# From *Lexical Functional Grammar* to Enhanced *Universal Dependencies*

Adam Przepiórkowski and Agnieszka Patejuk


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

UNIVERSITY OF WARSAW



LFG 2018
Vienna, 17 July 2018

# Introduction 1

**Universal Dependencies** (UD; Nivre *et al.* 2016; version 2.2 announcement):

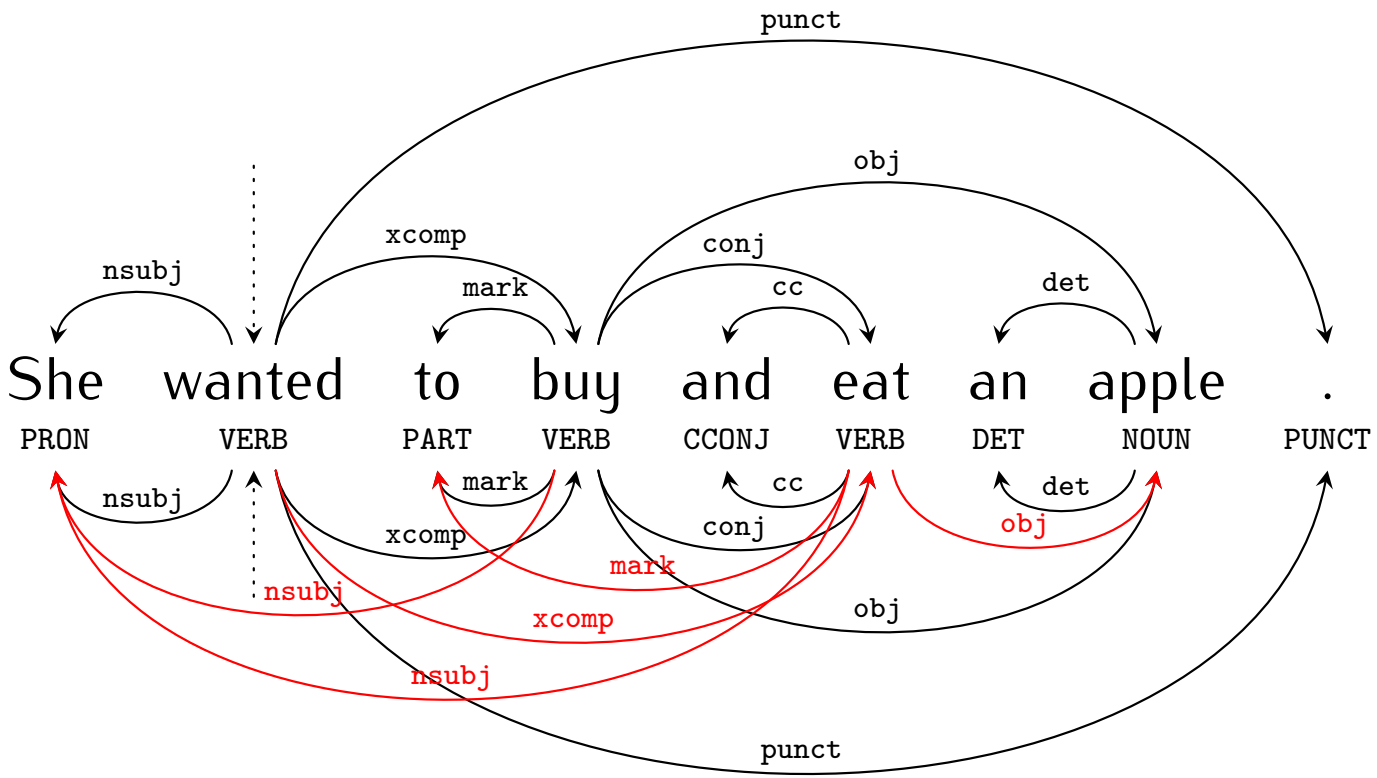
> *a project that seeks to develop **cross-linguistically consistent treebank annotation for many languages** with the goal of **facilitating multilingual parser development**, cross-lingual learning, and parsing research from a language typology perspective.*

**Version 2.2** (released on 8 July 2018) – **122 treebanks** of **71 languages**:

Afrikaans, Amharic, Ancient Greek, Arabic, Armenian, Basque, Belarusian, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Marathi, Naija, North Sami, Norwegian, Old Church Slavonic, Old French, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Swedish Sign Language, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri and Yoruba.

# Introduction 2

An example (`http://universaldependencies.org/u/overview/syntax.html`):



# Introduction 3

**New UD treebank of Polish**:

- **converted** from an LFG parsebank of Polish,
- officially available **since July 2018** (UD release 2.2) – unofficially since February 2018,
- **17,246 sentences** (130,967 tokens).

**Outline**:

- **conversion** in two stages:
  - from LFG structures **to LFG-like dependencies**,
  - from LFG-like dependencies **to enhanced UD**,
- what is **lost in translation**.

# Conversion to LFG-like dependencies 1

**Input to conversion**:

- **f-structures**,
- **c-structures** – only terminals (tokens) and preterminals (categories).

**Example 1**:

- – Słowo daję, że się nie gniewam.
  word.ACC give.1.SG that RM NEG be_angry.1.SG
  'I give you my word that I am not angry.'

- PRED 'dawać<[10:pro], [6:słowo], [2:gniewać_się]>'
  COMP 2: PRED 'gniewać_się<[4:pro]>', SUBJ 4: PRED 'pro'
  OBL-STR 6: PRED 'słowo'
  SUBJ 10: PRED 'pro'
  0

# Conversion to LFG-like dependencies 2

Recall:

- – Słowo daję, że się nie gniewam.
  word.ACC give.1.SG that RM NEG be_angry.1.SG
  'I give you my word that I am not angry.'

Dependencies **read off the f-structure**:

- COMP (between f-structures 0 and 2),
- OBL-STR (between f-structures 0 and 6),
- ROOT (to f-structure 0).

Which **tokens** correspond to these f-structures?

- 0: ***daję*** 'give', but also the initial dash and the final period,
- 2: ***gniewam*** 'be angry', but also *że* 'that', *się* RM, *nie* NEG and the comma,
- 6: ***słowo*** 'word'.

