zadaniu precyzja i pełność wyników różnią się ±1%
- podajemy tylko $F$ – średnią harmoniczną precyzji i pełności

# Ewaluacja

- $F_L$ – zgodność jednostek nieterminalnych na atrybucie *category*
- $F_A$ – zgodność jednostek nieterminalnych na wszystkich atrybutach
- ULAS – zliczenia dobrze wybranych krawędzi w drzewie zależnościowym

# Ewaluacja

- $F_L$ – zgodność jednostek nieterminalnych na atrybucie *category*
- $F_A$ – zgodność jednostek nieterminalnych na wszystkich atrybutach
- ULAS – zliczenia dobrze wybranych krawędzi w drzewie zależnościowym

- 10-krotna walidacja krzyżowa

# Punkt odniesienia

- cel: ocena użyteczności otrzymanych wyników
- metoda: naśladowanie procesu ręcznego ujednoznaczniania
- model „małpy stukającej w maszynę do pisania”

# Punkt odniesienia

- cel: ocena użyteczności otrzymanych wyników
- metoda: naśladowanie procesu ręcznego ujednoznaczniania
- model „małpy stukającej w maszynę do pisania”
  - małpa wybierająca interpretacje – z równym prawdopodobieństwem każda dostępna możliwość
  - złośliwa małpa – najpierw odrzuca wybór poprawny, wybiera z pozostałych

# Punkt odniesienia

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| złośliwa małpa | 0,859 | 0,696 | 0,808 |

# Punkt odniesienia

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| złośliwa małpa | 0,859 | 0,696 | 0,808 |

$\sim$70% wierzchołków jest jednoznacznych

# Probabilistyczna gramatyka bezkontekstowa

Gramatyka bezkontekstowa z prawdopodobieństwami przypisanymi do każdej z produkcji.


$$[zdanie] \longrightarrow [fw][ff][fw] \qquad 0{,}23$$

Prawdopodobieństwo każdego z możliwych rozbiorów wypowiedzenia jest iloczynem prawdopodobieństw występujących w nim produkcji.

