Agnieszka Patejuk
Adam Przepiórkowski

From Lexical Functional Grammar to Enhanced Universal Dependencies

Linguistically informed treebanks of Polish

Institute of Computer Science
Polish Academy of Sciences

Warszawa 2018

Scientific Editors at the Institute of Computer Science PAS:
Prof. dr hab. Jan Mielniczuk
Prof. dr hab. Wojciech Penczek

Reviewers:
Dr. Paul Meurer (University of Bergen)
Prof. Dr. Stephan Oepen (University of Oslo)

Authors' address:
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
{aep;adamp}@ipipan.waw.pl

ISBN 978-83-63159-26-9

First edition

Typeset in XeLaTeX by Adam Przepiórkowski

Contents

Preface

Syntactically annotated corpora, or 'treebanks', belong to the most heterogeneous kinds of linguistic resources. They differ not only in the general kind of approach they adopt (constituency or dependency), but also in the number of representation levels they assume (often one, but sometimes two or more) and in the extent to which they follow an established linguistic theory (if at all). Also, even within one kind of approach, the representation of a particular phenomenon may differ widely between treebanks (see, e.g., Popel et al. 2013 for the treatment of coordination in various dependency treebanks).

In treebank development, there is a clear tension between theoretical accuracy within a treebank and utilitarian consistency between treebanks of the same or different languages. On the one hand, utterances should be annotated with linguistically accurate and precise descriptions, and one way to achieve this is by following a specific linguistic theory, one with a well-defined terminology, good formal background and a body of carefully justified analyses of many phenomena of typologically diverse languages. An example of such a theory is Lexical Functional Grammar (LFG; Bresnan 1982; Dalrymple 2001; Bresnan et al. 2015; Dalrymple et al. 2018). However, LFG is not the only theory of this kind, and even within one theory, similar phenomena may receive very different representations, reflecting different traditions or different weights assigned to pieces of evidence supporting one or another analysis. So this theoretically-oriented approach to treebank development inevitably leads to the creation of treebanks with very diverse annotation schemes, which are often comprehensible only to a limited number of followers of a given linguistic theory.

On the other hand, especially in the context of multilingual natural language processing (NLP), treebanks should ideally follow a common annotation scheme, one that is intelligible to a much broader group of treebank consumers than professional linguists working within a given theory. Moreover, similar phenomena and constructions should receive analogous representations, even if there are subtle – from the point of view of practical applications – differences suggesting dissimilar analyses. A recent attempt at such a comprehensive syntactic annotation scheme is Universal Dependencies (UD; `http://universaldependencies.org/`, Nivre et al. 2016). As a practical solution, UD aims at providing a maximally simple syntactic representation, one that is useful for various NLP applications, even if at the cost of linguistic precision.

This monograph presents two treebanks of Polish which follow the two approaches, as well as the procedure of converting one to the other. Part I describes an LFG treebank, which – given that each utterance is annotated not only with a constituency tree but also with a non-arboreous functional structure – is called 'structure bank' below. Both structures adhere to the principles of Lexical Functional Grammar, but many aspects of the two representations

are specific to Polish and to the LFG grammar which underlies the treebank (see Chapter 1); the role of particular attributes occurring in functional structures is described and illustrated in Chapter 2, while the role of different labels of syntactic nodes in constituency structures is explained in Chapter 3.

Part II describes the procedure of converting this LFG structure bank to a UD treebank. The input to the conversion, an intermediate representation, and the output are presented in Chapter 4. The following Chapter 5 discusses some differences in tokenisation between the two resources. Further, Chapter 6 is devoted to the differences between the morphosyntactic levels of the two treebanks. In order to comply with UD guidelines, it has been necessary to infer grammatical classes (e.g., that of determiner) and syntactic categories (e.g., that of mood) which are not explicitly represented in the input LFG structure bank. Conversely, it has also been useful to add to the usual UD categories a few language-specific features in order to preserve detailed information available in the input (e.g., that of the three masculine 'sub-genders' or emphatic forms of some broadly pronominal lexemes). Finally, the longest chapter of this part, Chapter 7, presents – in excruciating detail – the two stages of the conversion of syntactic LFG structures to dependency representations assumed in UD. First, the derivation of a dependency representation which closely mirrors the input LFG structures is described in Section 7.1. Second, the consecutive transformations of this intermediate representation resulting in the final fully UD-compliant structure are discussed in Section 7.2.

Part III consists of the sole Chapter 8, which offers a stand-alone presentation of the resulting UD trebank of Polish. Apart from describing the kinds of morphosyntactic and syntactic information available in the treebank, it also characterises the underlying data and gives quantitative information about the size of the corpus and the kinds of texts it contains. As this is not the first UD treebank of Polish, this chapter also contains a comparison of this LFG-derived UD treebank to an earlier treebank of Polish, itself the result of (a few steps of) conversion from a constituency treebank. The most conspicuous difference – apart from the larger size of the LFG-derived treebank – is the fact that the treebank presented here makes extensive use of the enhanced representation scheme made available in the current version 2 of Universal Dependencies. As discussed in Chapter 9, concluding the monograph, this feature of UD makes it possible to preserve various kinds of syntactic information normally not expressible in simple dependency trees, including information about grammatical control and about sharing of dependents in coordinate structures.

While the concluding Chapter 9 presupposes some knowledge of the material of the previous chapters, the three main parts of this monograph are meant to be self-contained. This is especially true about Parts I and III, which present the two resources in a way that does not assume the knowledge of the other resource or of the conversion procedure. An attempt was also made to present the conversion procedure in Part II independently of the presentation of the two resources, although prior exposition to LFG and UD will certainly make reading this part easier.

Both the creation of the original LFG corpus and the conversion into UD have been partially supported by the Polish Ministry of Science and Higher Education within the CLARIN ERIC programme 2016–2018 (`http://clarin.eu/`). The original LFG structure bank has been developed under the supervision of Agnieszka Patejuk and has been converted to UD by Adam

Przepiórkowski, in collaboration with Agnieszka Patejuk. We would like to cordially thank Joakim Nivre and Dan Zeman for their infinite patience in answering a myriad of diverse UD-related questions during the development of this treebank, and the reviewers of this monograph, Paul Meurer and Stephan Oepen, for their comments, which led to some important improvements. The data, lemmata and original morphosyntactic tags come from the manually annotated subcorpus of the National Corpus of Polish (`http://nkjp.pl/`), whose development – within a project led by Adam Przepiórkowski – was financed by the Polish Ministry of Science and Higher Education in 2007–2011, and – to a lesser extent – from the Corpus of 1960s Polish (`http://clip.ipipan.waw.pl/PL196x`). Parts of this monograph were written and revised during our fellowship at the Oslo Center for Advanced Study (CAS) at the Norwegian Academy of Science and Letters (`https://cas.oslo.no/`), within the group "SynSem: From Form to Meaning – Integrating Linguistics and Computing" led by Dag Haug and Stephan Oepen. It is very possible that, if not for our involvement in CAS, neither the UD treebank of Polish presented here, nor this monograph, would ever see the light of day.

Part I

LFG Structure Bank of Polish

Chapter 1

Polish in LFG: grammar and structure bank

1.1 Lexical Functional Grammar

Lexical Functional Grammar (LFG; Bresnan 1982; Dalrymple 2001; Bresnan et al. 2015; Dalrymple et al. 2018) is a linguistic theory which assumes two syntactic levels of representation (in addition to other, non-syntactic levels): constituency structure (c-structure) and functional structure (f-structure). In the case of the Polish sentence (1.1),[1] from the multilingual LFG test-suite ParGramBank (Parallel Grammar Treebank; Sulger et al. 2013), the c-structure is given in Figure 1.1 and the f-structure – in Figure 1.2.

(1.1) Kierowca zapala traktor.
driver.NOM.SG.M ignites.3SG tractor.ACC.SG.M
'The driver starts the traktor.'

LFG constituency and functional structures shown in this monograph are visualisations of such structures produced by the INESS system (Infrastructure for the Exploration of Syntax and Semantics; `http://clarino.uib.no/iness/`; Rosén et al. 2007, 2012), which hosts ParGramBank and the LFG structure bank of Polish described in subsequent chapters, among other treebanks.

According to the c-structure in Figure 1.1, the whole utterance (1.1) consists of a finite sentence (`S/IP`; such labels are explained in Chapter 3) and the final period. The sentence in turn consists of a nominal phrase (`ARG/NP/N/SUBST`) and a verbal phrase (`Ibar`) containing the finite

[1] Abbreviations of grammatical properties largely follow the Leipzig Glossing Rules (`https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf`; version marked as 'Leipzig, last change: May 31, 2015'), with the following exceptions: 1) internal morphological structure of particular tokens is not indicated explicitly, 2) only some of the morphosyntactic information is given explicitly, e.g., *zapala* 'ignites, starts' in (1.1) is only marked as third person singular, but not as occurring in present tense or active voice, as this information is not immediately relevant in this case, 3) sometimes English forms in word-by-word glosses indicate relevant morphosyntactic information, e.g., *zapala* is glossed as 'ignites', indicating present tense. While there are three masculine genders in Polish, they are all glossed as M here. Additionally, morphologically impersonal forms of verbs (so-called *-no/-to* forms) are glossed as IMPS, and gerundial forms (so-called *-nie/-cie* forms) – as GER. Finally, the so-called reflexive marker *się* is not glossed as REFL, as it is rarely truly reflexive; instead, it is marked as RM, or with a specific role it plays in the sentence: INH (inherent, part of the verb) or IMPS (impersonal).

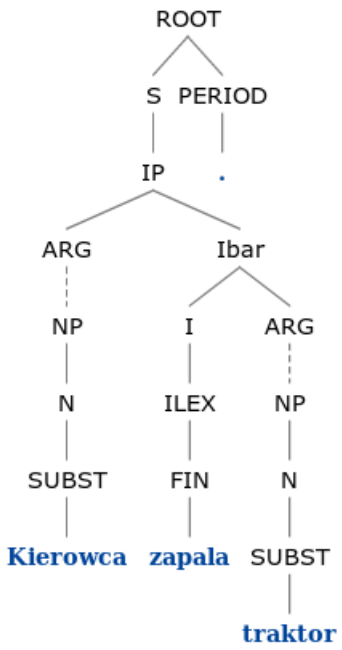


Figure 1.1: C-structure of (1.1)

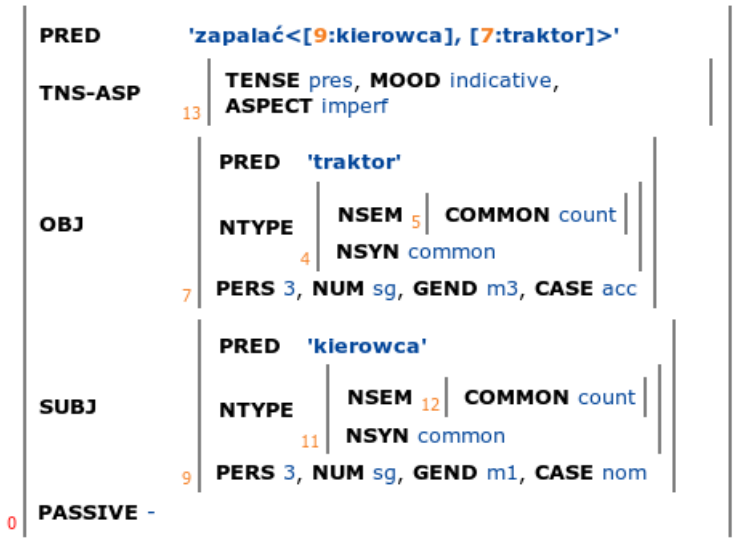


Figure 1.2: F-structure of (1.1)

verb (I/ILEX/FIN) and another nominal phrase (ARG/NP/N/SUBST). According to the f-structure in Figure 1.2, the main predicate (PRED; such attributes are explained in Chapter 2) of the whole utterance is ZAPALAĆ 'ignite, start', and it has two arguments: an object (OBJ) and a subject (SUBJ). The utterance is in the present tense, indicative mood and imperfective aspect (see the value of TNS-ASP), and in the active voice (cf. the negative value of PASSIVE). The object introduces the predicate TRAKTOR 'tractor', and it is described as a countable common noun in the accusative case, with the singular number, third person and 'masculine inanimate' gender (marked as M3). Similarly, the subject is a countable common noun in the nominative case, it has the singular number, third person and 'masculine human' gender (marked as M1), and it introduces the predicate KIEROWCA 'driver'.

Various levels of representation are related via structural correspondences (Kaplan 1995). In the case of c-structure and f-structure, there is a function, often called ϕ, from nonterminal nodes in c-structure to particular substructures in f-structure. For example, in the case of Figures 1.1–1.2, the leftmost nodes `NP`, `N` and `SUBST` in Figure 1.1 all map to the substructure with index 9 (i.e., the value of SUBJ) in Figure 1.2, the rightmost nodes `NP`, `N` and `SUBST` all map to the substructure with index 7 (i.e., the value of OBJ), and all the other nonterminals, including `ROOT`, `S`, `IP`, `Ibar` and `FIN` – to the whole f-structure with index 0.[2] In order to avoid clutter, such correspondences will not be explicitly shown in figures below, but they will be pointed out in the text, where needed.

1.2 Polish LFG grammar

While LFG is a linguistic theory, it is sufficiently formal to provide a basis for computer implementations of parsers, i.e., programs which automatically construct syntactic analyses of natural language sentences. More precisely, there exists a computational platform – XLE (Xerox Linguistic Environment; Crouch et al. 2011) – which reads an appropriately encoded LFG grammar of a given language and finds syntactic analyses of utterances of this language which comply with that grammar.

Such an implemented LFG grammar of Polish, POLFIE (`http://zil.ipipan.waw.pl/LFG/`), was developed in early 2010s and has since then been expanded in various ways. As described in more detail in Patejuk and Przepiórkowski 2012, grammatical rules used in early versions of POLFIE were written on the basis of two previous formal grammars of Polish: 1) the DCG (Definite Clause Grammar; Warren and Pereira 1980) grammar GFJP2 (based on the earlier GFJP grammar; Świdziński 1992) used by the parser Świgra (Woliński 2004) and 2) the HPSG (Head-driven Phrase Structure Grammar; Pollard and Sag 1994) grammar described in Przepiórkowski et al. 2002. While the former provided the basis for constituent structure rules, the latter was used as the basis for building f-structures. This initial LFG grammar of Polish has been extended with new analyses of various syntactic phenomena, including coordination, agreement, case assignment and negation; many of the implemented solutions are described and theoretically justified in various publications by the current authors in proceedings of consecutive editions of the international LFG conference since 2012 (`http://cslipublications.stanford.edu/LFG`).

Also the lexicon of POLFIE is heavily based on other resources. Morphosyntactic information is drawn from a morphological analyser of Polish, Morfeusz (Woliński 2006, 2014), or from corpora: the National Corpus of Polish (NKJP; `http://nkjp.pl/`; Przepiórkowski et al. 2011, 2012), the Corpus of 1960s Polish (`http://clip.ipipan.waw.pl/PL196x`; Kurcz et al. 1990; Bień and Woliński 2003; Ogrodniczuk 2003), or Składnica, a treebank of parses produced by the Świgra constituency parser (Świdziński and Woliński 2010; Woliński et al. 2011).[3] While some syntactic information is added manually to selected lexical entries – e.g., those of *wh*-words (such

[2] Apart from the whole f-structure, whose index in INESS visualisations is always 0, particular substructures have indices assigned in an arbitrary fashion.

[3] The annotation of LFG structure bank is independent of syntactic analyses found in Składnica – only morphosyntactic information is used (orthographic form, lemma, tag).

as *kto* 'who' or *dlaczego* 'why'), *n*-words (such as *nikt* 'nobody', *nigdy* 'never' or *żaden* 'none'), etc. – valency information is automatically converted from a large valency dictionary of Polish, Walenty (`http://walenty.ipipan.waw.pl/`; Przepiórkowski et al. 2014, 2017; Hajnicz et al. 2016); the conversion procedure is described in detail in Patejuk 2015: ch.8.

POLFIE is one of the largest implemented LFG grammars. The number of grammatical rules – 118 – is deceptively small, as XLE allows for only one rule defining any given non-terminal. This means that a typical XLE rule contains multiple right-hand side disjunctions and corresponds to many context-free rules. Perhaps a more telling measure is the number of lines of code. The pure grammar, without the valency dictionary, contains 19,878 lines of code. The dictionary itself has 2,142,129 lines of code, so the total number of lines is 2,162,007. The grammar is available at `http://zil.ipipan.waw.pl/LFG`.