# Conversion to LFG-like dependencies 3

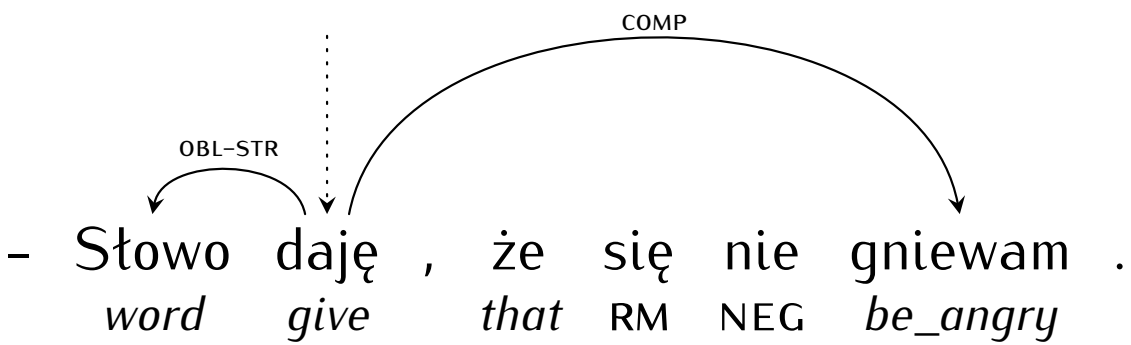
Definition:

- **co-heads**: tokens whose preterminals map to the same f-structure.

**First step** (of the first stage):

- **select true heads** among co-heads (as on the previous slide),
- **algorithm**: very simple, on the basis of preterminal labels,
- **example**: the verb ***gniewam*** 'be angry' wins with the complementiser *że*, the reflexive marker *się*, the negative particle *nie* and the comma.

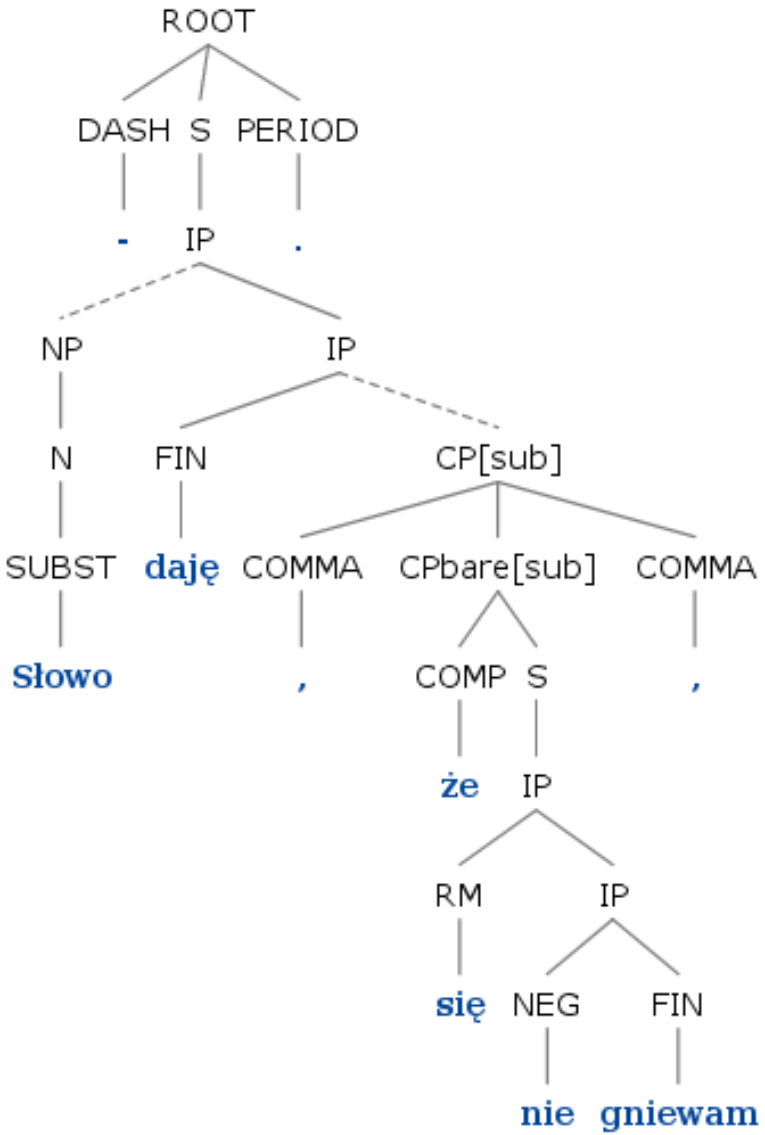
**Result**:

- 


**Second step**: other dependencies mirror c-structure preterminals.

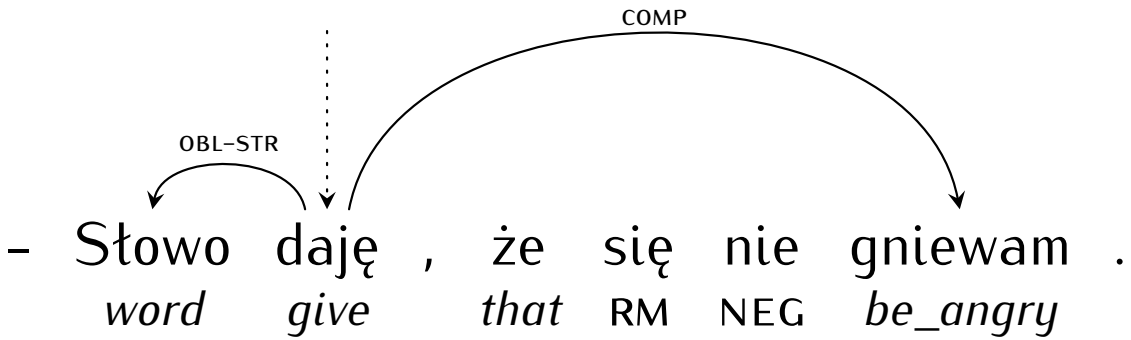
# Conversion to LFG-like dependencies 4

Full **c-structure**:

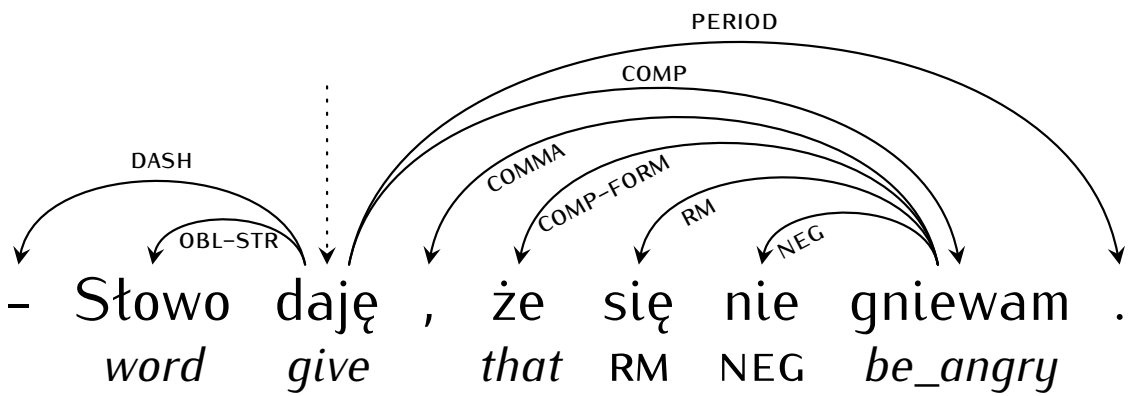


# Conversion to LFG-like dependencies 5

The **backbone** after the first step (repeated):