$$P(T) = \prod_{prod \in T} P(prod)$$

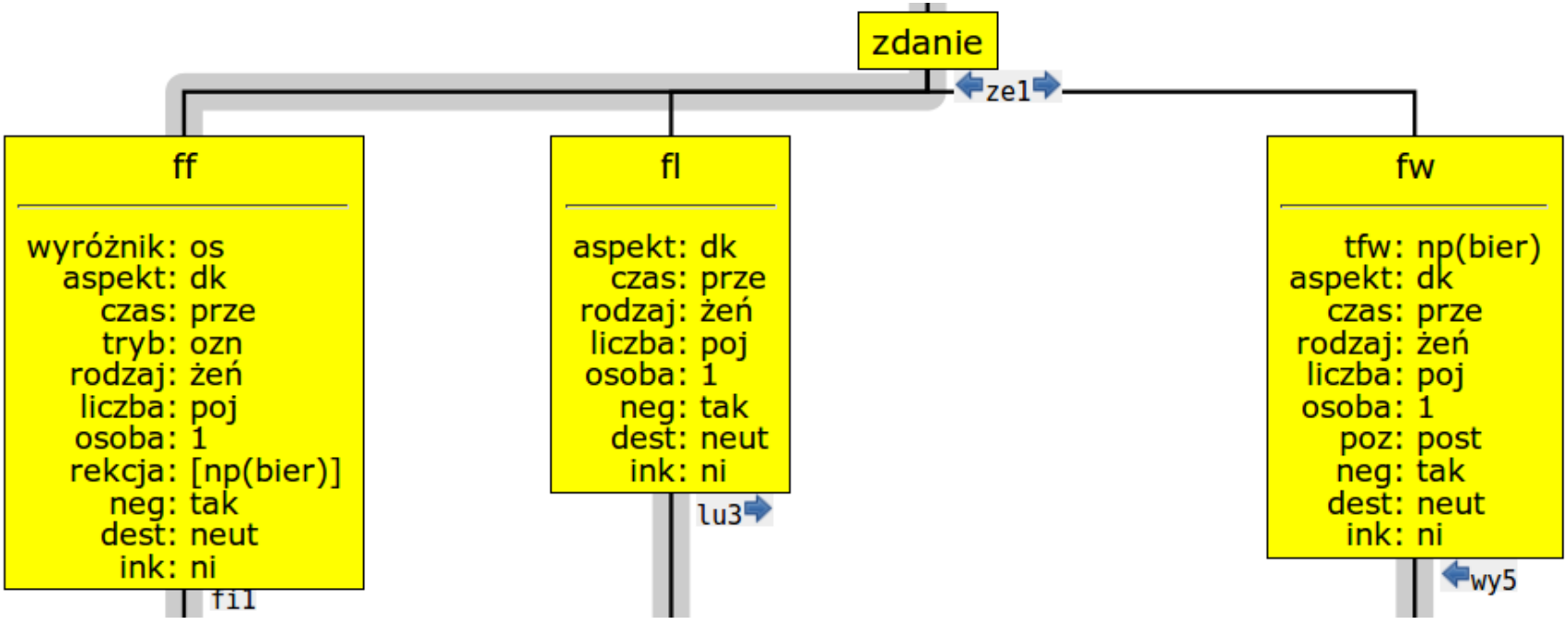
# Ujednoznacznianie

- ekstrakcja reguł gramatyki z danych uczących
- algorytm dynamiczny (bottom-up) dla obliczania prawdopodobieństw poszczególnych drzew

# Wyniki

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| złośliwa małpa | 0,859 | 0,696 | 0,808 |
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG | 0,923 | 0,833 | 0,878 |

# Rozszerzanie opisu produkcji gramatyki




- typ frazy wymaganej
  $[zdanie] \longrightarrow ([ff])[fl][fw@np(bier)]$
- rodzaj, liczba, osoba
  $[zdanie] \longrightarrow ([ff@e@poj@1])[fl@e@poj@1][fw@e@poj@1]$