1.3 Polish LFG structure bank

The main source of texts in the Polish LFG structure bank described in detail in the following two chapters is the National Corpus of Polish, the secondary source is the Corpus of 1960s Polish. Both corpora are manually annotated with morphosyntactic tags compliant with the tagset of the National Corpus of Polish, briefly described in Appendix A. These manually introduced tags are to a very large extent preserved in the LFG structure bank of Polish and they are reflected both in the names of c-structure preterminals and in various morphosyntactic attributes present in f-structures. In very rare cases, some of the original morphosyntactic information has been automatically modified to reflect LFG analyses of some phenomena. For example, the case of typical numeral subjects has been converted from nominative to accusative, in accordance with the arguments of Franks 1995 and Przepiórkowski 1999, 2004a (among others) and following the LFG analysis of Przepiórkowski and Patejuk 2012a, 2012b and Patejuk and Przepiórkowski 2014b.

Syntactic annotations in the LFG structure bank have been created semi-automatically. First, the sentences were parsed using the POLFIE grammar and the XLE system mentioned in the previous section. In effect, often multiple analyses were produced for many sentences, since any grammar of a reasonable size must be ambiguous. After this automatic process, analyses were manually disambiguated by a group of trained linguists – to ensure the high quality of the resulting structure bank, each sentence was disambiguated independently by two annotators,[4] whose analyses were subsequently inspected by the superannotator (for every single sentence), who could agree with the annotators or choose a different solution. During annotation, the annotators were not allowed to individually communicate or to see each other's comments. On the other hand, they could communicate via a mailing list accessible to all of them, to the superannotator and to the developers of the grammar. The process was supervised by the chief grammar writer, who responded to questions, and by the superannotator, who replied to annotators' numerous comments.

[4] As in the case of the manual annotation of NKJP (Przepiórkowski and Murzynowski 2011), pairs of annotators were not constant; instead annotators were shuffled so as to avoid co-learning the same mistakes.

Relatively high speed of annotation could be attained thanks to the use of the INESS infrastructure – mentioned in Section 1.1 – for building structure banks. Figure 1.3 presents a screenshot of the system for sentence (1.2).

(1.2) Jak wygląda przepiórka?
how looks.3SG quail.NOM.SG.F
'What does a quail look like?'
'How does a quail look out?'

This sentence is syntactically and semantically ambiguous: *wygląda* is a form of an ambiguous lexeme WYGLĄDAĆ, whose meanings include the bivalent 'look like' and the possibly monovalent 'look out' (as in looking out of a window, etc.). In both cases *przepiórka* 'quail' is the subject of this verb, but the initial question word *jak* 'how' is interpreted either as the second argument, in the case of the 'look like' meaning, or as a manner adjunct, in the case of the 'look out' meaning.

Both the c-structure and the f-structure are shown in a compact format encompassing a number of analyses (here, two) at the same time. For example, in the c-structure in the middle of the screenshot, the choice is at the level of the highest `IP` node: should it be rewritten to `ADVP IP` (the analysis marked as `[a2]`) or to `IP XPsem` (analysis `[a1]`, with the order of nodes reversed, as the lower `IP` is shared between these two analyses)? The correct parse may be selected by the annotator by clicking on one of the two rules in the bottom left corner of the screenshot: `IP –> XPsem IP` or `IP –> ADVP IP`.

This choice at the level of c-structure is correlated with a choice at the level of f-structure. For example, the f-structure will contain the feature `ADJUNCT` only if `a2` is selected. Otherwise, if `a1` is chosen, it will contain the feature `OBL-MOD`. So, instead of relying on c-structure discriminants in the table at the bottom left corner of this figure, annotators may rely on f-structure discriminants in the table above it, and select either the third row of the table, mentioning `OBL-MOD 'jak'`, or the fifth row, mentioning `ADJUNCT $ 'jak'`. In fact, the choice boils down to whether the verb WYGLĄDAĆ 'look like' is a two-argument verb (see the first row in this table) or a one-argument verb (see the second row). As the first of these options seems correct, the annotator may disambiguate this sentence by clicking on the first row or – equivalently – on the third row. The result of choosing the latter discriminant is shown in Figure 1.4.

The fully disambiguated part of the Polish LFG structure bank contains 21,732 utterances and is searchable via the INESS infrastructure at `http://clarino.uib.no/iness/`, where it is called `pol-lfg`.

Figure 1.3: Annotation of (1.2) before disambiguation

Figure 1.4: Annotation of (1.2) after disambiguation

Chapter 2

F-structure

Of the two kinds of syntactic structures assumed in LFG, functional structures are in various ways more important than constituency structures. One reason is that functional structures are more universal: languages with widely different constituency syntax – e.g., highly configurational languages such as English and highly non-configurational languages such as Warlpiri – may have very different c-structures of translation equivalents, while having rather similar f-structures. Another – related – reason is that f-structures are 'closer to semantics', in the sense that semantic representations of sentences may be constructed to a large extent on the basis of information contained directly in f-structures (cf., e.g., Andrews 2007 and references therein).

Consider again the simple sentence (1.1), repeated below, and its f-structure in Figure 1.2, repeated below as Figure 2.1.

(1.1) Kierowca zapala traktor.
driver.NOM.SG.M ignites.3SG tractor.ACC.SG.M
'The driver starts the traktor.'

One of the attributes in such f-structures, PRED, is directly related to semantics: its values are so-called semantic forms (Dalrymple 2001: 219–221), i.e., predicates introduced by particular content words together with their argument structures. For example, the value of the top-level PRED in Figure 2.1 says that the main predicate of the sentence is ZAPALAĆ 'ignite, start' and that this predicate takes two arguments: one represented by the substructure with index 9, i.e., by the value of the SUBJ attribute, and another represented by substructure 7, i.e., the value of OBJ. The values of PRED within these substructures, i.e., KIEROWCA 'driver' and TRAKTOR 'tractor', are predicates with empty argument structures.

Apart from PRED, other attributes can be roughly split into two classes. The first contains attributes such as TNS-ASP, MOOD, PASSIVE, NTYPE, PERS, CASE, etc., i.e., attributes representing mainly morphosyntactic information. Such attributes are briefly characterised in Section 2.1. The second class consists of attributes representing relations between parts of the sentence, especially, grammatical functions such as subject (SUBJ) or object (OBJ). Attributes belonging to this class are described in more detail in Sections 2.2–2.17.

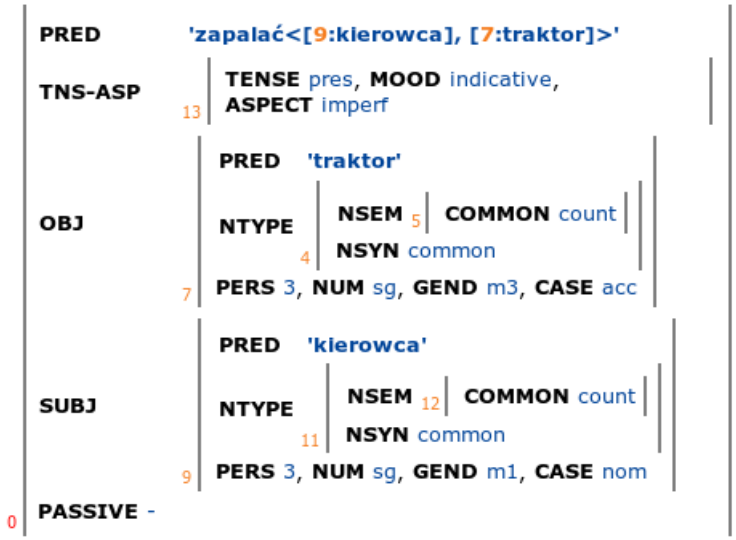


Figure 2.1: F-structure of (1.1)

2.1 Morphosyntactic attributes

As mentioned in Section 1.3, utterances in the LFG structure bank are manually annotated with morphosyntactic tags compliant with the tagset of the National Corpus of Polish (see Appendix A), and many f-structure attributes directly reflect these legacy tags. This is true of the following attributes, *inter alia*:

- ASPECT (grammatical aspect):
  - IMPERF: imperfective
  - PERF: perfective
- CASE (grammatical case):
  - ACC: accusative
  - DAT: dative
  - GEN: genitive
  - INST: instrumental
  - LOC: locative
  - NOM: nominative
  - VOC: vocative
- DEGREE (grammatical degree – analytic or synthetic):
  - COMPARATIVE
  - POSITIVE
  - SUPERLATIVE
- GEND (grammatical gender):
  - F: feminine
  - M1: 'human' masculine
  - M2: 'animate' masculine
  - M3: 'inanimate' masculine
  - N: neuter

- NUM (grammatical number):
  - SG: singular
  - PL: plural
- PERS (grammatical person):
  - 1: first
  - 2: second
  - 3: third

The only attribute that requires a comment is GEND, with its five values, including three masculine genders, following Mańczak 1956. Despite the descriptive names of these masculine genders, suggesting that they differ in the semantic feature of animacy, they can be distinguished purely formally (on the basis of agreement facts) and their correlation with semantic animacy is far from perfect.[1]

The above attributes are standard in the sense that LFG grammars for many different languages are expected to have them (even if, as in the case of GEND, the repertoire of possible values varies from language to language). Other morphosyntactic attributes corresponding to the legacy tagset are more parochial, specific to Polish. Such attributes are 'hidden' within the values of a special attribute, CHECK, which is normally suppressed in INESS visualisations. For example, in the case of sentence (2.1), the complete f-structure, with CHECK values shown, is given in Figure 2.2 (compare with Figure 2.10 on page 24, where CHECK is suppressed).

(2.1) Bezprawnie ją aresztowano!
unlawfully she.ACC.SG.F arrested.IMPS
'She was arrested unlawfully!'

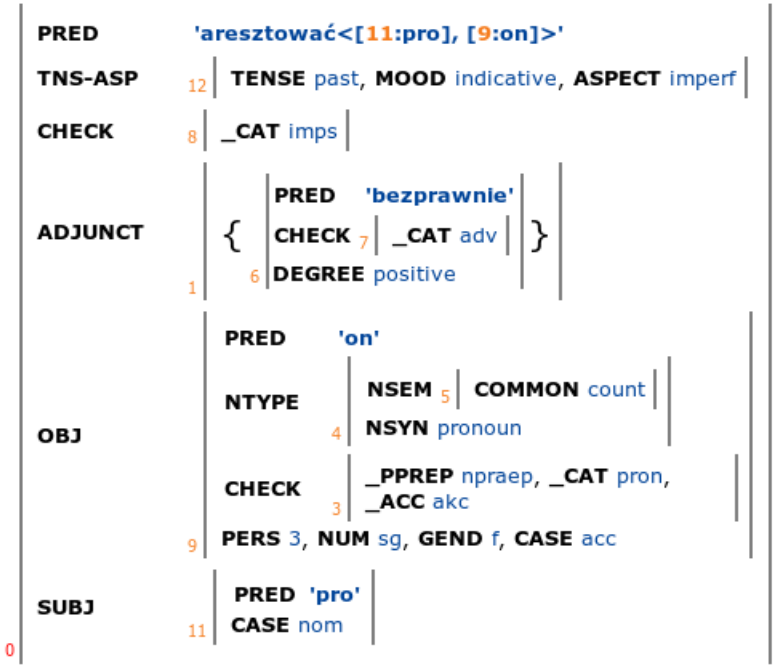


Figure 2.2: F-structure of (2.1) with CHECK values displayed

[1]Note that, in the examples given in the text, such as (1.1), we simplify morphosyntactic information and mark all masculine forms as M (masculine), rather than as M1, M2 or M3.

Here, the f-structure with index 9, representing the pronoun *ją* 'her' (a form of the lexeme ON 'he'), contains the CHECK attribute with a value introducing three parochial attributes mirroring those of the legacy tagset: _CAT, _PPREP and _ACC. Altogether, five such 'legacy attributes' are relegated to CHECK:

- _CAT: fine-grained syntactic class (part of speech) of the main head – its values are not formally restricted and they include the following values occurring in Figure 2.2:
  - IMPS: impersonal (so-called *-no*/*-to*) form of a verb
  - ADV: adverb
  - PRON: personal pronoun
- _ACC (accentability: can the pronoun be stressed?):
  - AKC: accentable (strong)
  - NAKC: non-accentable (weak)
- _PPREP (post-prepositionality: does the pronoun only occur as a dependent of a preposition?):
  - PRAEP: post-prepositional
  - NPRAEP: non-post-prepositional
- _AGL (agglutination: does this verbal form only occur adjacent to a mobile inflection?):
  - AGL: agglutinative (only adjacent to a mobile inflection)
  - NAGL: non-agglutinative (only non-adjacent to a mobile inflection, if any)
- _VOC (vocalicity: does this form of a preposition or a mobile inflection differ from another form of the same preposition or mobile inflection only in an additional vowel?):
  - WOK: vocalic
  - NWOK: non-vocalic

Another attribute corresponding directly to the legacy tagset is ACM:

- ACM (accommodability: does the numeral agree in case with its nominal dependent or does it assign the genitive case?):
  - CONGR: agreeing
  - REC: governing

While this attribute is as specific to Polish morphosyntax as the above CHECK attributes, it appears (e.g., in Figures 2.4–2.5 below) outside of CHECK, as it is important for the syntactic analysis of numeral phrases.

Other attributes in this class do not directly correspond to morphosyntactic categories defined in the legacy tagset. Two of them, MOOD and TENSE, are grouped together with ASPECT within the values of TNS-ASP:

- MOOD (grammatical mood):
  - IMPERATIVE
  - INDICATIVE
  - CONDITIONAL
- TENSE (grammatical tense):
  - FUT: future
  - PAST: past
  - PRES: present

- TNS-ASP: groups MOOD, TENSE and ASPECT

A related attribute marks passive participles:

- PASSIVE:
  - +: passive participle form
  - −: not passive participle form

A group of attributes subclassifies various types of constituents:

- CLAUSE-TYPE (clause type, applies to embedded clauses):
  - DECL: declarative
  - INT: interrogative
  - IMP: imperative
  - REL: relative
  - OR: oratio recta
- PTYPE (preposition type):
  - SEM: semantic
  - NOSEM: non-semantic
- ATYPE (adjective type):
  - ATTRIBUTIVE
  - PREDICATIVE
- NTYPE (noun type): it contains the attributes NSYN and NSEM
- NSYN (syntactic noun type):
  - COMMON
  - PRONOUN
  - PROPER
- NSEM (semantic noun type): it contains the attribute COMMON
- COMMON (common nouns):
  - COUNT: countable
  - GERUND: gerund
- TYPE (pronoun type):
  - INT: interrogative
  - REL: relative
  - NEG: n-word
  - ANY: *-kolwiek* '-ever' type
  - ALL: universal
  - RES: resumptive

Additionally, four attributes mark the presence of specific kinds of constituents:

- COMITATIVE (comitative coordination):
  - +: yes
- CORRELATIVE (correlative pronoun):
  - +: yes
- PARTITIVE (partitive dependent):
  - +: yes
- _PREDICATIVE (predicative dependent of any category; this attribute occurs within CHECK):
  - +: yes

Two attributes record the presence of eventuality and constituent negation (Przepiórkowski and Patejuk 2015):

- NEG (eventuality negation):
  - +: yes (present)
- CNEG (constituent negation):
  - +: yes (present)

Some attributes record the particular function lexeme (or an equivalence class of mutually substitutable function lexemes, as in the case of COMP-FORM and (PRE)COORD-FORM):

- COMP-FORM: complementiser form (not restricted)
- PFORM: non-semantic preposition form (not restricted) – a preposition which does not introduce a temporal, locative, etc., semantic relation, i.e., which acts as a 'case marker'
- COORD-FORM: conjunction form (not restricted)
- PRECOORD-FORM: preconjunction form (not restricted)

There is also a group of attributes marking the presence and the function(s) of the so-called reflexive marker SIĘ. As the analysis of this small but fascinating word changed at one point in the underlying LFG grammar (Patejuk and Przepiórkowski 2015a), two different representations of SIĘ may be found in the structure bank. According to the original analysis, every SIĘ is either indicated with the IMPERSONAL attribute, when it marks an impersonal construction (as in Figure 2.11 on page 25), or with the misleadingly named REFLEXIVE attribute, in all other cases:

- IMPERSONAL (SIĘ marks an impersonal construction):
  - +: yes
- REFLEXIVE (a non-impersonal SIĘ is present):
  - +: yes

On that analysis, the reflexive and reciprocal uses of SIĘ may be distinguished from inherent uses, where SIĘ is a meaningless part of the verb, only on the basis of the PRED value: in the case of an inherent use, the predicate ends in _SIĘ. For example, in Figure 2.7 on page 21, the '+'-valued REFLEXIVE occurs at the same f-structure as PRED with the predicate name WYDAWAĆ_SIĘ, so SIĘ does not have a reflexive (or reciprocal) meaning there.

The newer representation of the various functions of SIĘ is more explicit:

- SIE (topmost SIĘ-related attribute); it may contain the following attributes:
- INH (inherent SIĘ):
  - +: yes
- REFL (reflexive SIĘ):
  - +: yes
- RECIP (reciprocal SIĘ):
  - +: yes
- IMP (impersonal SIĘ):
  - +: yes
- PRESENT (SIĘ is present locally):
  - +: yes

For example, two uses of impersonal SIĘ are marked as such with the use of the IMP attribute in Figure 2.12 on page 26, while an occurrence of SIĘ is marked as inherent with the use of the INH attribute in (a substructure in) Figure 2.24 on page 38.