The **LFG-like dependency representation** after the second step (i.e., after the first stage):

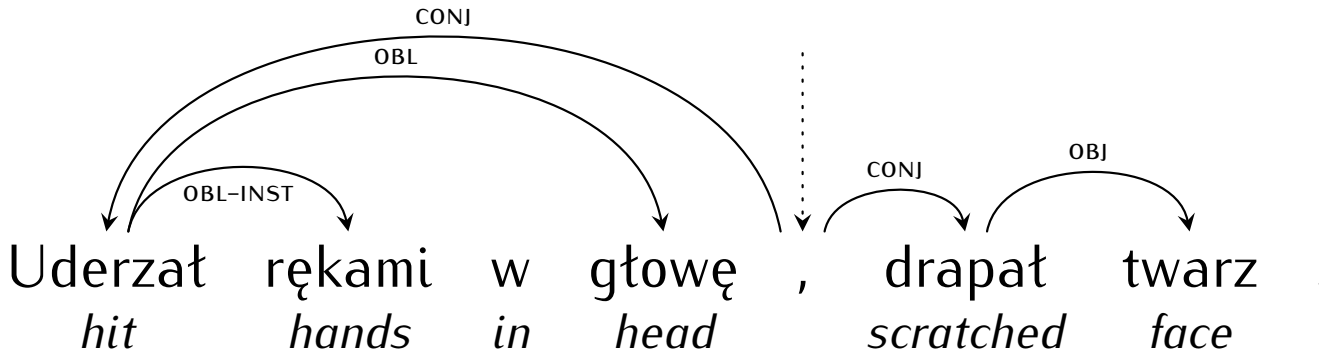


# Conversion to LFG-like dependencies 6

**Example** 2, involving coordination:

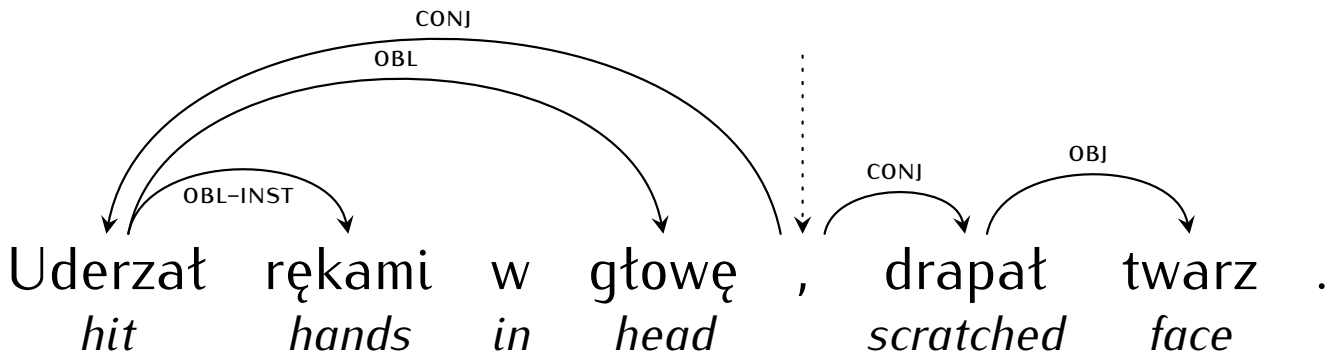
- Uderzał rękami w głowę, drapał twarz.
  hit.3.SG.M hands.INST in head.ACC scratched.3.SG.M face.ACC
  'He pounded his head with his fists, scratched his face.'
- 0 { 47 PRED 'drapać<[81:pro], [50:twarz]>', OBJ 50 PRED 'twarz', SUBJ 81 PRED 'pro' }, 1 PRED 'uderzać<[81:pro], [39:ręka], [38:głowa]>', OBL-INST 39 PRED 'ręka', OBL 38 PRED 'głowa', SUBJ [81]

**Set membership** in coordination translated into CONJ, resulting in the following **backbone** (result of first step):

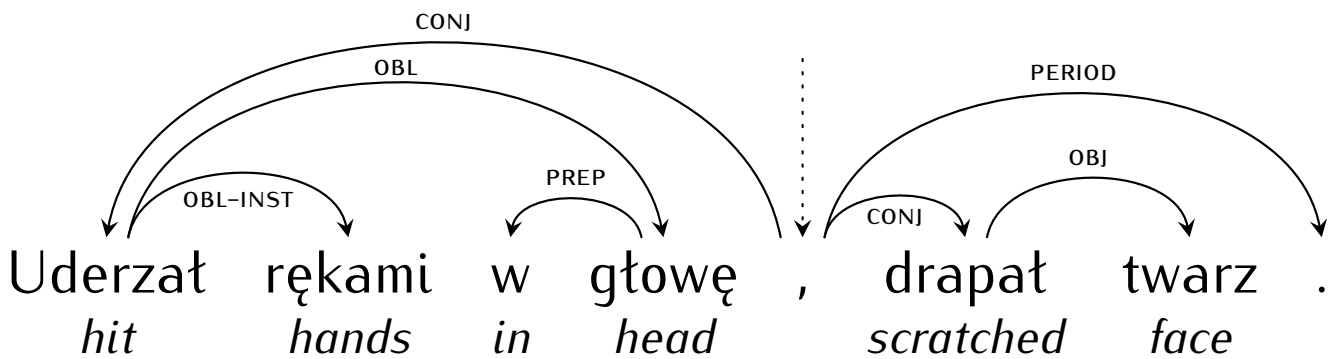
- 

# Conversion to LFG-like dependencies 7

Result of first step repeated:

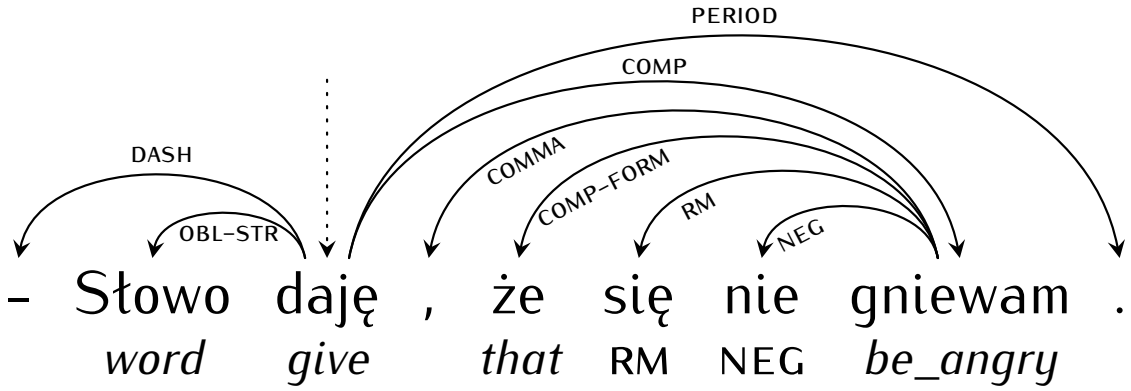



**After the second step** (of the first stage) – note that *w* 'in' is an asemantic preposition here:

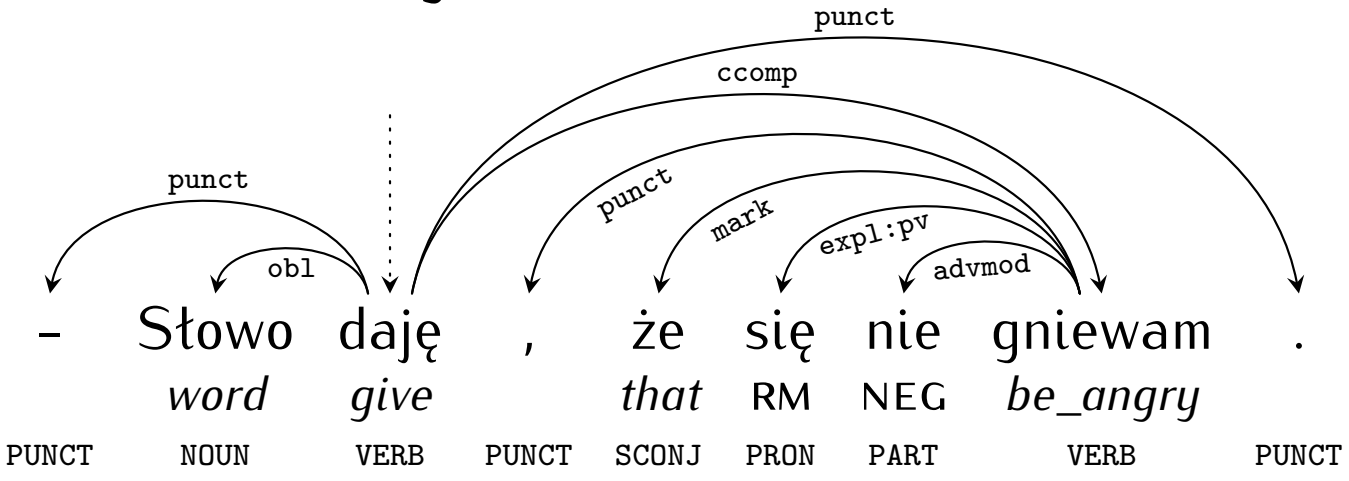


# Conversion to enhanced UD 1

**Second stage** – in the simplest (but very rare) case, it is sufficient to translate dependency labels. Example 1 **after the first stage** (repeated):




**After the second stage**:



# Conversion to enhanced UD 2


Usually, the **dependency graph needs to be rearranged**:

- to **rearrange coordination dependencies**,
- more generally, to **reverse dependencies between function words and content words**.

**Coordination in LFG-like dependencies**:

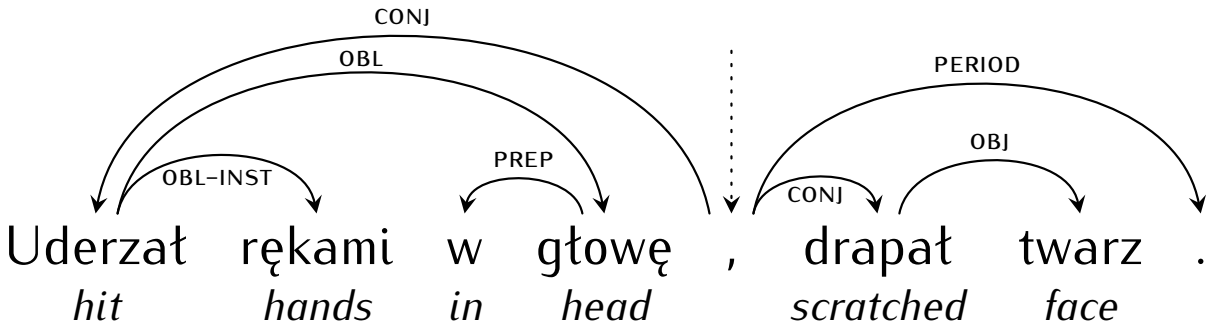
- headed by the conjunction,
- conjuncts are its CONJ dependents.

**Coordination in UD**:

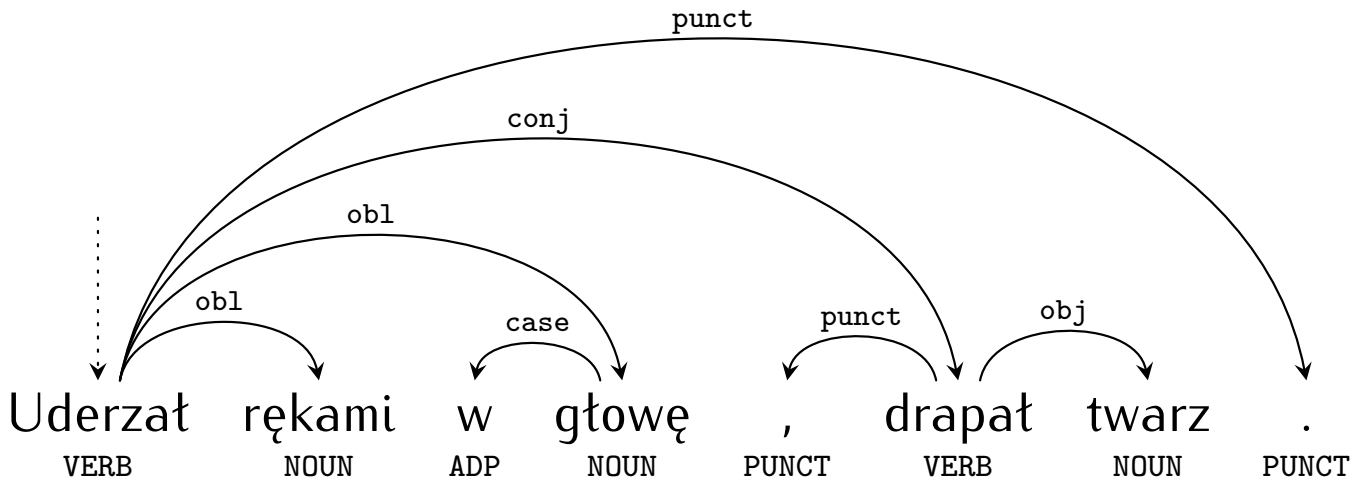
- headed by the 1st conjunct,
- all other conjuncts are its CONJ dependents,
- conjunction is a CC dependent of the following conjunct.