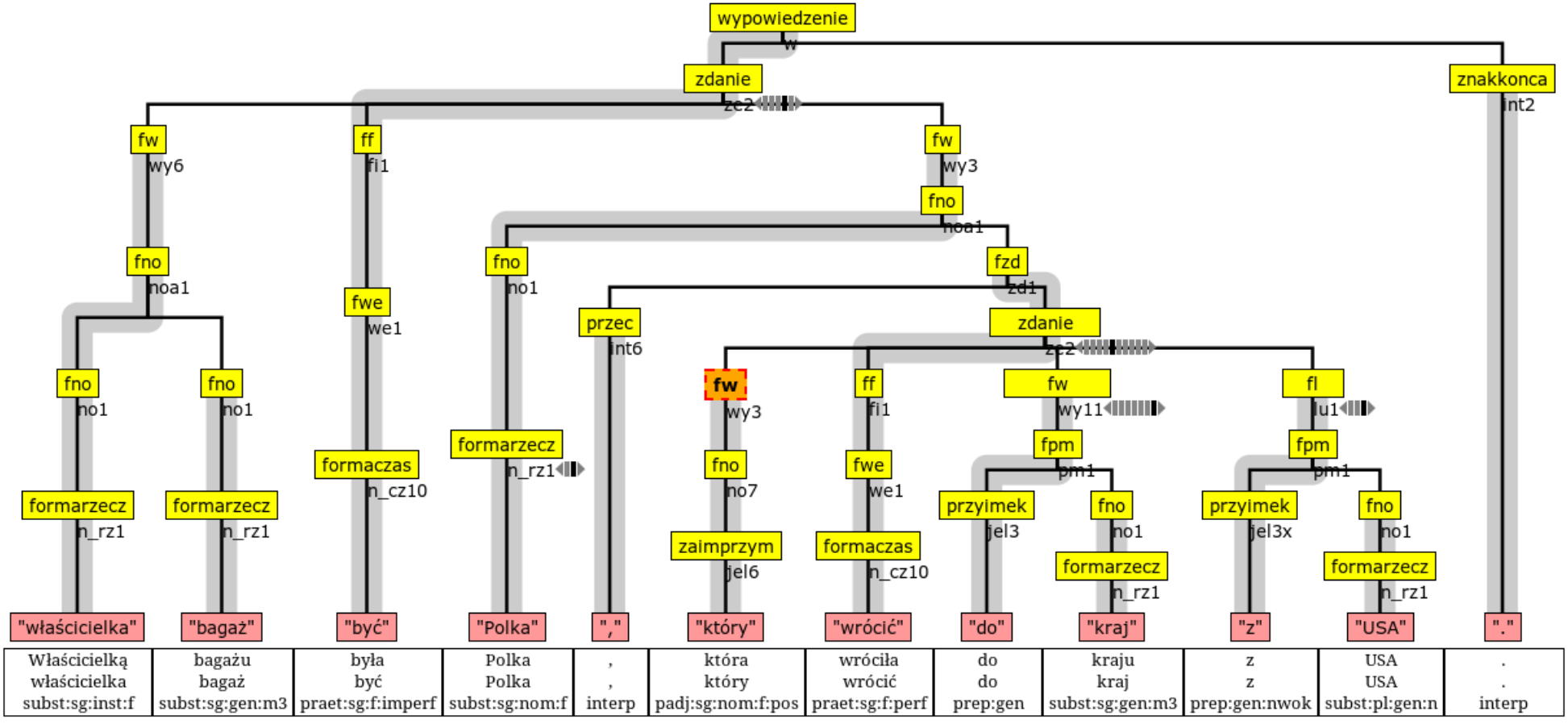
# Wyniki

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| złośliwa małpa | 0,859 | 0,696 | 0,808 |
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG | 0,923 | 0,833 | 0,878 |
| PCFG+typ-fw | **0,941** | 0,875 | **0,921** |
| PCFG+RLO | 0,936 | **0,876** | 0,915 |
| PCFG+typ-fw+RLO | 0,932 | 0,873 | 0,914 |