2.2 Grammatical functions

The ensuing sections describe the following repertoire of grammatical functions in the Polish LFG structure bank:

- SUBJ: subject
- OBJ: direct object
- OBJ-TH: indirect object (in the dative case)
- OBL-<CASE>: other complements (i.e., non-subject arguments) marked for various cases:
  - OBL-STR: structural case
  - OBL-GEN: genitive case (lexical)
  - OBL-INST: instrumental case
- OBL, OBL2: non-semantic prepositional phrase
- OBL-AG: prepositional phrase expressing the oblique agent with passive participles
- OBL-<SEM>: various semantically defined complements (regardless of category):
  - OBL-COMPAR: prepositional phrase expressing a comparison
  - OBL-ABL: ablative phrase of any category
  - OBL-ADL: adlative phrase of any category
  - OBL-PERL: perlative phrase of any category
  - OBL-LOCAT: locative phrase of any category
  - OBL-MOD: manner phrase of any category
  - OBL-TEMP: temporal phrase of any category
  - OBL-DUR: durative phrase of any category
- OBL-ADV: adverbial oblique
- COMP: closed clausal complement (headed by a verbal predicate)
- XCOMP: open (controlled) clausal complement (headed by a verbal predicate)
- XCOMP-PRED: open (controlled) predicative complement, regardless of category (NP, AP, PP, PAP)
- ADJUNCT: closed adjunct
- XADJUNCT: open (controlled) adjunct (secondary predicate, adverbial participle)
- POSS: possessive dependent (genitive)
- APP: appositive dependent

2.3 Subject (SUBJ)

The following subsections describe possible values of the SUBJ attribute, i.e., possible subjects in the LFG structure bank: broadly nominal (Section 2.3.1), broadly verbal (Section 2.3.2), covert (implicit; Section 2.3.3), as well as subjects shared by coordinated predicates (Section 2.3.4).

2.3.1 Nominal

While typically Polish subjects are nominative and agree with the verb, this is not always the case: there are also non-agreeing numeral subjects and genitive subjects of gerunds.

Nominative

The verb *milczał* 'kept silent' in (2.2) takes a nominative subject, *chłopak* 'boy, lad', which agrees with the verb in number (SG in the glosses) and gender (M in the glosses). The f-structure in Figure 2.3 shows that the predicate MILCZEĆ 'keep silent' in the main f-structure (with index 0), contains a SUBJ attribute, whose value is the substructure with index 2. The main predicate of this substructure is CHŁOPAK 'boy, lad', and the value of CASE is NOM.

(2.2) Chłopak milczał.
boy.NOM.SG.M kept silent.3SG.M
'The boy kept silent.'

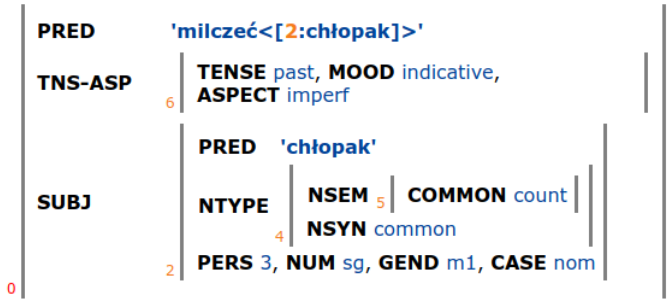


Figure 2.3: F-structure of (2.2)

Numeral: agreeing vs. non-agreeing

The verb *zginęły* 'were killed, died' in (2.3) takes a nominative numeral subject, *cztery osoby* 'four people', which agrees with the verb, as shown in glosses (PL.F). The f-structure in Figure 2.4 shows that the predicate ZGINĄĆ 'be killed, die', 0, contains a SUBJ attribute, 2, filled by the predicate CZTERY 'four', whose value of CASE is NOM. Also, its value of ACM is CONGR, marking that it is an agreeing numeral form: CZTERY, 2, takes the predicate OSOBA 'person', 8, as the value of its OBJ attribute, and both have the same value of CASE (NOM).

(2.3) Cztery osoby zginęły.
four.NOM.PL.F person.NOM.PL.F died.3PL.F
'Four people died.'

By contrast, the verb *głosowało* 'voted' in (2.4) takes an accusative numeral subject (Franks 1995; Przepiórkowski 1999, 2004a), *390*, which does not agree with the verb – while the subject is plural and 'human' masculine, the verb displays 'default agreement' (Dziwirek 1990), as shown in glosses (3SG.N). The f-structure in Figure 2.5 shows that the predicate GŁOSOWAĆ

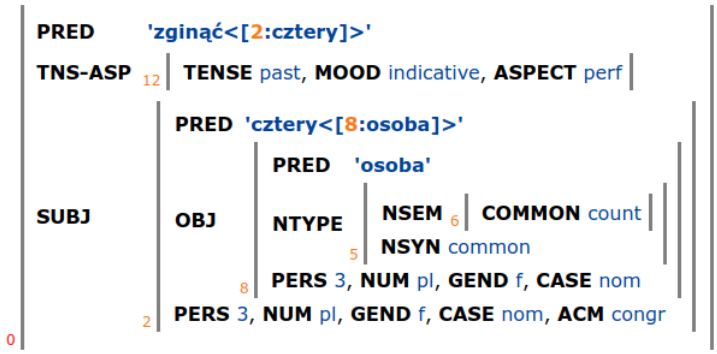


Figure 2.4: F-structure of (2.3)

'vote' in 0 contains a SUBJ attribute, 17, filled by the predicate 390, whose value of CASE is ACC and whose value of ACM is REC, marking that it is a non-agreeing numeral form: the accusative *390*, 17, takes as the value of its OBJ attribute the genitive form of the predicate POSEŁ 'member of parliament', 8.

(2.4) Głosowało 390 posłów.
voted.3SG.N 390.ACC.PL.M MPs.GEN.PL.M
'390 MPs voted.'

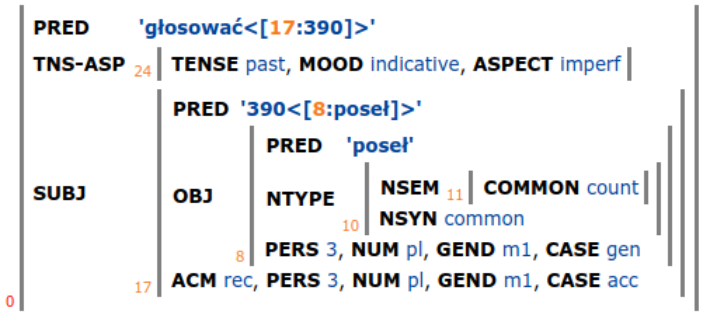


Figure 2.5: F-structure of (2.4)

Genitive subject of a gerund

The gerund *stuknięcia* 'knocks' in (2.5) takes a genitive subject, *młotka* 'hammer'. The f-structure in Figure 2.6 shows that the predicate STUKNĄĆ 'knock', 25, contains a SUBJ attribute, 43, filled by the predicate MŁOTEK 'hammer', whose value of CASE is GEN.

(2.5) Lekkie stuknięcia młotka przyniosły mu
gentle.NOM.PL.N knock.GER.NOM.PL.N hammer.GEN.SG.M brought.3PL.N he.DAT.SG.M
spokój.
peace.ACC.SG.M
'Gentle knocks of the hammer brought him peace.'

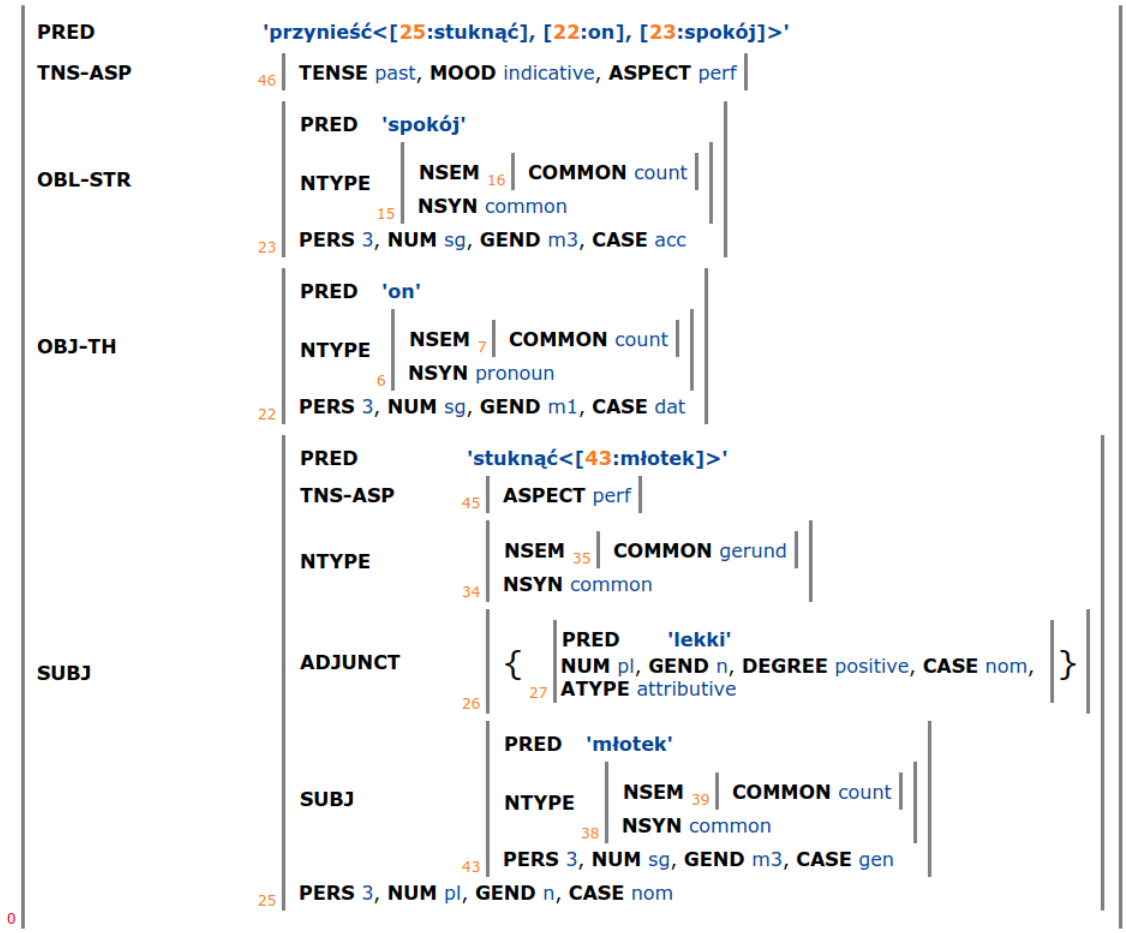


Figure 2.6: F-structure of (2.5)

2.3.2 Verbal

Though typically the subject is nominal in Polish, there are also instances of verbal subjects: these may be clausal or infinitival.

Clausal

The verb *wydaje się* 'seems' in (2.6) takes a clausal subject, *że praca polega tylko na kopaniu rowów* 'that work is only about digging ditches', which does not agree with the verb – instead, the verb *wydaje się* displays 'default agreement', as shown in glosses (3SG).[2] The f-structure in Figure 2.7 shows that the predicate WYDAWAĆ_SIĘ 'seem' (WYDAWAĆ requires inherent SIĘ, so it is included in the verb's lemma), 0, contains a SUBJ attribute filled by the predicate POLEGAĆ 'be about, consist in', 14. This last substructure has the COMP-FORM attribute with value ŻE, which indicates the type of subordinating conjunction used in the clause (the type ŻE corresponds to lexemes ŻE 'that', as in (2.6), and the equivalent but rarer IŻ).

[2] As the verb is in the present tense, it does not overtly mark gender.

(2.6) Tobie się wydaje, że praca polega tylko na kopaniu rowów?
you.DAT.SG INH seems.3SG that work consists.3SG only at digging ditches
'Do you think that work is only about digging ditches?'

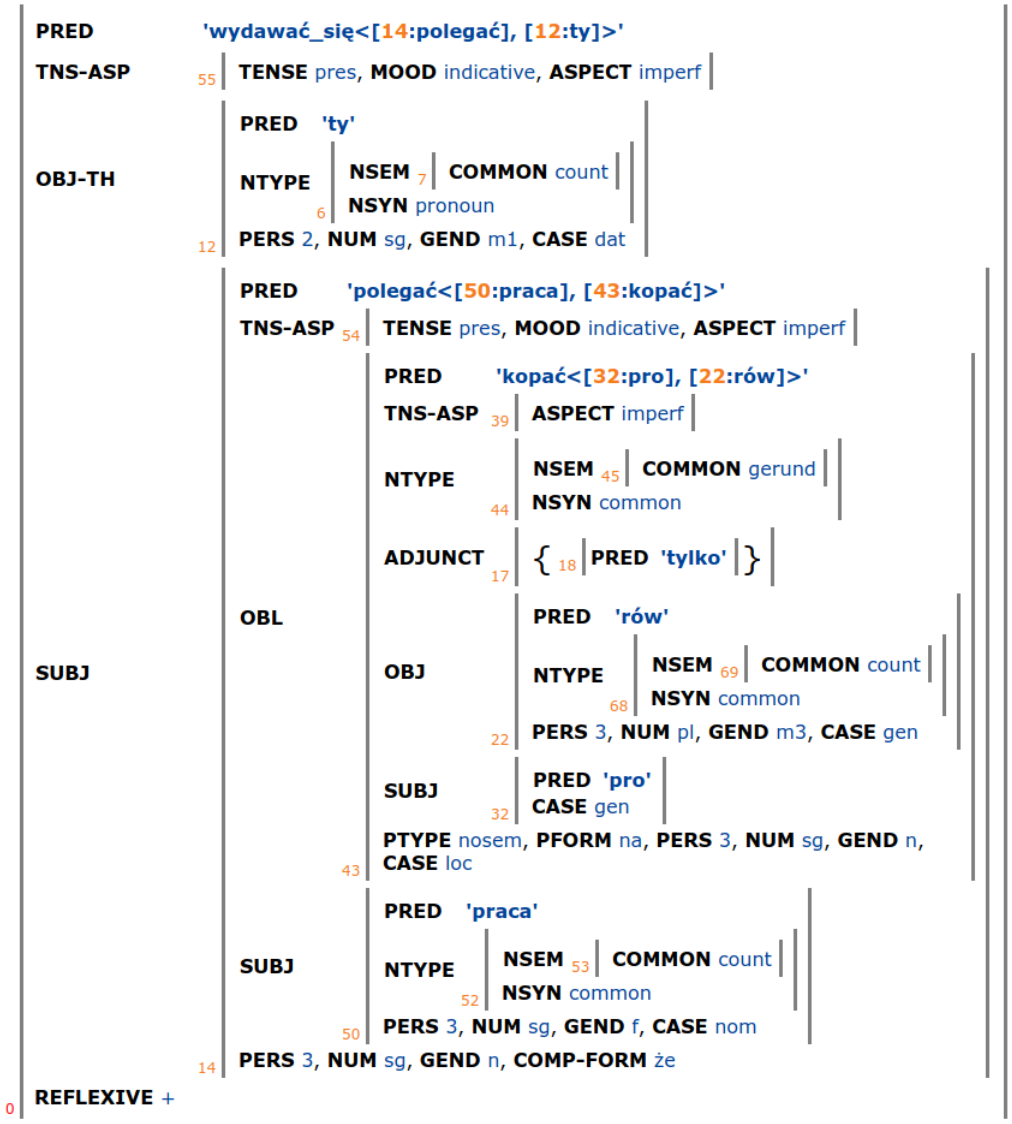


Figure 2.7: F-structure of (2.6)

Infinitival

The verb *wystarczy* 'suffice' in (2.7) takes an infinitival subject consisting of two coordinated phrases, *otworzyć kopertę i spisać dane* 'open the envelope and write down the data', which does not agree with the verb – instead, the verb *wystarczy* displays 'default agreement', as shown in glosses (3SG). The f-structure in Figure 2.8 shows that the predicate WYSTARCZYĆ 'suffice', 0, contains a SUBJ attribute, 41, filled by a set (enclosed in curly brackets) containing two elements: the predicate OTWORZYĆ 'open', 42, and the predicate SPISAĆ 'write down, record', 1.

(2.7) Wystarczy otworzyć kopertę i spisać dane.
suffices.3SG open.INF envelope.ACC and record.INF data.ACC
'It is enough to open the envelope and write down the data.'

2.3.3 Implicit subject (*pro*)

While the subject may be realised lexically, as discussed so far, it may also be implicit: though it does not appear on the surface (there is no corresponding branch in c-structure), it is included in f-structure representation – such implicit subjects fill the SUBJ attribute with PRO. Implicit subjects are used in three environments: plain *pro*-drop (subjects may typically be dropped in Polish), with morphological impersonals (*-no*/*-to* forms), where the lexical subject must not be used, and with constructional impersonals involving SIĘ, where lexical subjects also cannot appear.

