Simultaneous error detection at two levels of syntactic annotation

Adam Przepiórkowski and Michał Lenart

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

13 July 2012

LAW VI @ ACL 2012

Aim

Given:

- the National Corpus of Polish, i.e.,
- a corpus annotated syntactically at two levels:
    - **shallow syntax** (S),
    - **deep syntax** (D).

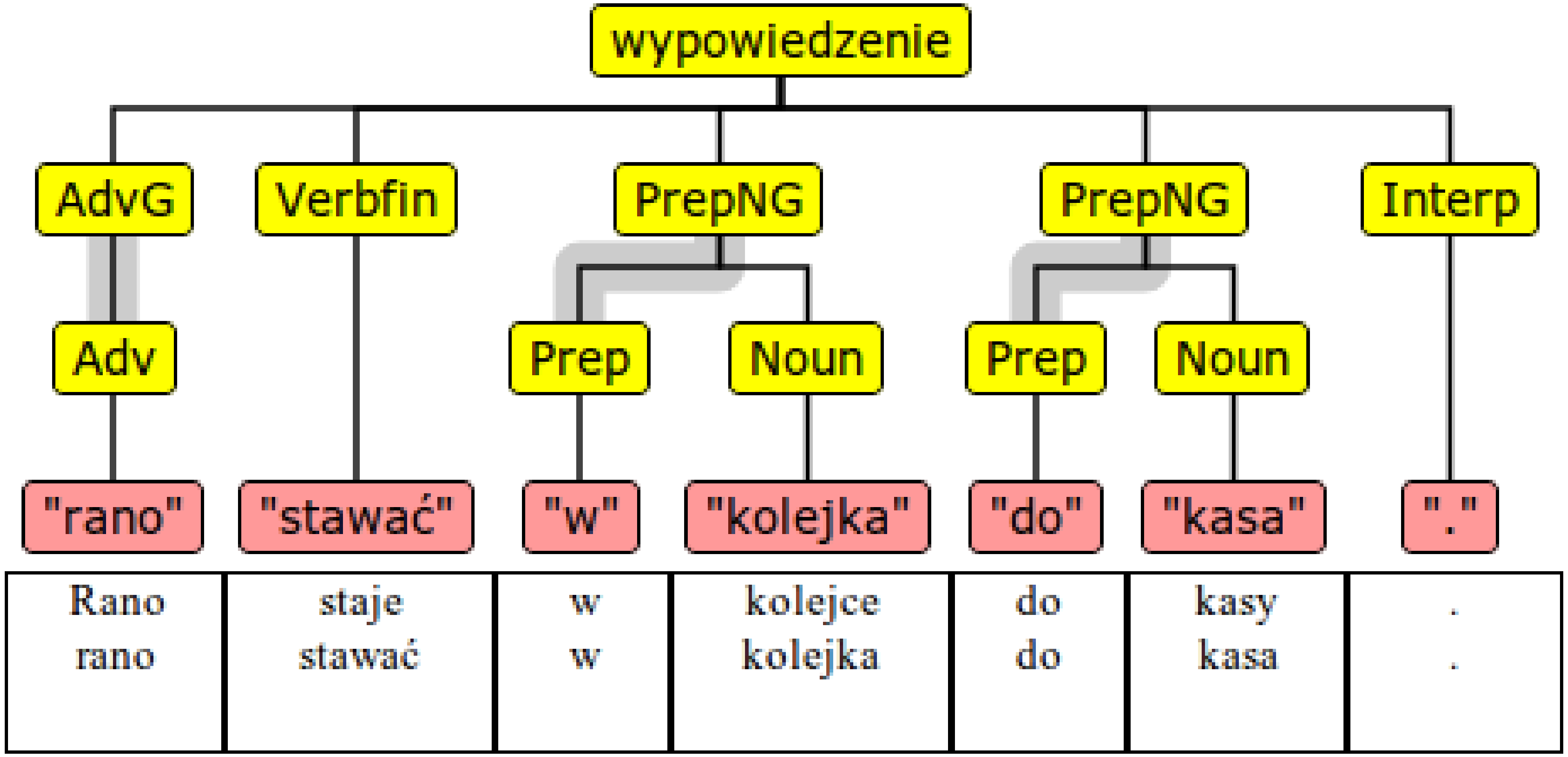
Aim:

- **find errors** at both levels of annotation.

Problem:

- very **different linguistic assumptions** at both levels.