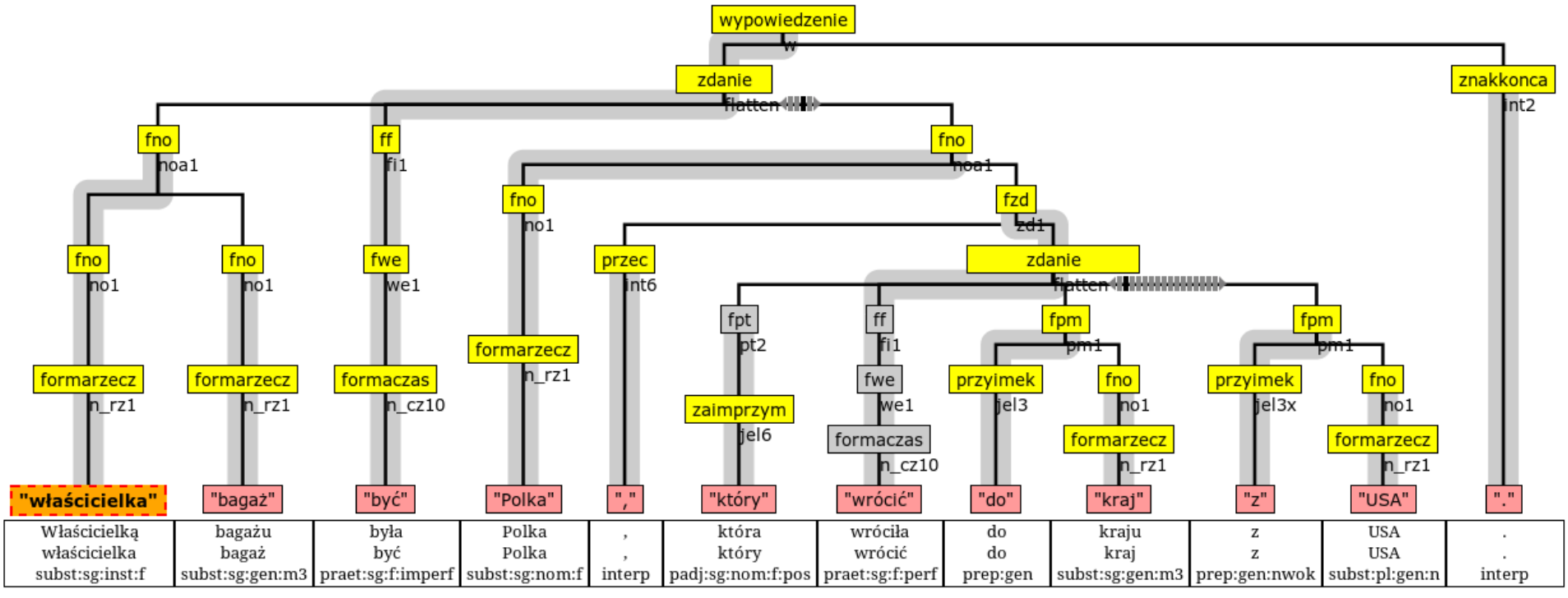
# Nierozpoznane produkcje

| | liczba typów | liczba wystąpień |
|---|---|---|
| PCFG | 3,434 | 171,130 |
| PCFG+typ-fw | 15,472 | 248,946 |
| PCFG+typ-fw+RLO | 61,281 | 416,605 |

# Eksperyment: frazy wymagane i luźne



# Eksperyment: frazy wymagane i luźne



# Eksperyment: frazy wymagane i luźne

fw & fl

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG | 0,923 | 0,833 | 0,878 |
| PCFG+typ-fw | 0,941 | 0,875 | 0,921 |

bez fw & fl

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,935 | 0,890 | 0,831 |
| PCFG | 0,960 | 0,922 | 0,890 |
| PCFG+RLO | 0,943 | 0,925 | 0,859 |

# Granularność wyborów

Błędne wybory potomków w wierzchołkach reprezentujących zdania:

| | za mało składników | za dużo składników |
|---|---|---|
| PCFG +typ-fw | 4,2% | 15,0% |
| PCFG bez fw/fl | 2,1% | 26,3% |

# Model maksymalizacji entropii

$$p_M(t|h) = \frac{\prod_i \alpha_i f_i(t,h)}{\sum_{t' \in \tau(h)} \prod_i \alpha_i f_i(t',h)}$$

- lepszy od PCFG – nie zakłada niezależności cech
- liczba cech nie jest ograniczona jak w PCFG
- zmiana podejścia z lokalnego na globalne

# Model maksymalizacji entropii

- wybór znaczących cech
- wydajny obliczeniowo (?)
- problem z wykładniczą złożonością rozpakowanego lasu
- Miyao Yusuke, Tsujii Jun'ichi. *Maximum entropy estimation for feature forests* In: Proceedings of the second international conference on Human Language Technology Research, 2002, 292–297.

# Wstępne wyniki

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| złośliwa małpa | 0,859 | 0,696 | 0,808 |
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG + typ-fw | 0,941 | 0,875 | 0,921 |
| MaxEnt: podstawowa | 0,958 | 0,905 | 0,921 |
| MaxEnt: podstawowa + typ-fw | 0,958 | 0,905 | 0,922 |