Plain *pro*-drop

The verb *chrapie* 'snores' in (2.8) takes an implicit subject. Since there is no overt subject, information about the agreement features of the implicit subject can only be inferred on the basis of the verb form used: 3SG, as shown in glosses. The f-structure in Figure 2.9 shows that the predicate CHRAPAĆ 'snore', 0, contains a SUBJ attribute, 2, filled by the predicate PRO, which marks the use of an implicit argument, whose value of PERS is 3, NUM is SG (as in glosses) and CASE is NOM (because implicit subjects are assumed to be nominative). If the verb were in the past form, the implicit subject would also be specified for gender. However, in (2.8) gender is unspecified (as indicated in glosses and in the free translation).

(2.8) - Chrapie.
snores.3SG
'– He/she/it snores.'

Morphological impersonal

Polish has a class of morphological impersonals ending in *-no*/*-to* – since they must not have a lexical subject, they use an implicit subject. For example, the verb *aresztowano* 'arrested' in (2.1) (repeated below) is a morphological impersonal form, as shown in glosses. The f-structure in Figure 2.10 shows that the predicate ARESZTOWAĆ 'arrest', 0, contains a SUBJ attribute, 16, filled by the predicate PRO, which marks the use of an implicit argument. Perhaps somewhat controversially, the value of CASE is NOM – subjects of morphological impersonals are assumed in the LFG structure bank to be nominative.

(2.1) Bezprawnie ją aresztowano!
unlawfully she.ACC.SG.F arrested.IMPS
'She was arrested unlawfully!'

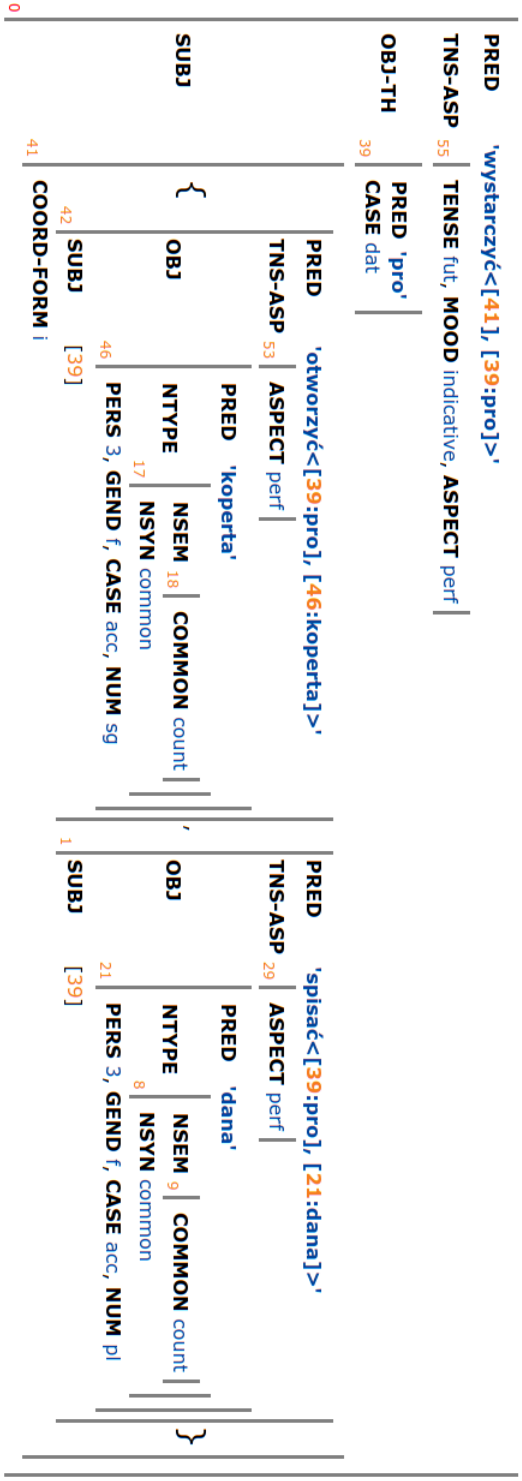


Figure 2.8: F-structure of (2.7)

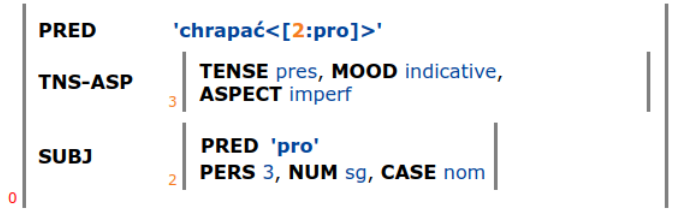


Figure 2.9: F-structure of (2.8)

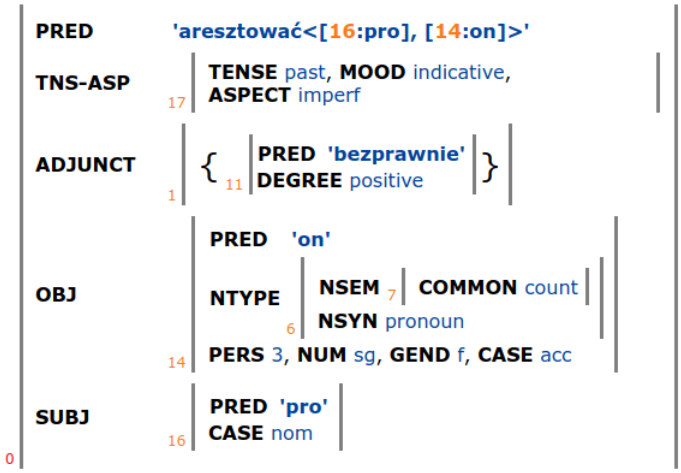


Figure 2.10: F-structure of (2.1)

SIĘ impersonal

An alternative to morphological impersonals in Polish is a constructional impersonal formed by using SIĘ together with a default agreement form of the verb (3SG or 3SG.N): see (2.9) and (2.10), the latter of which includes two instances of impersonal SIĘ under coordination. As explained in Section 2.1, the analysis of such constructions in the LFG structure bank has changed over time, so there are two types of representation, the recent one (Patejuk and Przepiórkowski 2015a) being more detailed and more expressive. The f-structure in Figure 2.11 corresponds to (2.9) and it provides the older representation, whereby the IMPERSONAL attribute is used to distinguish impersonal SIĘ. As with morphological impersonals, the subject is implicit and it is assumed to bear nominative case.

(2.9) Odpowiada się na każde pytanie.
answers.3SG IMPS for every question
'One answers every question.'

By contrast, the f-structure in Figure 2.12,[3] which corresponds to (2.10), uses the complex SIE attribute – both predicates (UKŁADAĆ 'lay' and MALOWAĆ 'paint') contain the SIE attribute whose values in turn contain the '+'-valued attributes IMP and PRESENT. IMP means that SIĘ has the impersonal function, while PRESENT means that SIĘ is local to the relevant predicate (placed in the same clause; it may occur non-locally in infinitival constructions).

[3]The ordering of set elements in INESS visualisations does not always follow the linear order of corresponding constituents in the sentence – it does not in this figure.

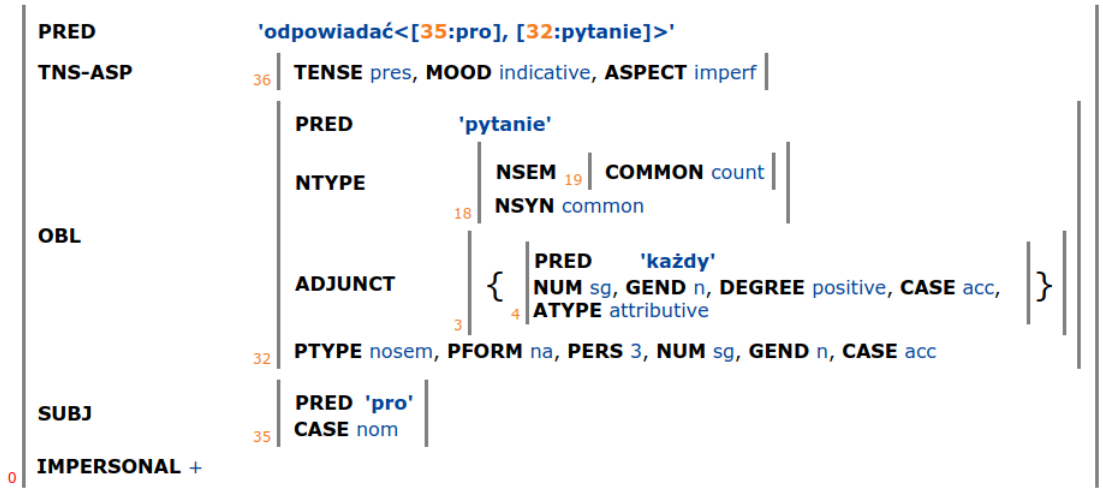


Figure 2.11: F-structure of (2.9)

(2.10) Układa się podłoża pod posadzki i maluje się ściany.
lays.3SG IMPS ground under flooring and paints.3SG IMPS walls
'One lays the ground for the flooring and one paints the walls.'

2.3.4 Subject shared under coordination

The following subsections discuss situations where the subject is a shared dependent under coordination: it may either be lexical or implicit.

Overt subject shared under coordination

In (2.11), the coordinated verbs *uciekł* 'escaped' and *opowiedział* 'relayed, told' take a shared nominative subject, *chłopak* 'boy, lad' (placed to their left), which agrees with both verbs, as shown in glosses (SG.M). The f-structure in Figure 2.13 shows that the topmost f-structure, 0, contains a set containing two predicates: UCIEC 'escape', 33, and OPOWIEDZIEĆ 'relay, tell', 1. Both contain a SUBJ attribute filled by the predicate CHŁOPAK 'boy, lad', 65, whose value of CASE is NOM. Multiple occurrences of the index 65 explicitly indicate that the subject is shared by the two coordinated predicates. Though by convention the contents of 65 (the attribute-value pairs) are fully expanded only in one place in INESS visualisations, all occurrences of the same index point to the same functional substructure.

(2.11) Chłopak uciekł i opowiedział wszystko policji.
boy.NOM.SG.M escaped.3SG.M and told.3SG.M everything.ACC police.DAT
'The boy escaped and told everything to the police.'

Similarly to (2.11) discussed above, (2.12) features verbal coordination with a shared subject: the verbs *zerwał się* 'started' and *uderzał* 'hit' take a shared nominative subject, *wiatr* 'wind', which agrees with both verbs, as shown in glosses (SG.M). The difference with respect to (2.11)

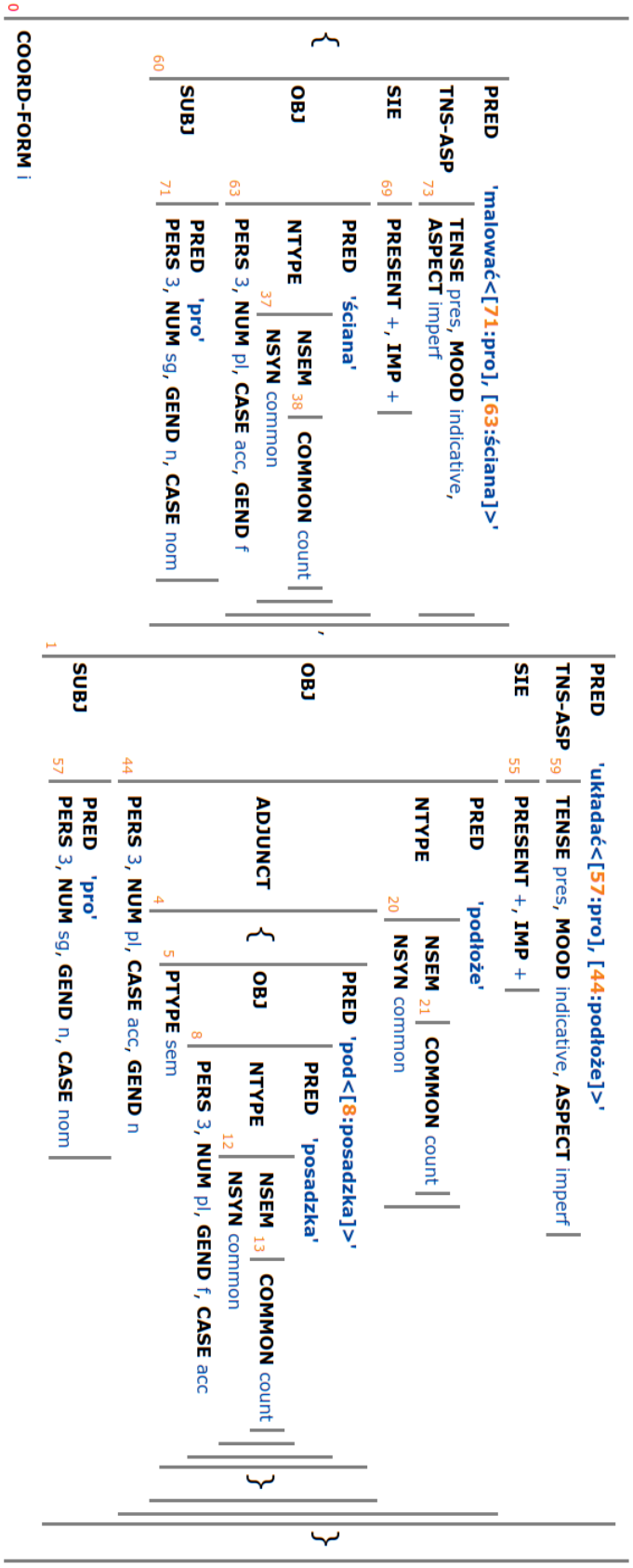


Figure 2.12: F-structure of (2.10)

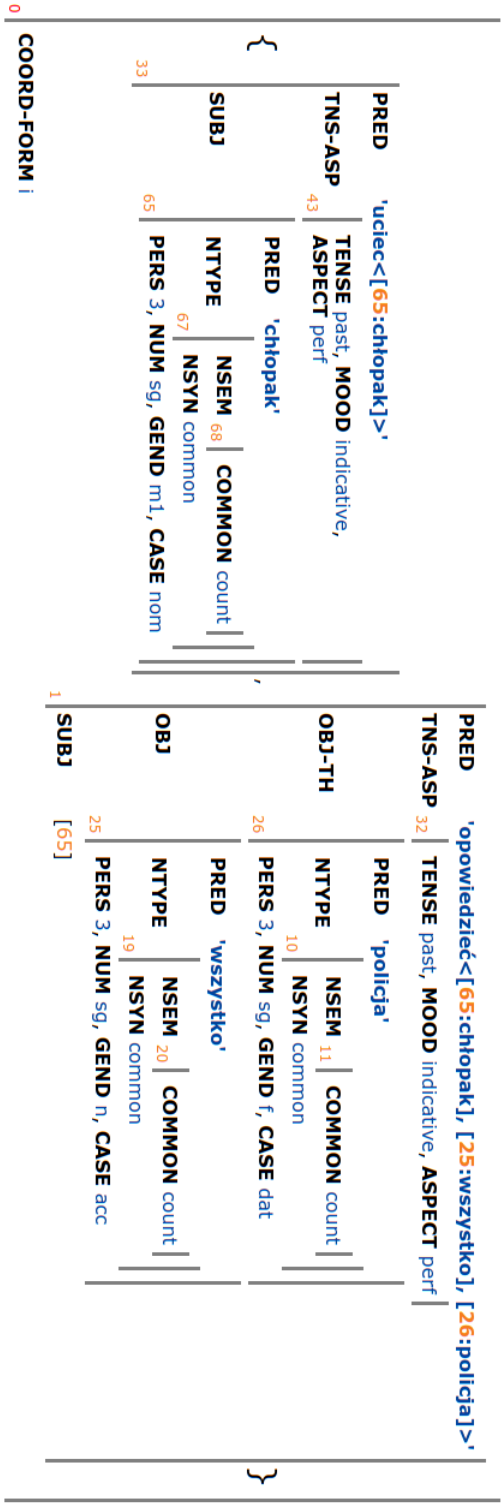


Figure 2.13: F-structure of (2.11)

is that the shared subject in (2.12) is not placed to the left of both verbs – instead, it is placed inside the first conjunct, to the right of the first verb. This, however, does not preclude the subject from being shared, which is shown in the f-structure in Figure 2.14. There, the topmost f-structure, 0, contains a set, which in turn contains two predicates: ZERWAĆ_SIĘ 'start', 54, and UDERZAĆ 'hit', 1. Both feature a SUBJ attribute filled by the predicate WIATR 'wind', 75, whose value of CASE is NOM.

(2.12) Zerwał się wiatr i uderzał ich coraz bardziej w
started.3SG.M INH wind.NOM.SG.M and hit.3SG.M they.ACC increasingly more in
policzek.
cheek
'The wind started and hit them in the cheek harder and harder.'