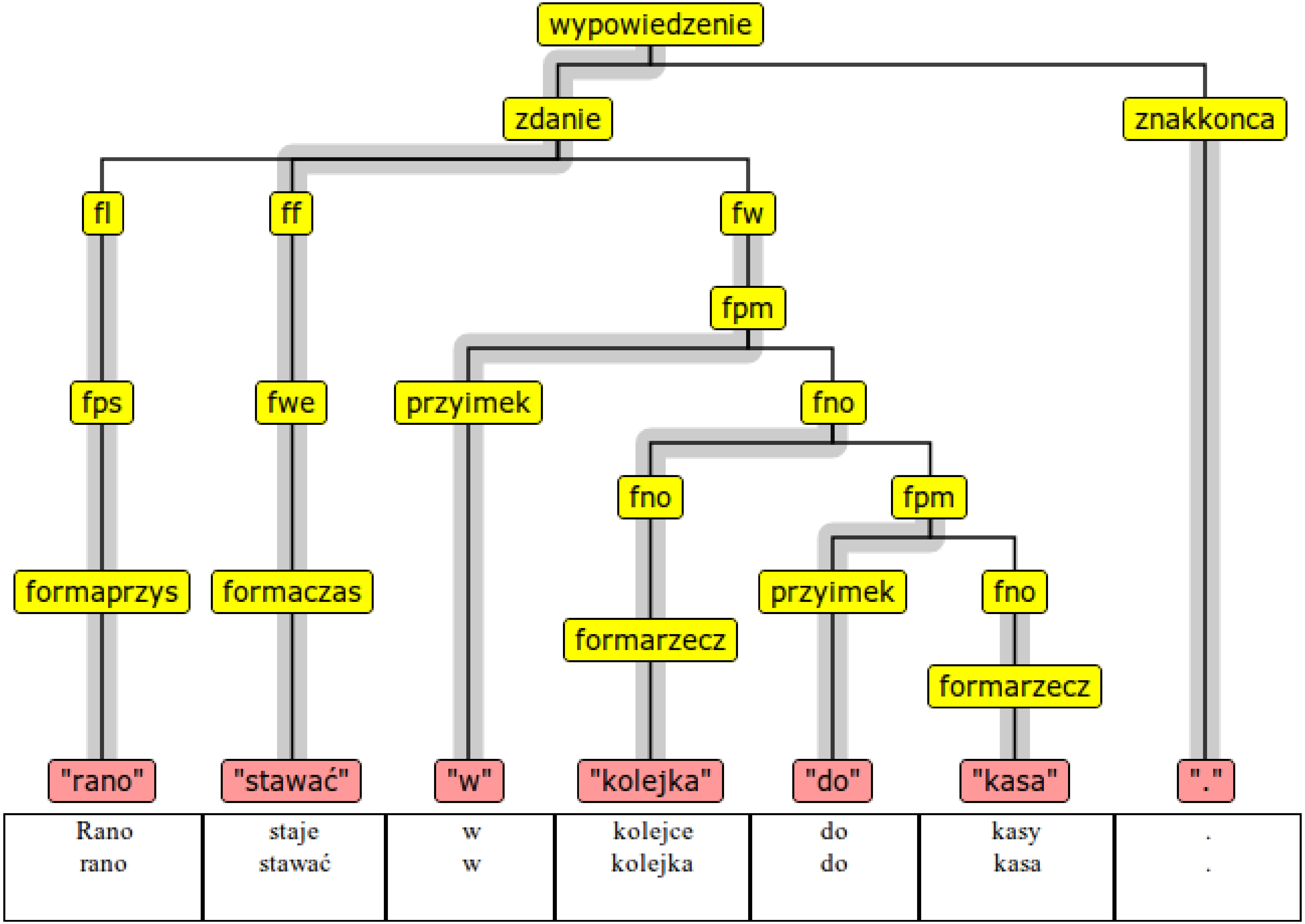
Example: shallow annotation



Rano rano	staje stawać	w w	kolejce kolejka	do do	kasy kasa	. .

Rano staje w kolejce do kasy.
morning join.3.sg in queue to cash desk
'In the morning, (s)he queues to the cash desk.'

Example: deep annotation



Rano rano	staje stawać	w w	kolejce kolejka	do do	kasy kasa	. .

Measures 1

Approach:

- **define precision** of one level against the other,
- **find fragments violating precision**.

Shallow precision P_s:

$$P_s = \frac{|\{w : \exists G \;\; w \in \mathit{yield}(G) \wedge c(w, G)\}|}{|\{w : \exists G \;\; w \in \mathit{yield}(G)\}|},$$

where:

- w ranges over words,
- G ranges over (non-sentential) shallow groups,
- $c(w, G)$ is the compatibility predicate (of w across the two levels).

$c(w, G)$ is true iff there exists a deep phrase F such that:

- $w \in \mathit{yield}(F)$, and
- G and F have the same lexical heads.

Measures 2

Labelled shallow precision lP_s:

- additionally require that F and G have matching labels (e.g., both indicating a PP).

Deep precision (**unlabelled** P_d and **labelled** lP_d):

- as shallow precision,
- but only consider words w *more or less directly* contained in a phrase of a type corresponding to the types of shallow groups (e.g., NP, PP, but not sentential clause).

(More careful definitions in the paper.)

Experiment and results

Experiment:

- **7600 sentences** from the National Corpus of Polish,
- **manually annotated** at both levels.

Unlabelled results (all mean micro-average):

- $P_s = 98.7\%$ and $P_d = 93.4\%$,
- $P_d < P_s \implies$ more common for the shallow level to miss (parts of) deep-level constituents, than the other way round.

Analysis:

- 50 sentences with non-perfect matching examined manually,
- 104 word-level discrepancies found:
  1. false positives (over 50%),
  2. result of controversial design decisions at the shallow level (15%),
  3. real differences, i.e., possible errors (33%).

Results (contd.)

Errors discovered:

- wrong treatment of discontinuities (at **D**),
- different analyses of particles,
- different analyses of adverbs, etc.

Labelled results (all mean micro-average):

- $IP_s = 95.1\%$ and $IP_d = 91.1\%$.

Analysis of label differences:

- relative pronouns (marked as pronoun vs. NP, Adv, etc.),
- prepositional constructions (some marked as adverbials at **S**).

Estimation: out of 1882 non-matching sentences (out of 7600 examined), around 500 contain **real errors**.

Conclusion:

Useful for finding errors in manually annotated corpora.