# Konstrukcja wektora cech

Przykłady cech do klasyfikacji

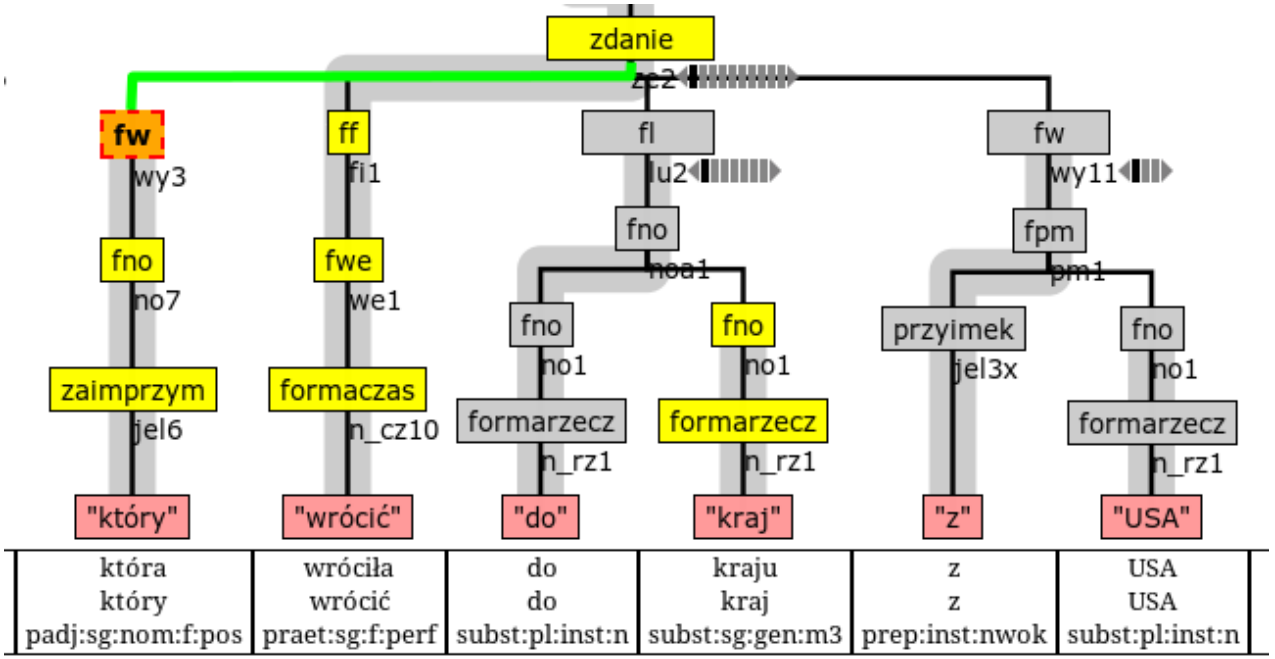
- modelowanie szyku zdania
- aktywne krawędzie
- lokalne konfiguracje (local configuration)
- informacja morfosyntaktyczna
- n-gramy cech
- długość produkcji

# Luźny szyk zdania

- prawa strona produkcji reprezentowana jako zbiór

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| MaxEnt: podstawowa + TFW | 0,958 | 0,905 | 0,922 |
| MaxEnt: szyk luźny | 0,946 | 0,870 | 0,897 |
| MaxEnt: podst. + szyk luźny | 0,958 | 0,890 | 0,936 |