Implicit subject shared under coordination

In (2.13), the coordinated verbs *otworzył* 'opened' and *wszedł* 'entered' take a shared implicit nominative subject, which agrees with both verbs, as shown in glosses (SG.M). The f-structure in Figure 2.15 shows that the topmost f-structure, 0, contains a set, which in turn contains two predicates: OTWORZYĆ 'open', 1, and WEJŚĆ 'enter', 31. Both contain a SUBJ attribute filled by the predicate PRO, 51, whose value of CASE is NOM (again, implicit subjects are assumed to be nominative).

(2.13) Otworzył drzwi i wszedł do sekretarek.
opened.3SG.M doors.ACC and entered.3SG.M to secretary.PL.F.GEN
'He opened the door and visited the secretaries.'

2.4 Passivisable object (OBJ)

In the LFG structure bank, direct objects are defined as those dependents of verbs which become subjects when the verb passivises. The most typical such objects – exemplified in Section 2.4.1 – are marked for the so-called structural case (Rouveret and Vergnaud 1980; Babby 1980b, 1980a; Przepiórkowski 1999; Przepiórkowski and Patejuk 2012a, 2012b; Patejuk and Przepiórkowski 2014b), i.e., very roughly, they occur in the accusative case in the absence of negation and in the genitive case in the presence of negation (see Przepiórkowski 2000 for details), and they also occur in the genitive as dependents of gerunds. However, as shown in Section 2.4.2, direct objects may also occur in so-called lexical cases. Potentially somewhat confusingly, in LFG – and, hence, also in the LFG structure bank of Polish – the OBJ attribute is also used for marking the sole arguments of prepositions, a practice extended in the current structure bank to numerals; such uses of OBJ are exemplified in Section 2.4.3.

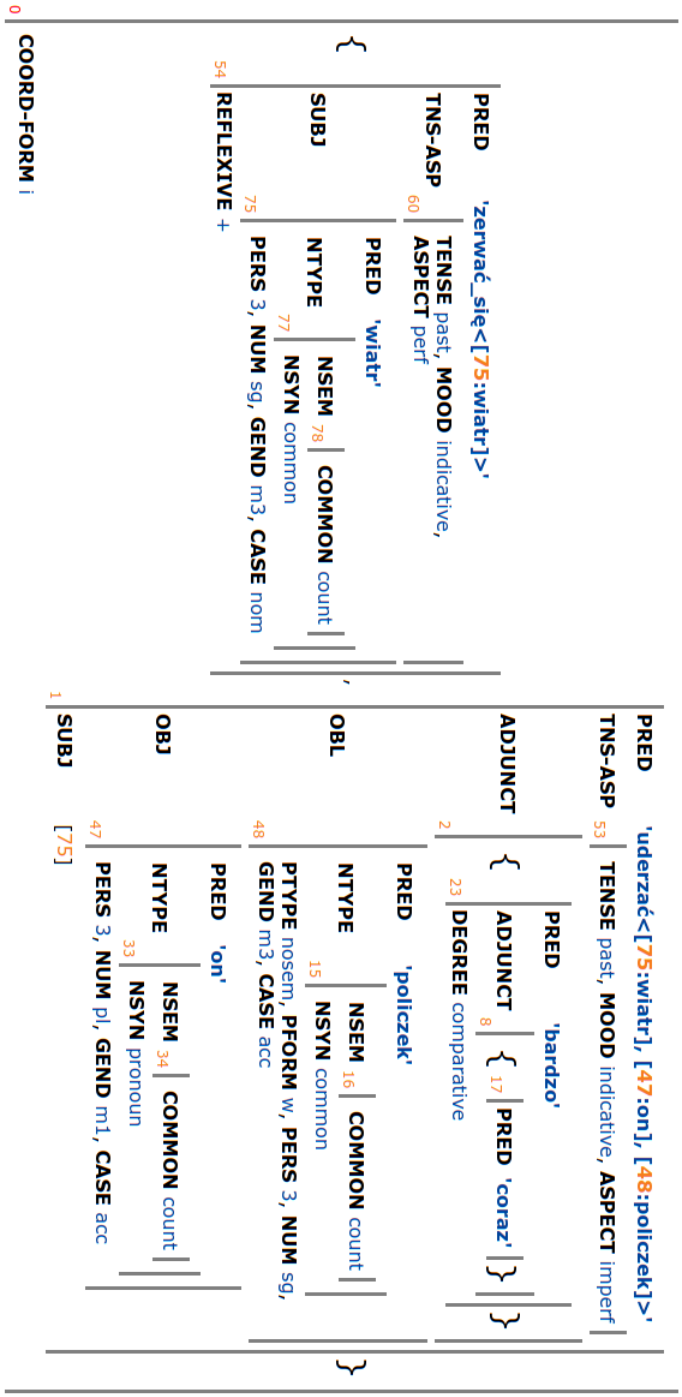


Figure 2.14: F-structure of (2.12)

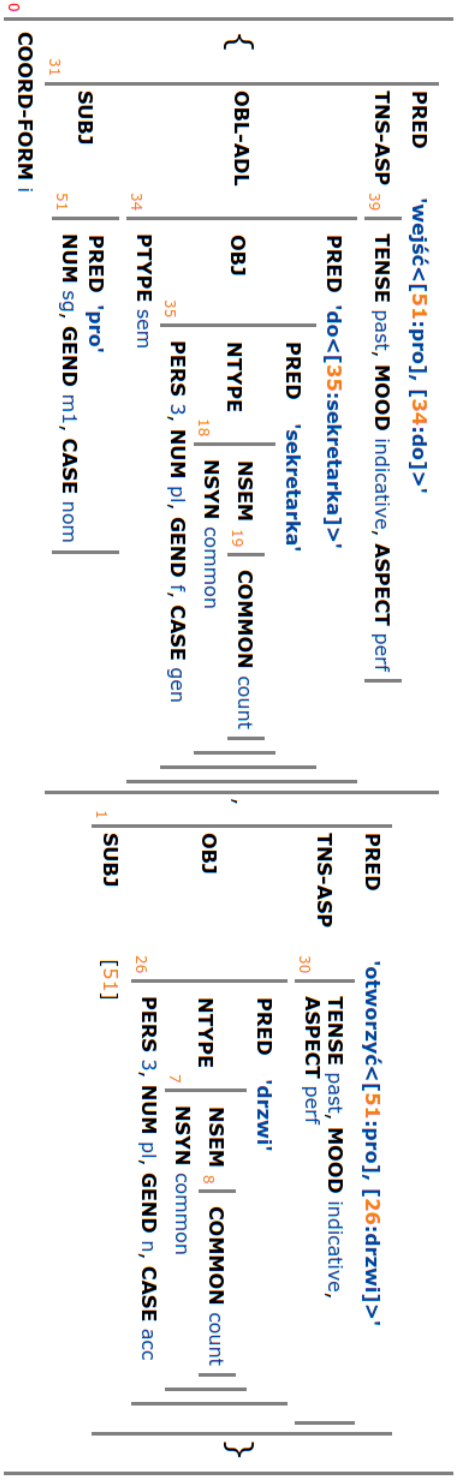


Figure 2.15: F-structure of (2.13)

2.4.1 Passivisable object marked for structural case

Accusative as structural case

The verb *akceptuje* 'accepts' in (2.14) is not negated, so its object, *decyzję* 'decision', bears accusative as the value of structural case, as shown in glosses. The f-structure in Figure 2.16 shows that the predicate AKCEPTOWAĆ 'accept', 0, contains an OBJ attribute, 23, filled by the predicate DECYZJA 'decision', whose value of CASE is ACC.

(2.14) Akceptuje naszą decyzję!
accepts.3SG our.ACC.SG.F decision.ACC.SG.F
'(He/she/it) accepts our decision!'

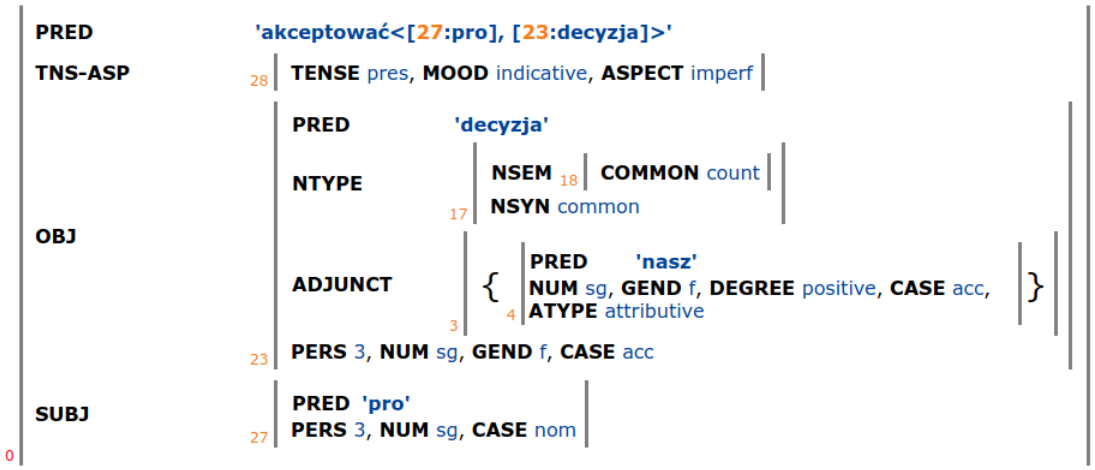


Figure 2.16: F-structure of (2.14)

Genitive as structural case under negation

The verb *akceptuje* 'accepts' in (2.15) is negated (the word *nie* 'not' is present), so its object, *homoseksualizmu* 'homosexuality', bears genitive as the value of structural case, as shown in glosses. The f-structure in Figure 2.17 shows that the predicate AKCEPTOWAĆ 'accept', 0, contains a NEG attribute with value + and an OBJ attribute, 2, filled by the predicate HOMOSEKSUALIZM 'homosexuality', whose value of CASE is GEN.

(2.15) Nie akceptuje homoseksualizmu.
NEG accepts.3SG homosexuality.GEN.SG.M
'(He/she/it) does not accept homosexuality.'

Genitive as structural case with gerund heads

The gerund *pozyskanie* 'gaining, acquisition' in (2.16) takes an object, *sponsora*, which bears genitive as the value of structural case (regardless of the presence of negation), as shown in

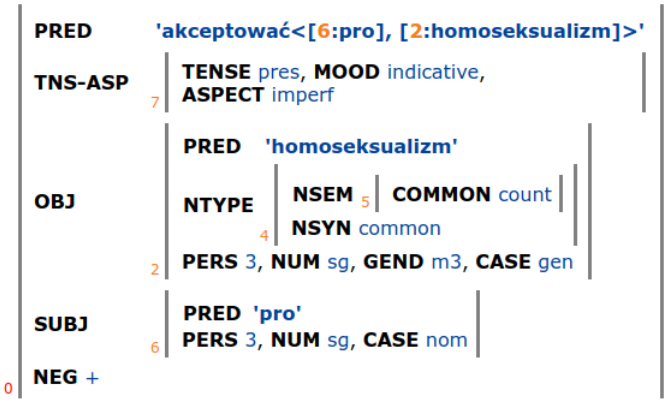


Figure 2.17: F-structure of (2.15)

glosses. The f-structure in Figure 2.18 shows that the predicate POZYSKAĆ 'gain, acquire', 4, contains an OBJ attribute, 8, filled by the predicate SPONSOR 'sponsor', whose value of CASE is GEN.

(2.16) Ale kluczowe jest pozyskanie sponsora.
but crucial.NOM.SG.N is.3SG gain.GER.NOM.SG.N sponsor.GEN.SG.M
'But the crucial thing is to find a sponsor.'

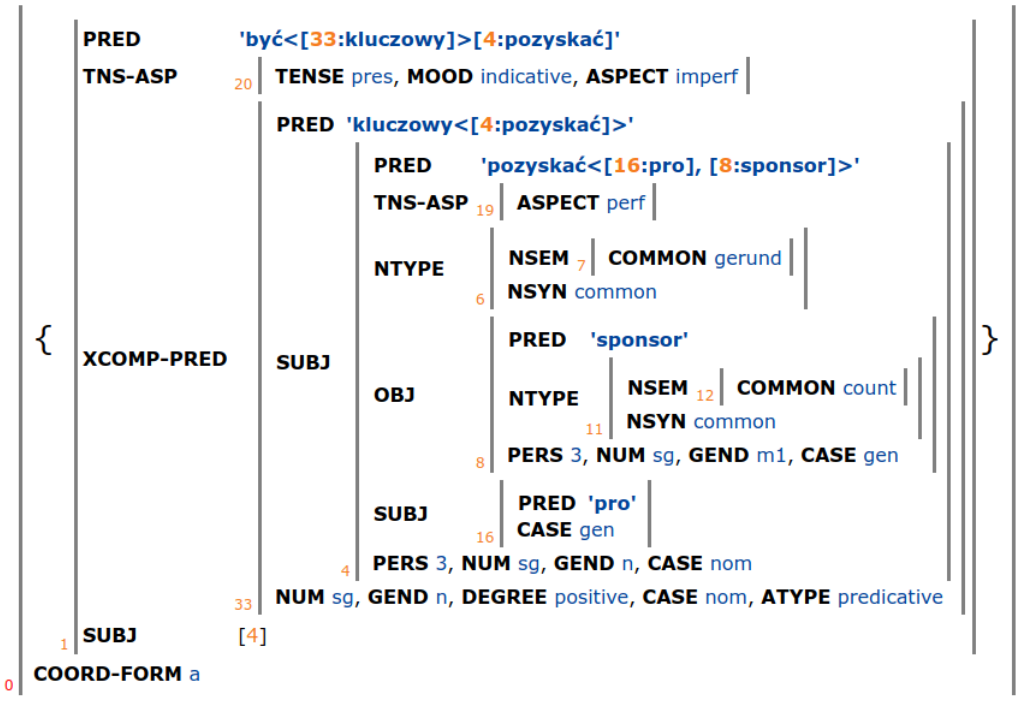


Figure 2.18: F-structure of (2.16)

2.4.2 Passivisable object marked for lexical case

Genitive

The verb *dokonało* 'accomplished' in (2.17) takes an object, *tego* 'this', which is marked for lexical genitive case (as opposed to structural genitive case discussed above), as shown in glosses. The f-structure in Figure 2.19 shows that the predicate DOKONAĆ 'accomplish', 0, contains an OBJ attribute, 20, filled by the predicate TO 'this', whose value of CASE is GEN.

(2.17) Dokonało tego dwóch młodzieńców.
accomplished.3SG.N this.GEN.SG.N two.ACC.PL.M youngsters.GEN.PL.M
'Two young people accomplished this.'

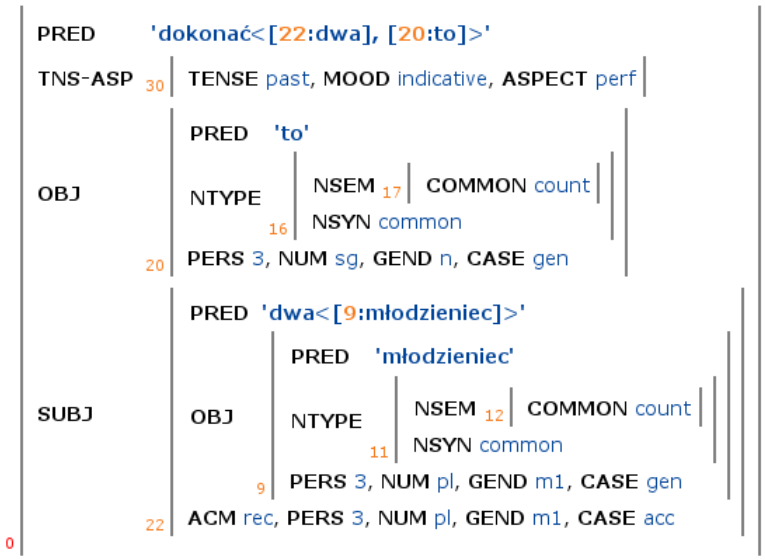


Figure 2.19: F-structure of (2.17)

Instrumental

The verb *kierował* 'lead, run' in (2.18) takes an object, *towarzystwem* 'association, society', which is marked for instrumental case (which is always lexical), as shown in glosses. The f-structure in Figure 2.20 shows that the predicate KIEROWAĆ 'lead, run', 0, contains an OBJ attribute, 42, filled by the predicate TOWARZYSTWO 'association, society', whose value of CASE is INST.[4]

(2.18) Będzie kierował on towarzystwem do 2007 roku.
will.3SG lead.3SG.M he.NOM.SG.M association.INS.SG.N to 2007 year
'He will run the association until 2007.'

[4]Note that in the glosses we follow the Leipzig Glossing Rules and, hence, abbreviate instrumental to INS, while in the LFG structure bank we follow the legacy tagset and, hence, abbreviate instrumental to INST.