# Conversion to enhanced UD 3

**after 1st stage:**




**after 2nd stage:**



## Conversion to enhanced UD 4

**Reversing dependencies between function words and content words**; in UD:

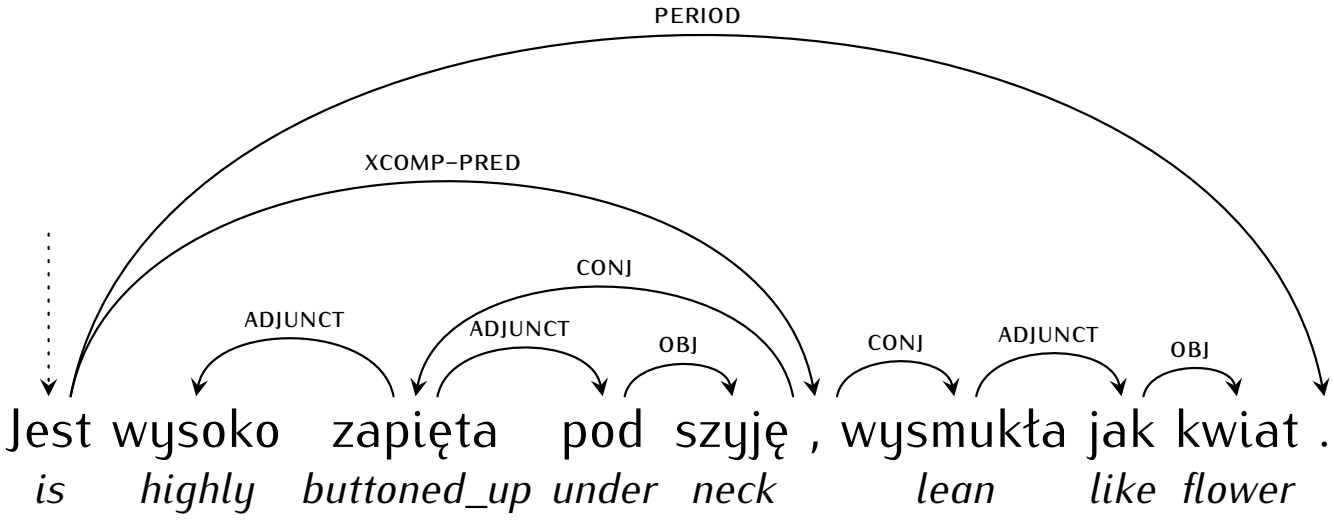
- **prepositions** (both: asemantic and semantic) are dependents of nouns,
- **numerals** are dependents of nouns (contrary to morphosyntactic tests on headedness in Polish),
- **auxiliaries and copulas** are dependents of verbs.

**Example 3**:

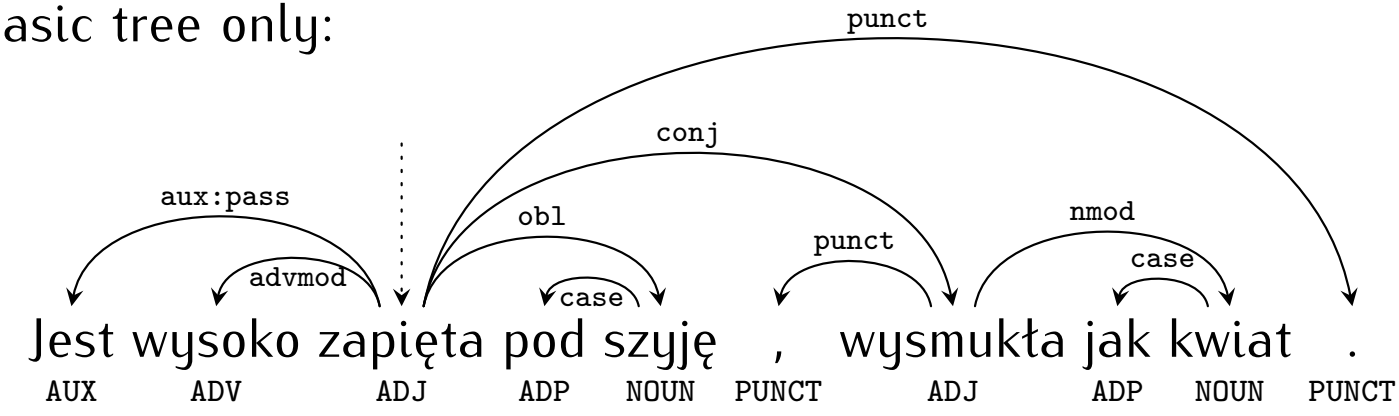
- Jest wysoko zapięta pod szyję, wysmukła
  is.3.SG highly buttoned_up.NOM.SG.F under neck.ACC lean.NOM.SG.F
  jak kwiat.
  like flower.NOM.SG.M
  'She is buttoned up high to the neck, lean like a flower.'