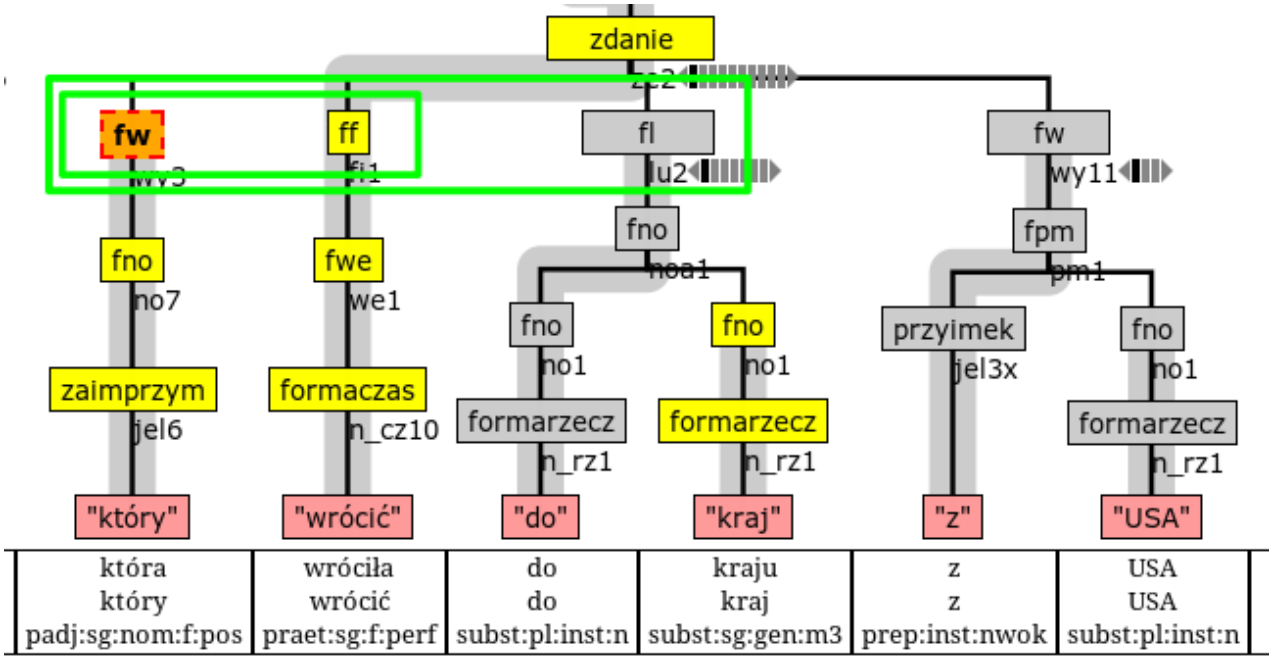
# Aktywne krawędzie



# Aktywne krawędzie

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG + typ-fw | 0,941 | 0,875 | 0,921 |
| MaxEnt: podst. + typ-fw | 0,957 | 0,889 | 0,934 |
| MaxEnt: rozsz. + szyk luźny + a.k. | 0,960 | 0,893 | 0,945 |