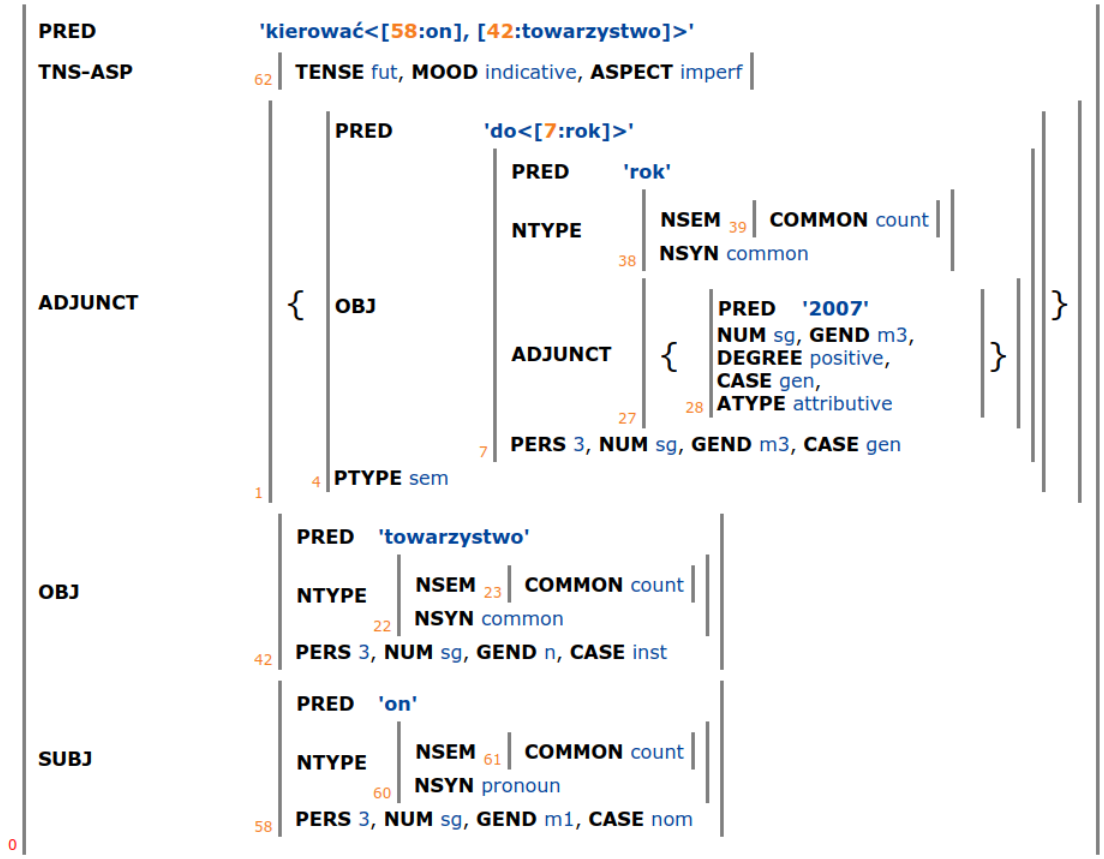


Figure 2.20: F-structure of (2.18)

2.4.3 Other uses of the OBJ attribute

Traditionally, the OBJ attribute is used in LFG also to mark the arguments of prepositions. This is illustrated in Figure 2.15 on page 30. There, the first element of the coordination, with index 31, corresponding to the constituent *wszedł do sekretarek* 'visited secretaries' (lit. 'entered to secretaries') in sentence (2.13), contains a dependent, 34, corresponding to the prepositional phrase *do sekretarek* 'to secretaries'. The main predicate of this dependent is the semantic preposition DO 'to', and its argument 35 is represented as the value of OBJ, even though this argument is not a direct object in the sense defined above.

In the LFG structure bank, this use of OBJ, where it does not mark a direct object, is extended to numeral phrases. For example, in Figure 2.19, the value of SUBJ, 22, represents the numeral phrase *dwóch młodzieńców* 'two youngsters'; the main predicate is the numeral DWA 'two', and its nominal dependent 9 is marked as an OBJ, even though it is not a direct object. While potentially misleading, such uses of OBJ are constrained to dependents of prepositions and numerals, so they are easy to distinguish from the standard uses of OBJ, as direct objects of verbs.

2.5 Dative indirect object (OBJ-TH)

The verb *przyniosły* 'brought' in (2.5), repeated below, takes an indirect object, *mu* 'him', which is marked for dative case (which is always lexical), as shown in glosses. The f-structure in Figure 2.6, repeated below as Figure 2.21, shows that the predicate PRZYNIEŚĆ 'bring', 0, contains an OBJ-TH attribute, 22, filled by the predicate ON 'he', whose value of CASE is DAT.

(2.5) Lekkie stuknięcia młotka przyniosły mu spokój.
gentle.NOM.PL.N knock.GER.NOM.PL.N hammer.GEN.SG.M brought.3PL.N he.DAT.SG.M peace.ACC.SG.M
'Gentle knocks of the hammer brought him peace.'

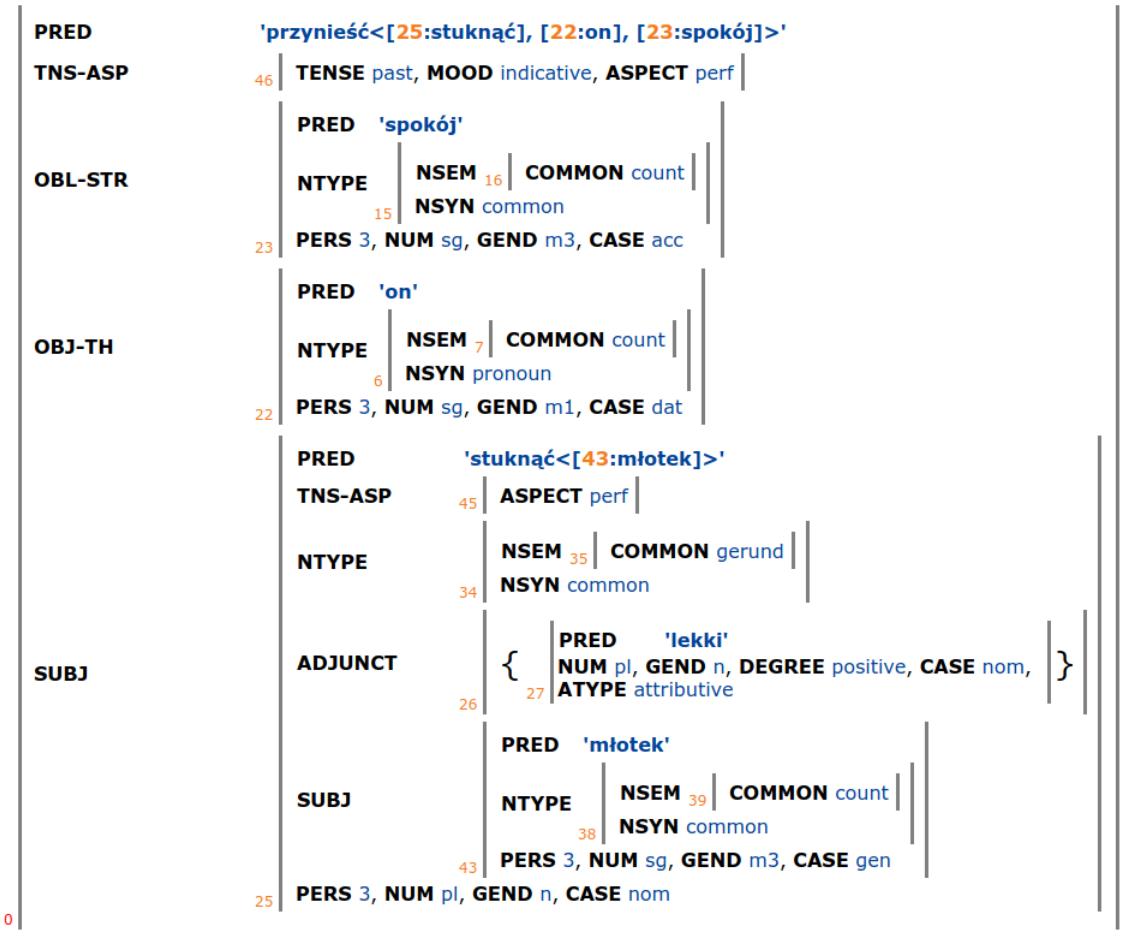


Figure 2.21: F-structure of (2.5)

2.6 Other non-passivisable complements (OBL-<CASE>)

2.6.1 Non-passivisable complement marked for structural case (OBL-STR)

The verb *ma* 'has' in (2.19) is not negated, so its non-passivisable complement, *naturę* 'nature', bears accusative case, as shown in glosses. The f-structure in Figure 2.22 shows that the predicate MIEĆ 'have', 0, contains an OBL-STR attribute, 23, filled by the predicate NATURA 'nature', whose value of CASE is ACC.

(2.19) Świat ma naturę hierarchiczną.
world.NOM.SG.M has.3SG nature.ACC.SG.F hierarchical.ACC.SG.F
'The world's nature is hierarchical.'

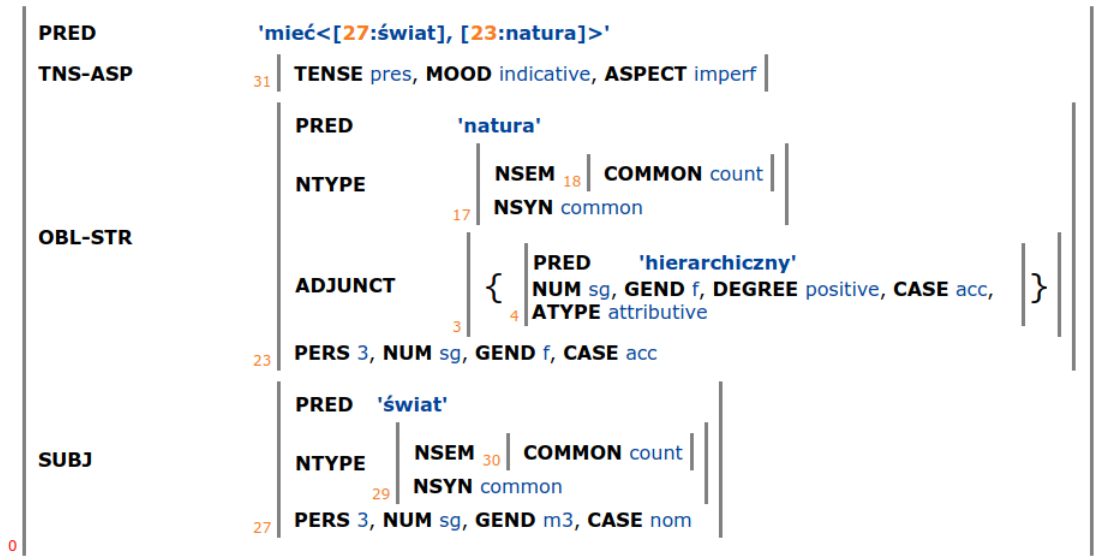


Figure 2.22: F-structure of (2.19)

By contrast, the verb *mają* 'have' in (2.20) is negated (the word *nie* is present), so its non-passivisable complement, *wyboru* 'choice', bears genitive case, as shown in glosses. The f-structure in Figure 2.23 shows that the predicate MIEĆ 'have', 0, contains a NEG attribute with value + and an OBL-STR attribute, 2, filled by the predicate WYBÓR 'choice', whose value of CASE is GEN.

(2.20) Nie mają wyboru.
NEG have.3PL choice.GEN.SG.M
'They have no choice.'

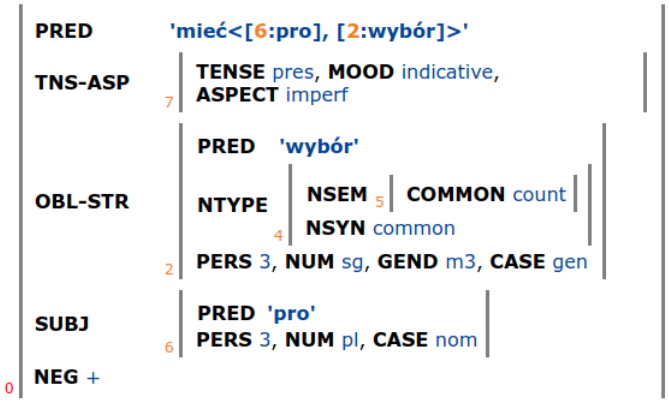


Figure 2.23: F-structure of (2.20)

2.6.2 Non-passivisable complement marked for lexical genitive case (OBL-GEN)

The verb *bać się* 'fear, be afraid' in (2.21)[5] takes a non-passivisable complement *przedterminowych wyborców* 'snap voters' bearing genitive case, as shown in glosses – it is lexical genitive (regardless of the syntactic context, as opposed to genitive as a value of structural case). The f-structure in Figure 2.24 shows that the predicate BAĆ_SIĘ 'fear, be afraid', 15, contains an OBL-GEN attribute, 17, filled by the predicate WYBORCA 'voter', whose value of CASE is GEN.

(2.21) Partia opozycyjna nie powinna bać się przedterminowych
party.NOM.SG.F opposition.NOM.SG.F NEG should.3SG.F fear.INF INH early.GEN.PL.M
wyborców.
voter.GEN.PL.M
'The opposition party should not be afraid of snap voters.'

2.6.3 Non-passivisable complement marked for instrumental case (OBL-INST)

The verb *wzruszył* 'shrugged' in (2.22) takes a non-passivisable complement *ramionami* 'shoulders' bearing instrumental case, as shown in glosses. The f-structure in Figure 2.25 shows that the predicate WZRUSZYĆ 'shrug', 0, contains an OBL-INST attribute, 12, filled by the predicate RAMIĘ 'arm, shoulder', whose value of CASE is INST.

(2.22) Chłopiec wzruszył ramionami.
boy.NOM.SG.M shrugged.3SG.M shoulder.INS.PL.N
'The boy shrugged his shoulders.'

[5]It is not clear whether (2.21) contains a typo (*wyborców* 'voters' instead of *wyborów* 'elections'). This, however, has no bearing on the issues discussed here.

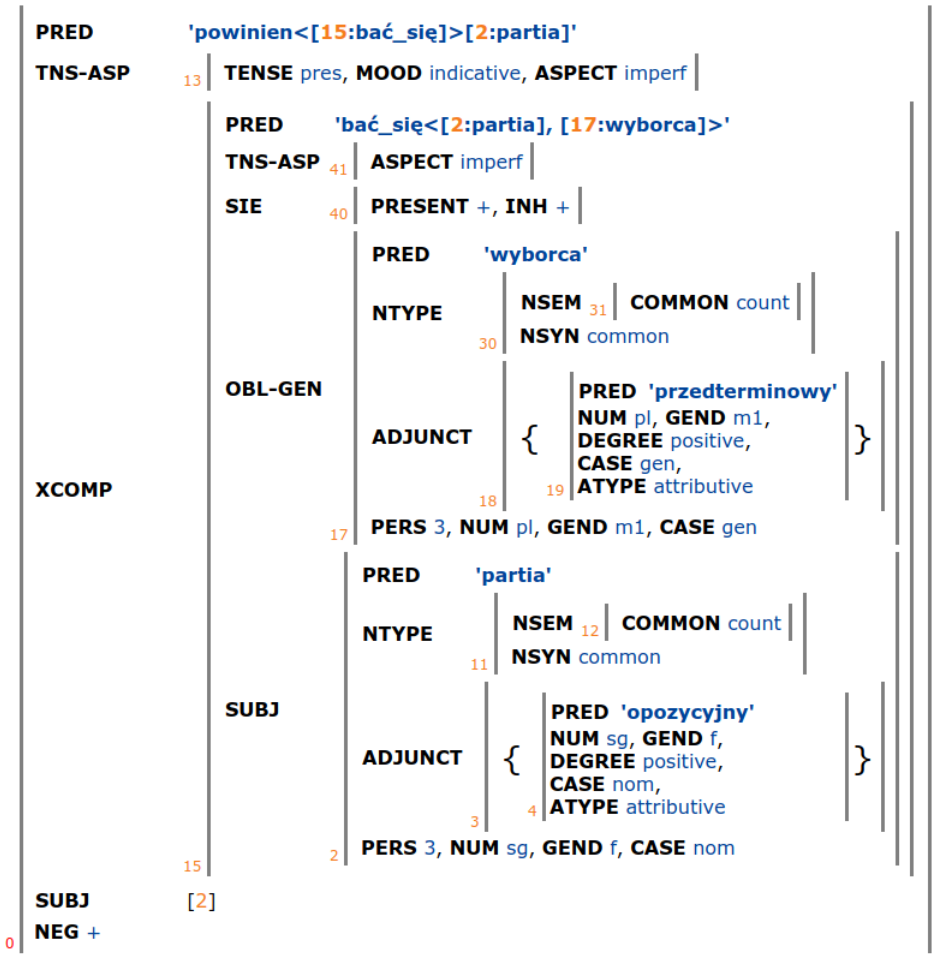


Figure 2.24: F-structure of (2.21)

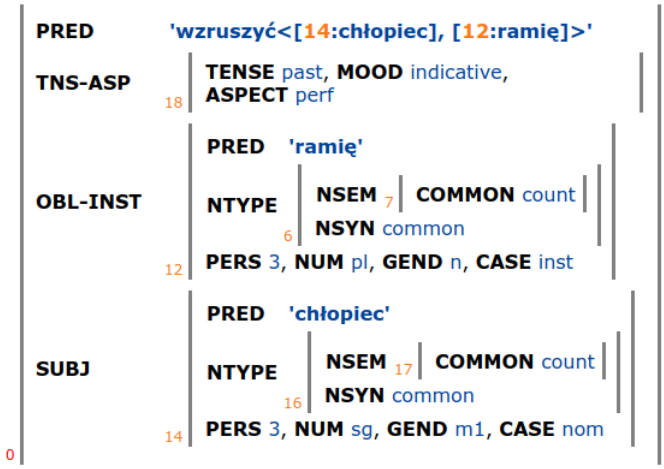


Figure 2.25: F-structure of (2.22)

2.7 Non-semantic obliques (OBL, OBL2)

The verb *rozmawiałeś* '(you) talked' in (2.23) takes two non-semantic oblique complements: one requires the preposition O 'about' taking locative case, while the other requires the preposition Z 'with' taking instrumental case. The f-structure in Figure 2.26 shows that the predicate ROZMAWIAĆ 'talk', 0, contains two obliques: OBL and OBL2, both of which are filled by prepositional phrases. In LFG, non-semantic prepositional phrases do not introduce a PRED attribute of their own (because they are non-semantic) – instead, they introduce a PFORM attribute whose value corresponds to the lemma of the preposition used; moreover, the value of their PTYPE is NOSEM. As a result, the PRED of non-semantic obliques is contributed by the nominal: OBL, 82, is filled by the predicate TO 'this' which bears locative case, as required by the preposition O 'about', which contributes its PFORM. Similarly, OBL2, 100, is filled by the predicate PRZYJACIEL 'friend' bearing instrumental case, as required by the preposition Z 'with', which contributes its PFORM.