# Conversion to enhanced UD 5

**after 1st stage**:

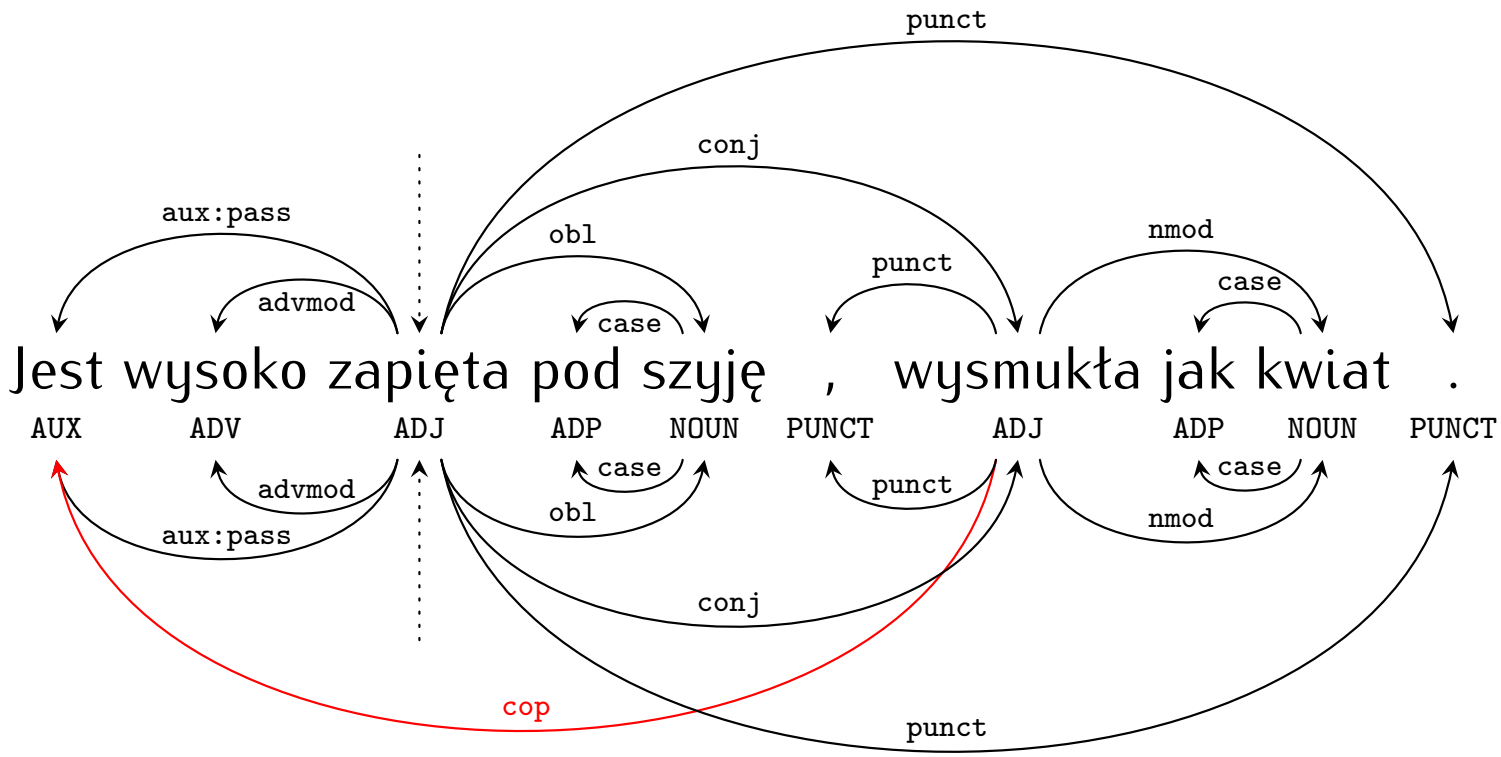



**after 2nd stage** – basic tree only:



# Conversion to enhanced UD 6

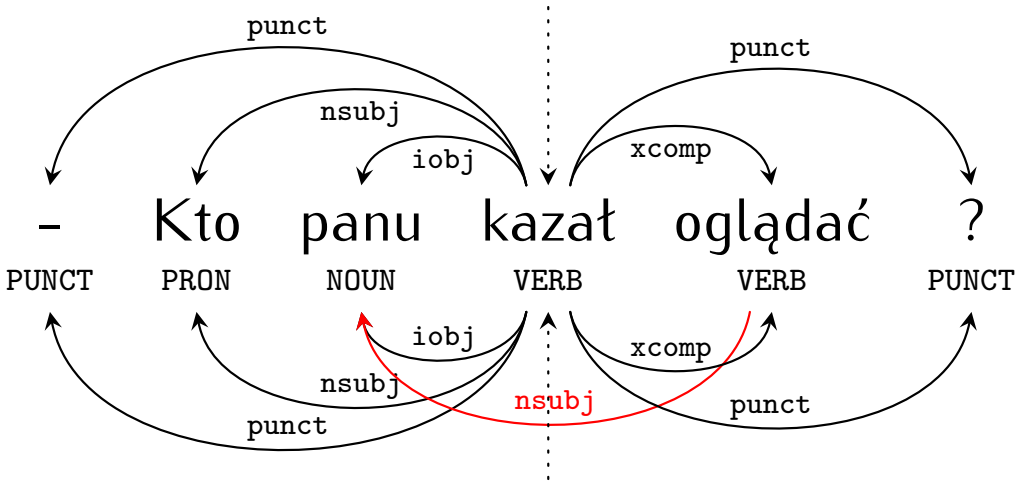
**after 2nd stage** – also (partial) enhanced structure:




Note: the auxiliary/copula *jest* 'is' is **a shared dependent** of the passive participle (aux:pass) and the predicative adjective (cop).

# Lost in translation

**Dependent-sharing** is not a problem for enhanced UD (also in control).

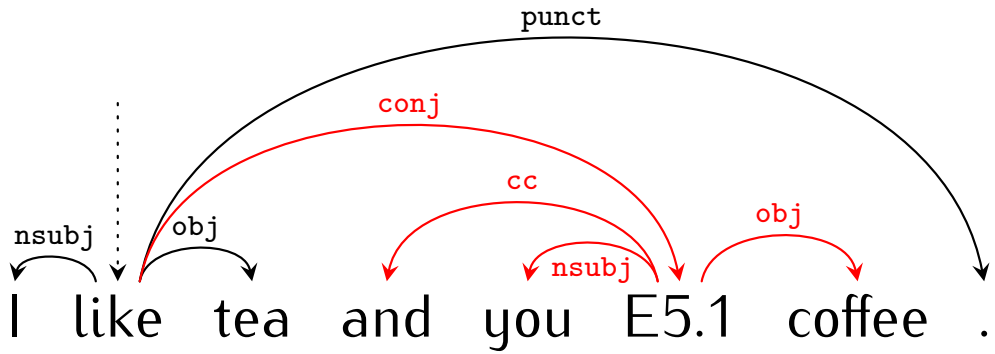
- – Kto panu kazał oglądać?
  who.NOM.SG.M you.DAT ordered.3.SG.M watch.INF
  'Who asked you to watch?'
- 0 PRED 'kazać<[14:kto], [12:pan], [20:oglądać]>'
  XCOMP 20 PRED 'oglądać<[12:pan]>'
  SUBJ 12 PRED 'pan'
  OBJ-TH [12]
  SUBJ 14 PRED 'kto'
- 

# Lost in translation – *pro*-drop 1

So **what kind of information** – if any – is **lost in translation** from LFG to UD?

The main reason for loss of information: **prohibition on empty dependents**.

Note: there is **no general prohibition on empty nodes in UD**, e.g.:




http://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis

# Lost in translation – *pro*-drop 2

Problem in expressing **control** relations, etc.:

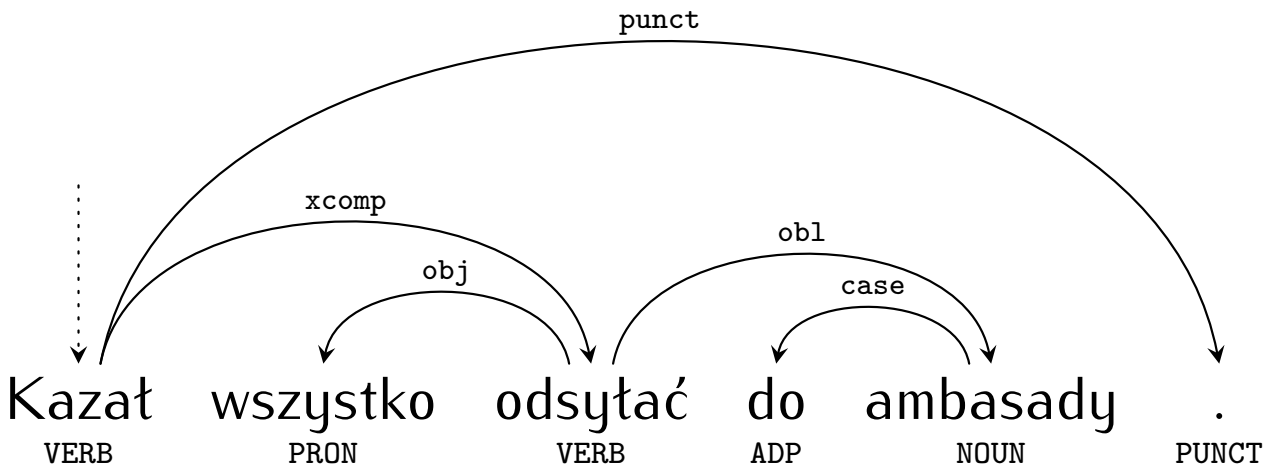
- Kazał wszystko odsyłać do ambasady.
  ordered.3.SG.M all.ACC send_back.INF to embassy
  'He ordered to send everything back to the embassy.'

- 

# Lost in translation – *pro*-drop 2

Problem in expressing **control** relations, etc.:

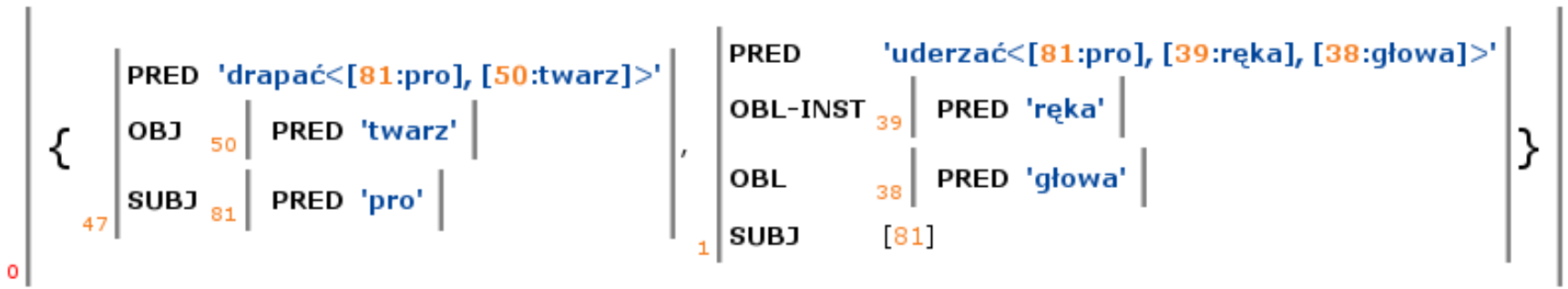
- Kazał wszystko odsyłać do ambasady.
  ordered.3.SG.M all.ACC send_back.INF to embassy
  'He ordered to send everything back to the embassy.'



# Lost in translation – *pro*-drop 3

Problem in expressing shared *pro*-dropped dependents in **coordination**:

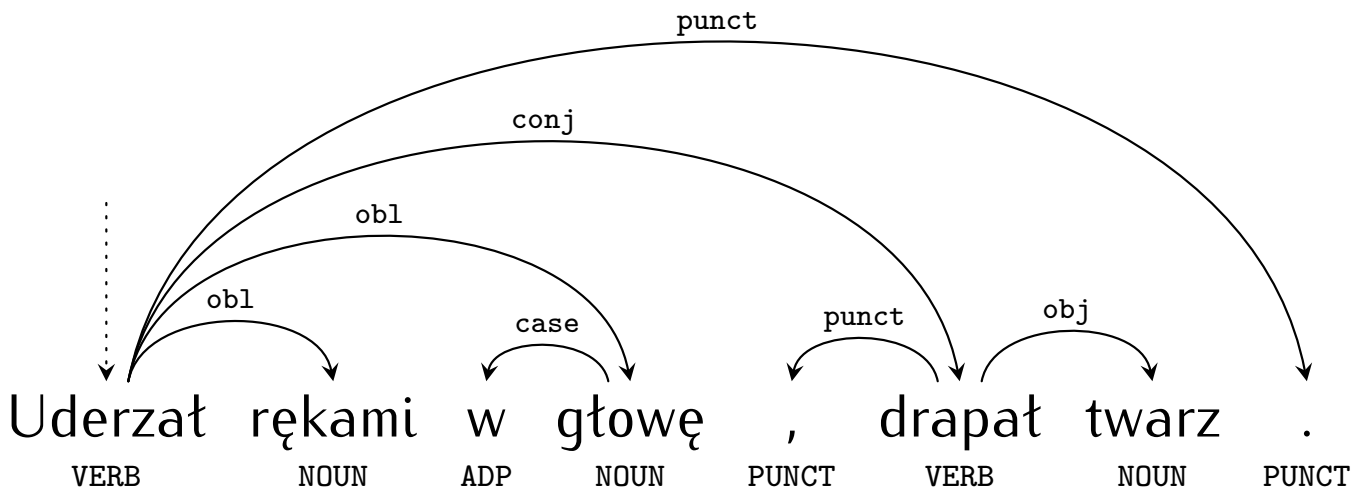
- Uderzał rękami w głowę, drapał twarz.
  hit.3.SG.M hands.INST in head.ACC scratched.3.SG.M face.ACC
  'He pounded his head with his fists, scratched his face.'

- 

# Lost in translation – *pro*-drop 3

Problem in expressing shared *pro*-dropped dependents in **coordination**:

- Uderzał rękami w głowę, drapał twarz.
  hit.3.SG.M hands.INST in head.ACC scratched.3.SG.M face.ACC
  'He pounded his head with his fists, scratched his face.'



# Lost in translation – multiple edges

Another structural **prohibition**: **up to one edge from A to B**.

Problem in cases of the **haplology of the reflexive marker** (Kupść 1999, Patejuk and Przepiórkowski 2015):

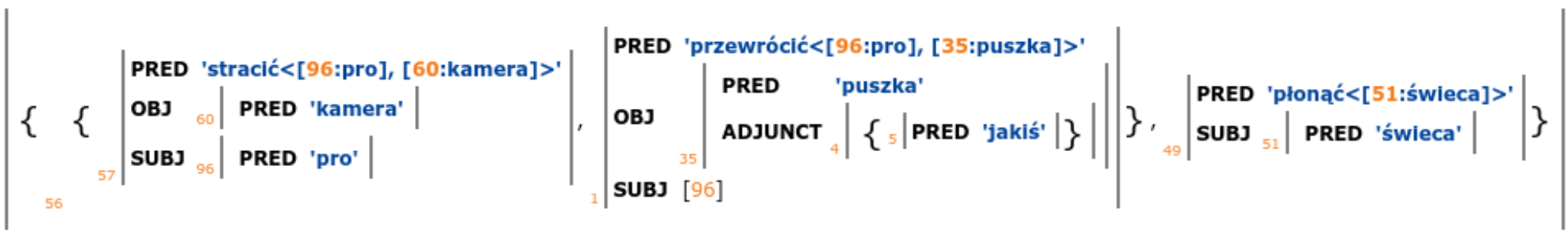
- W Laskach w liturgii uczestniczyło się przez cały dzień i
  in Laski in liturgy participated.3.SG.N RM for whole day and
  modliło się wszędzie.
  prayed.3.SG.N RM everywhere
  'In Laski, one would participate in the liturgy for the whole day and one would pray everywhere.'