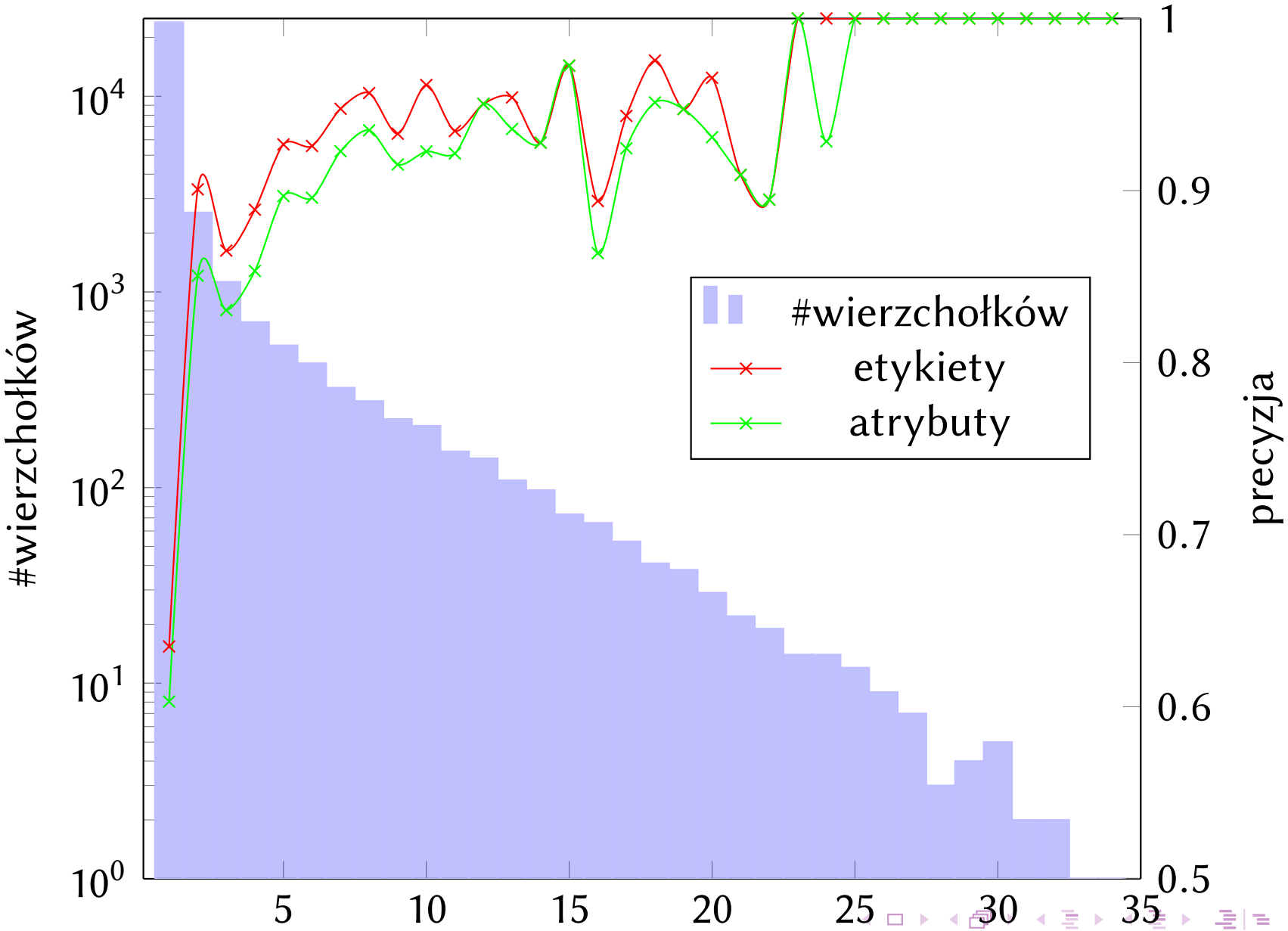
# Lokalne konfiguracje



# Lokalne konfiguracje

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG + typ-fw | 0,941 | 0,875 | 0,921 |
| MaxEnt: rozsz. | 0,960 | 0,893 | 0,945 |
| MaxEnt: rozsz. + lok. konf. | 0,957 | 0,889 | 0,934 |

# Rozpoznane wierzchołki w zależności od ich zasięgu



# Współwystępowanie atrybutów dla produkcji o długości 1

| | $F_L$ | $F_A$ | ULAS |
|---|---|---|---|
| małpa | 0,877 | 0,759 | 0,832 |
| PCFG + typ-fw | 0,941 | 0,875 | 0,921 |
| MaxEnt: rozsz. | 0,960 | 0,893 | 0,945 |
| MaxEnt: rozsz. + wsp. | **0,963** | **0,914** | **0,947** |

# Wagi atrybutów obliczone w wyniku uczenia modelu

| cecha | $\alpha_i$ |
|---|---|
| fw@subj → fno | 3,479327 |
| fno → fpt fno fno fpm | 3,165473 |
| fw@prepnp(omiej) → fpm | 3,079992 |
| fno → fpt fno fno | 2,840276 |
| fw@prepnp(obier) → fpm | 2,440277 |
| zdanie → zdanie przec zdanie | 1 |
| zdanie → fw fw fl fw ff fl fw | 1 |
| zdanie → fw fw fl fw ff fw fl | 1 |
| zdanie → fw@np(dop) fl ff fw@subj fl fw@np(bier) | 1 |
| fno → fpm fno | 0,415269 |
| fno @1 | 0,2770649 |
| fw@prepnp(wmiej) → fpm | 0,2583525 |
| fno → fpt | 0,2253499 |

# Podsumowanie i perspektywy

- `http://swigra.nlp.ipipan.waw.pl/`
- dotychczasowe eksperymenty były prowadzone w wyidealizowanych warunkach ręcznie ujednoznacznianego fleksyjnie NKJP1M
- najlepsze wyniki: $F_L$ **0,963**, $F_A$ **0,914**
- ULAS **0,947** lepszy niż w eksperymencie opisanym w Wróblewska and Woliński, 2012 – 0,922
- wada: pracujemy na zdaniach zaakceptowanych przez Świgrę
- zaleta: pracujemy na strukturach zaakceptowanych przez Świgrę

# Podsumowanie i perspektywy

Przyszłe prace

- analiza typów popełnianych błędów
- lepszy wybór cech
- powiększanie Składnicy
- dalszy rozwój Świgry

# Dobrze ujednoznacznione drzewa

| | $F_L$ | $F_A$ | ULAS | $F_L$ | $F_A$ |
|---|---|---|---|---|---|
| MaxEnt: podstawowa | 0,958 | 0,905 | 0,921 | 0,605 | 0,513 |
| MaxEnt: podst. + typ-fw | 0,958 | 0,905 | 0,922 | 0,605 | 0,514 |
| MaxEnt: rozszerzona | 0,963 | 0,914 | 0,947 | 0,631 | 0,557 |