(2.23) Czy kiedykolwiek rozmawiałeś o tym z twoimi przyjaciółmi?
Q ever talked.2SG.M about this.LOC.SG.N with your.INS.PL.M friend.INS.PL.M
'Have you ever talked about this to your friends?'

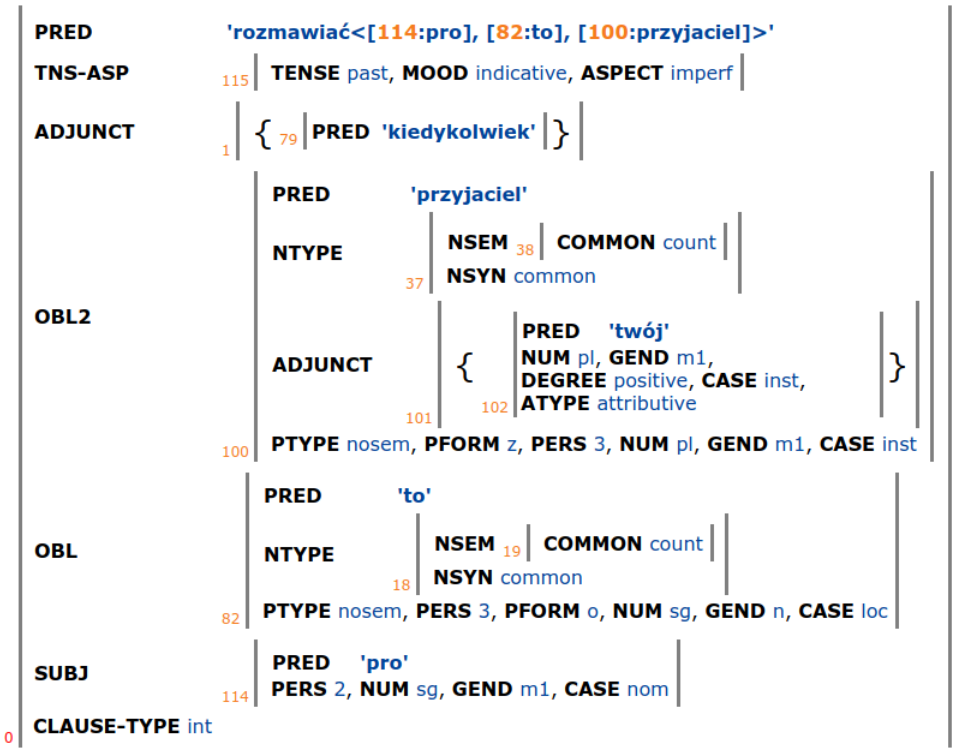


Figure 2.26: F-structure of (2.23)

2.8 Agent oblique (OBL-AG)

The OBL-AG grammatical function represents the agent of passive participles. Such passive participles occur in passive constructions, where the participle acts as a predicative item (XCOMP-PRED; cf. Section 2.13). This is illustrated by (2.24). Alternatively, passive participles may also function as modifiers (ADJUNCT; cf. Section 2.14), as in (2.25).

(2.24) Sad ten został założony przez
orchard.NOM.SG.M this.NOM.SG.M became.3SG.M established.NOM.SG.M by
mego Ojca.
my.ACC.SG.M father.ACC.SG.M
'This orchard was established by my father.'

(2.25) Zostawiono drzewa wskazane przez specjalistów.
left.IMPS tree.ACC.PL.N selected.ACC.PL.N by specialist.ACC.PL.M
'They left the trees selected by specialists.'

The passive participle *założony* 'established, set up' in (2.24) takes an agent oblique, *przez mego ojca* 'by my father', which is a prepositional phrase consisting of the preposition *przez* 'by' and the nominal *ojca* 'father' (modified by *mego* 'my'). The f-structure in Figure 2.27 shows that the predicate ZAŁOŻYĆ 'establish, set up', 85, contains an OBL-AG attribute, 71, filled by the predicate OJCIEC 'father' which bears accusative case, as required by the preposition PRZEZ 'by' which contributes its PFORM (since, by convention, it is assumed to be non-semantic). (2.24) is a canonical example of passive voice: *został* 'became', the auxiliary verb, provides the main predicate (it contributes the value of PRED, namely, ZOSTAĆ 'become'), while the passive participle *założony* 'established, set up' is analysed as its predicative complement (XCOMP-PRED); see the LFG analysis of Polish passive in Patejuk and Przepiórkowski 2014a.

By contrast, (2.25) does not feature the auxiliary – the passive participle is used as a modifier. In spite of this difference, the passive participle *wskazane* 'indicated, selected' also takes an agent oblique – the prepositional phrase *przez specjalistów* 'by specialists', which consists of the preposition *przez* 'by' and the nominal *specjalistów* 'specialists'. The f-structure in Figure 2.28 shows that the predicate WSKAZAĆ 'indicate, select', 4, contains an OBL-AG attribute, 15, filled by the predicate SPECJALISTA 'specialist', which bears accusative case, as required by the preposition PRZEZ 'by', which contributes its PFORM.

2.9 Semantic obliques (OBL-<SEM>)

The Polish LFG grammar which underlies the structure bank uses a wide range of semantic obliques. Unlike in the case of non-semantic obliques discussed above, semantic obliques do not have to be nominal or prepositional phrases – depending on the semantic type, a given semantic oblique may correspond to a variety of c-structure categories (including: PP, ADVP, CP and sometimes bare NP), which may often be coordinated. If the semantic oblique is a PP, the preposition is semantic, i.e., it contributes its PRED, the value of its PTYPE is SEM, and it takes a nominal OBJ (as described in Section 2.4.3 above).

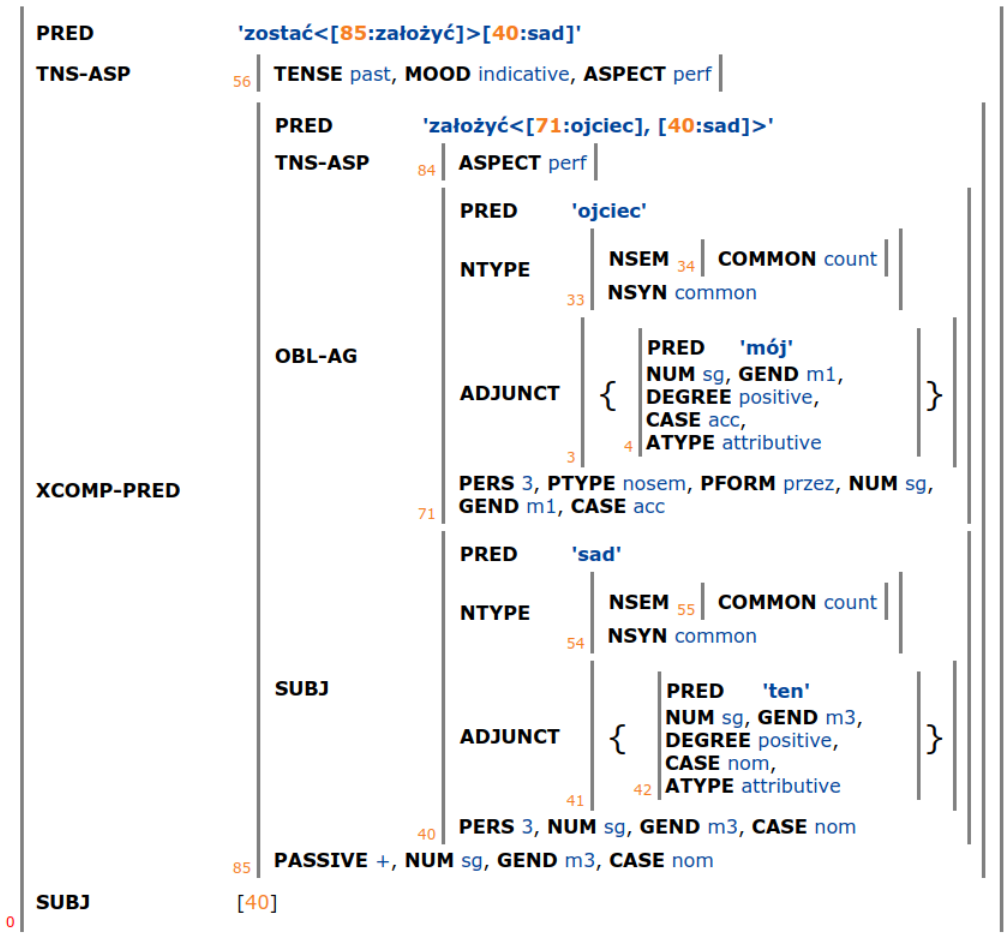


Figure 2.27: F-structure of (2.24)

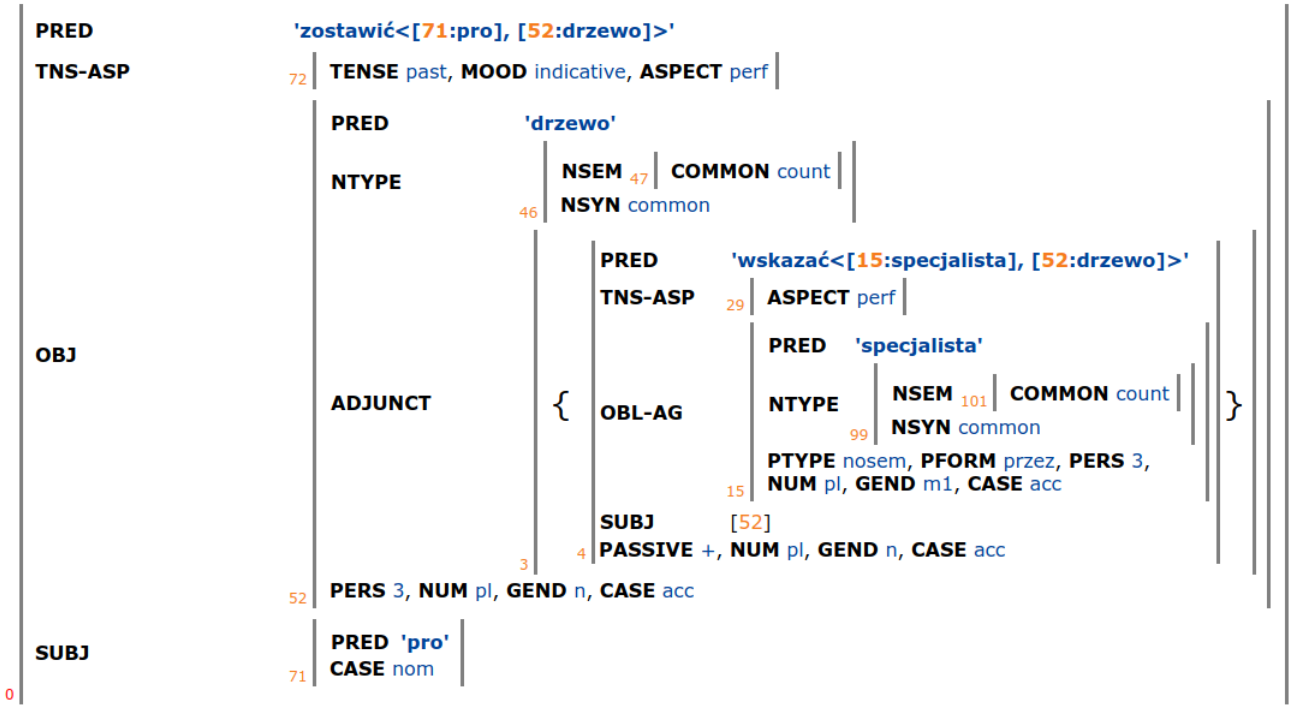


Figure 2.28: F-structure of (2.25)

2.9.1 Comparative oblique (OBL-COMPAR)

The adverb *bardziej* 'more' in (2.26) takes a comparative oblique, *od człowieka* 'than human', which is a prepositional phrase consisting of the preposition *od* 'from, than' and the nominal *człowieka* 'human'. The f-structure in Figure 2.29 shows that the predicate BARDZO 'very', 24, contains an OBL-COMPAR attribute, 25, filled by the predicate OD 'from, than', which takes the predicate CZŁOWIEK 'human', 8, as its genitive argument.

(2.26) Mechanizm okazał się bardziej trwały od człowieka.
mechanism.NOM.SG.M turned out.3SG.M INH more durable.3SG.M from human.GEN.SG.M
'The mechanism turned out to be more durable than a human.'

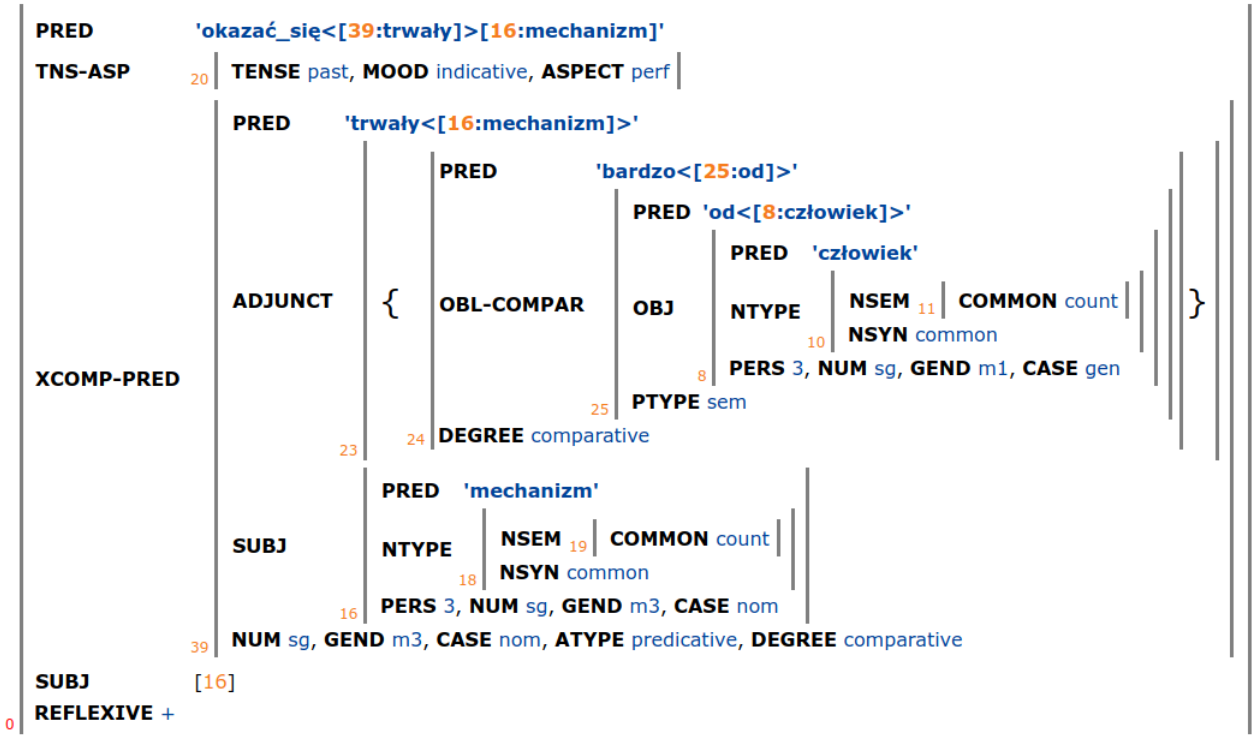


Figure 2.29: F-structure of (2.26)

2.9.2 Ablative oblique (OBL-ABL)

The verb *przywieziono* 'transported in, brought' in (2.27) takes the adverb *stamtąd* 'from there' as the ablative oblique. The f-structure in Figure 2.30 shows that the predicate PRZYWIEŹĆ 'transport in, bring', 0, contains an OBL-ABL attribute, 52, filled by the predicate STAMTĄD 'from there'.