- *się* in *uczestniczyło się* 'one would participate' – purely impersonal (`expl:impers`),
- *się* in *modliło się* 'one would pray' — both impersonal (`expl:impers`) and inherent (`expl:pv`) in the verb MODLIĆ SIĘ 'pray'.

# Lost in translation – coordination

Known structural problem in the UD representation of **coordination**:

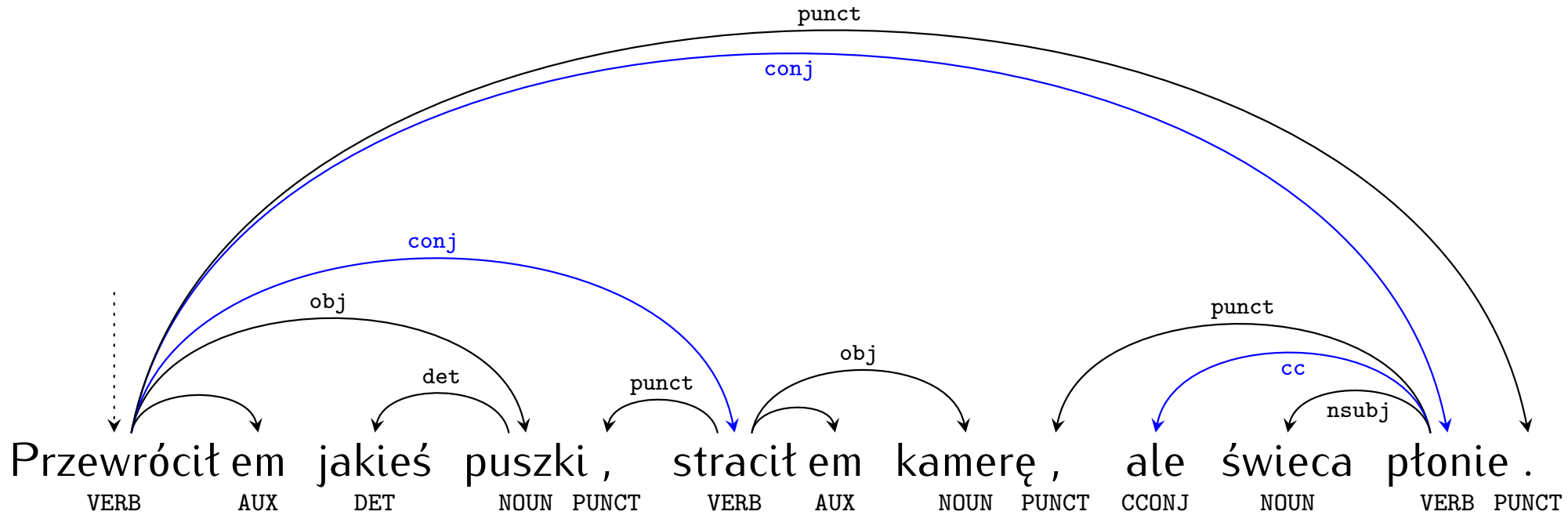
- Przewróciłem jakieś puszki, straciłem kamerę, ale
  overturned.1.SG.M some.ACC cans.ACC lost.1.SG.M camera.ACC but
  świeca płonie.
  candle.NOM.SG.F burns.3.SG
  'I overturned some cans, lost my camera, but the candle still burns.'



# Lost in translation – coordination

Known structural problem in the UD representation of **coordination**:

- Przewróciłem jakieś puszki, straciłem kamerę, ale
  overturned.1.SG.M some.ACC cans.ACC lost.1.SG.M camera.ACC but
  świeca płonie.
  candle.NOM.SG.F burns.3.SG
  'I overturned some cans, lost my camera, but the candle still burns.'



# Lost in translation – underspecified labels

Loss of structural information above. Also loss of information because of **underspecification of UD labels**:

- distinction between different kinds of oblique arguments (e.g., OBL-INST, OBL-ADL, etc.), and between **obliques and adjuncts**; UD treats all as `obl` (but subtypes of `obl` could be used to represent some distinctions; Zeman 2017),
- the different grammatical functions of **dependents of gerunds** (now all broadly nominal dependents of gerunds are marked as `nmod`, but they could be subtyped to `nmod:obj`, `nmod:obl`, etc.),
- the **distinction between controlled infinitivals and predicative complements**, both marked in UD as `xcomp` (e.g., by subtyping the latter to `xcomp:pred`),
- the **distinction between raising and control** (e.g., by representing raising via `xcomp:raising`),
- the distinction between **eventuality and constituent negation** (Przepiórkowski and Patejuk 2015), e.g., via the subtypes `advmod:eneg` and `advmod:cneg`,
- the distinction between **semantic and asemantic prepositions**, e.g., by subtyping the `case` relation in the former to `case:sem`; **etc.**

# Summary

Summary, conclusions:

- **largest UD treebank of Polish** (over 17K sentences, almost 131K tokens),
- conversion from LFG to enhanced UD **preserves much of structure-sharing**, etc.,
- main reasons for **loss of information**:
  - no representation of ***pro*-dropped dependents**,
  - **underspecification of labels**,
- statistically insignificant:
  - **ban on multiple relations** between same tokens,
  - representation of **coordination** which does not distinguish between (certain) flat and embedded structures.

Butt, M. and King, T. H., eds. (2015). *The Proceedings of the LFG'15 Conference*, Stanford, CA. CSLI Publications.

Kupść, A. (1999). Haplology of the Polish reflexive marker. In R. D. Borsley and A. Przepiórkowski, eds., *Slavic in Head-Driven Phrase Structure Grammar*, pp. 91–124. CSLI Publications, Stanford, CA.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pp. 1659–1666, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Patejuk, A. and Przepiórkowski, A. (2015). An LFG analysis of the so-called reflexive marker in Polish. In Butt and King (2015), pp. 270–288.

Przepiórkowski, A. and Patejuk, A. (2015). Two representations of negation in LFG: Evidence from Polish. In Butt and King (2015), pp. 322–336.

Zeman, D. (2017). Core arguments in Universal Dependencies. In S. Montemagni and J. Nivre, eds., *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing 2017)*, pp. 287–296, Pisa, Italy.