(2.27) Stamtąd przywieziono do Poznania obie rogówki zmarłego.
from there.ADV transported.IMPS to Poznań.GEN.SG.M both corneas deceased.GEN
'Both corneas of the deceased person were transported from there to Poznań.'

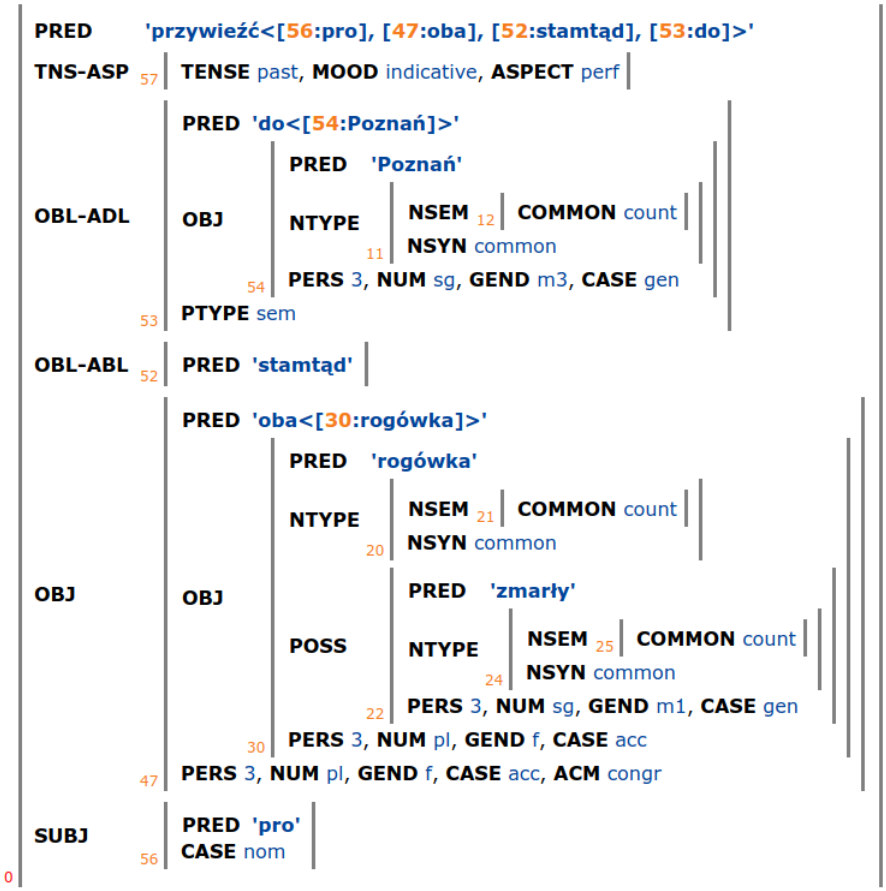


Figure 2.30: F-structure of (2.27)

2.9.3 Adlative oblique (OBL-ADL)

The verb *przywieziono* 'transported in, brought' in (2.27) above takes the prepositional phrase *do Poznania* 'to Poznań' as the adlative oblique, which consists of the preposition *do* 'to' and the nominal *Poznania* 'Poznań' (a Polish city). The f-structure in Figure 2.30 shows that the predicate PRZYWIEŹĆ 'transport in, bring', 0, contains an OBL-ADL attribute, 53, filled by the predicate DO 'to' which takes the predicate POZNAŃ, 54, as its genitive argument.

2.9.4 Perlative oblique (OBL-PERL)

The verb *płynie* 'flows' in (2.28) takes the adverb *tamtędy* 'that way' as the perlative oblique. The f-structure in Figure 2.31 shows that the predicate PŁYNĄĆ 'flow', 0, contains an OBL-PERL attribute, 6, filled by the predicate TAMTĘDY 'that way'.

(2.28) Tamtędy płynie Stobrawa.
that way.ADV flows.3SG Stobrawa.NOM.SG.F
'Stobrawa flows that way.'

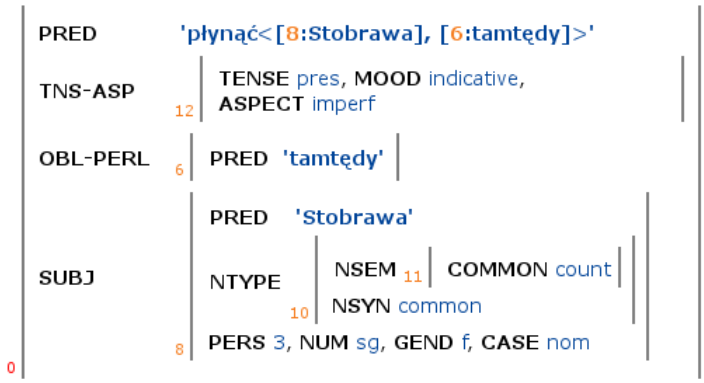


Figure 2.31: F-structure of (2.28)

2.9.5 Locative oblique (OBL-LOCAT)

The verb *znajdowało się* 'find itself, be located' in (2.29) takes the prepositional phrase *przy drzwiach wejściowych* 'by the entrance door' as the locative oblique, which consists of the preposition *przy* 'near, by' and the nominal *drzwiach* 'door' (modified by *wejściowych* 'entrance.ADJ'). The f-structure in Figure 2.32 shows that the predicate ZNAJDOWAĆ_SIĘ 'find itself, be located', 0, contains an OBL-LOCAT attribute, 19, filled by the predicate PRZY 'near, by' which takes the predicate DRZWI 'door', 20, as its locative argument.

(2.29) Źródło ogni znajdowało się przy drzwiach wejściowych.
source.NOM.SG.N fire.GEN found.3SG.N INH near door.LOC.PL.N entrance.ADJ.LOC.PL.N
'The source of fire was located near the entrance door.'

2.9.6 Manner oblique (OBL-MOD)

The verb *czuł się* 'felt' in (2.30) takes the adverb *fatalnie* 'terribly' as the manner oblique. The f-structure in Figure 2.33 shows that the predicate CZUĆ_SIĘ 'feel', 0, contains an OBL-MOD attribute, 8, filled by the predicate FATALNIE 'terribly'.

(2.30) Czuł się fatalnie.
felt.3SG.M INH terrible.ADV
'He felt terrible.'

2.9.7 Temporal oblique (OBL-TEMP)

The verb *odbędzie się* 'will happen, will take place' in (2.31) takes the prepositional phrase *w poniedziałek* 'on Monday' as the temporal oblique, which consists of the preposition *w* 'in, on' and the nominal *poniedziałek* 'Monday'. The f-structure in Figure 2.34 shows that the predicate ODBYĆ_SIĘ 'happen, take place', 0, contains an OBL-TEMP attribute, 16, filled by the predicate W 'in' which takes the predicate PONIEDZIAŁEK 'Monday', 17, as its accusative argument.

(2.31) Przesłuchanie odbędzie się w poniedziałek.
questioning.NOM.SG.N happens.3SG INH in Monday.ACC.SG.M
'The questioning will take place on Monday.'

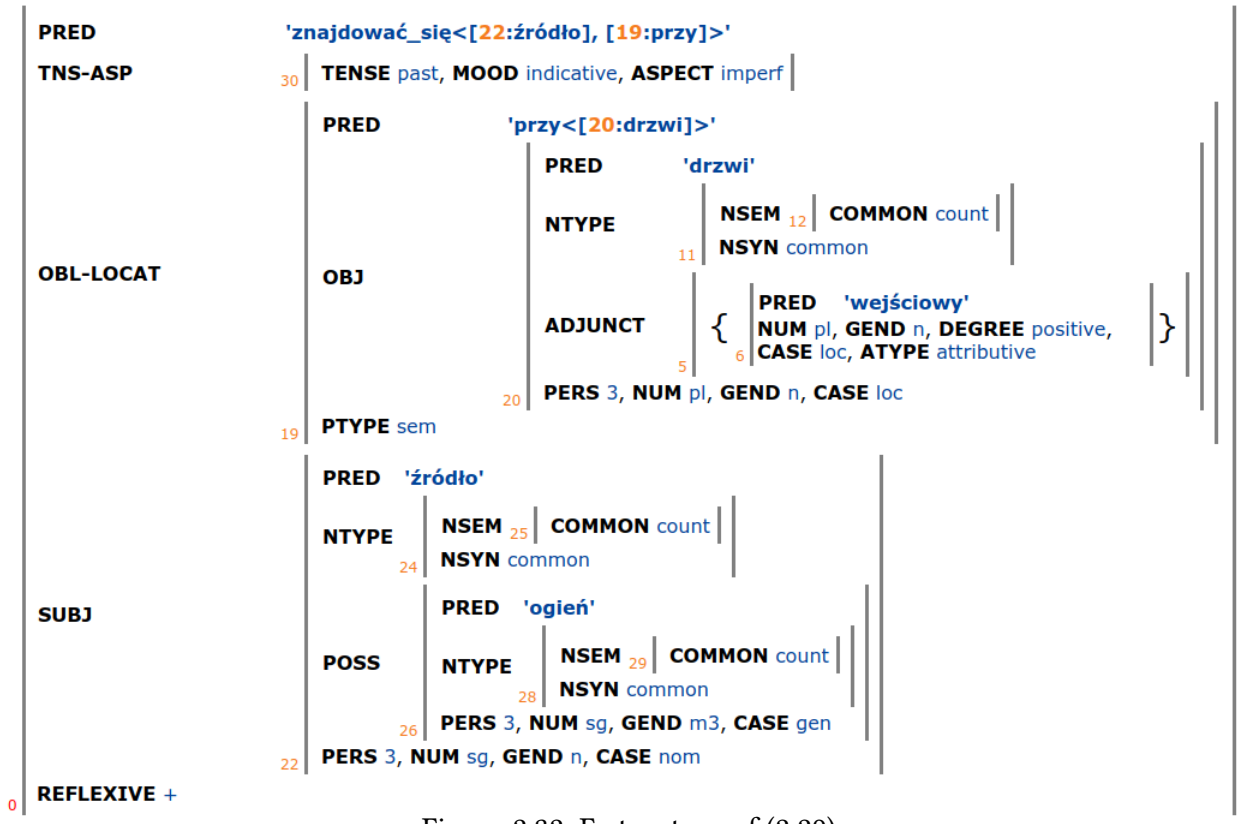


Figure 2.32: F-structure of (2.29)

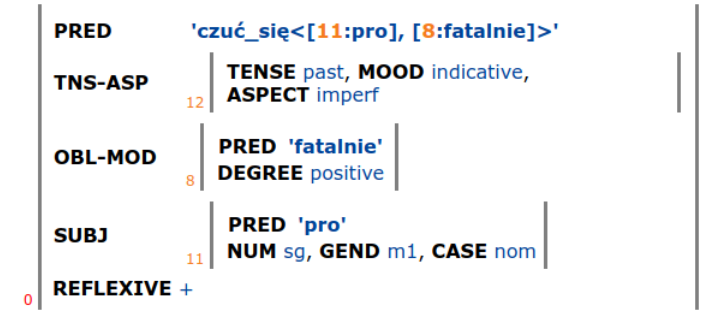


Figure 2.33: F-structure of (2.30)

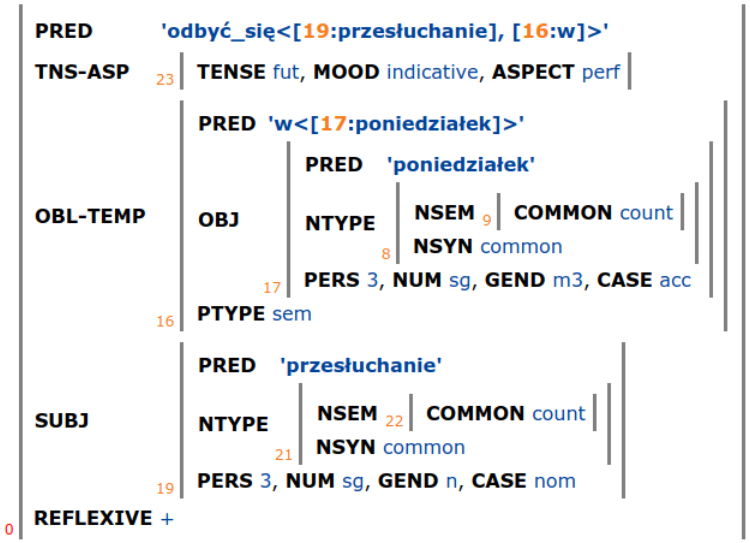


Figure 2.34: F-structure of (2.31)

2.9.8 Durative oblique (OBL-DUR)

The verb *trwały* 'lasted' in (2.32) takes the adverb *krótko* 'briefly, for a short time' as the durative oblique. The f-structure in Figure 2.35 shows that the predicate TRWAĆ 'last', 0, contains an OBL-DUR attribute, 8, filled by the predicate KRÓTKO 'briefly, for a short time'.

(2.32) Przygotowania trwały krótko.
preparation.NOM.PL.N lasted.3PL.N short.ADV
'The preparations were short.'

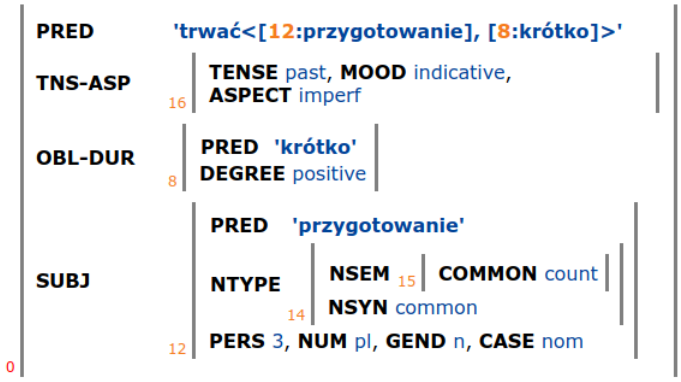


Figure 2.35: F-structure of (2.32)

2.10 Adverbial oblique (OBL-ADV)

The verb *sądzisz* '(you) think' in (2.33) takes *jak* 'how' as the adverbial complement. The f-structure in Figure 2.36 shows that the predicate SĄDZIĆ 'think, believe', 0, contains an OBL-ADV attribute, 7.

(2.33) Jak sądzisz?
how think.2SG
'What do you think?'

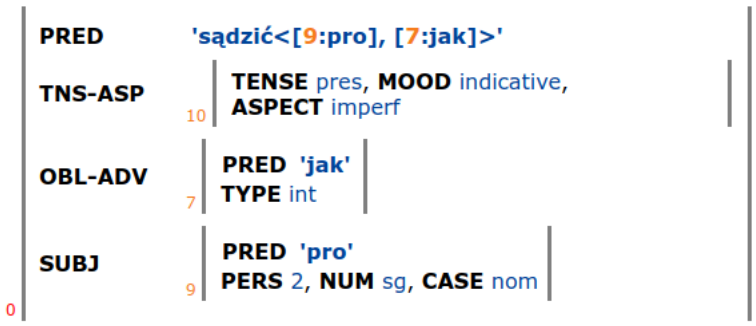


Figure 2.36: F-structure of (2.33)

2.11 Closed clausal complement (COMP)

There are two types of closed clausal dependents: introduced by a semantic complementiser or not. In the former case, the whole clause is typically an adjunct headed by the semantic complementiser, and the rest of the clause is a COMP dependent of that complementiser. Such constructions will be illustrated shortly.

In the other case, the whole clause – with a non-semantic complementiser, if any – is typically a COMP dependent of a higher head. For example, the verb *pamięta* 'remembers' in (2.34) takes the subordinate clause *że pił alkohol* 'that (he) drank alcohol' as the closed clausal complement, which features the non-semantic complementiser *że* 'that'. The f-structure in Figure 2.37 shows that the predicate PAMIĘTAĆ 'remember', 0, contains a COMP attribute, 2, filled by the predicate PIĆ 'drink', which contains the COMP-FORM attribute contributed by the complementiser.

(2.34) Pamięta, że pił alkohol.
remembers.3SG that drank.3SG.M alcohol.ACC.SG.M
'He remembers that he drank alcohol.'

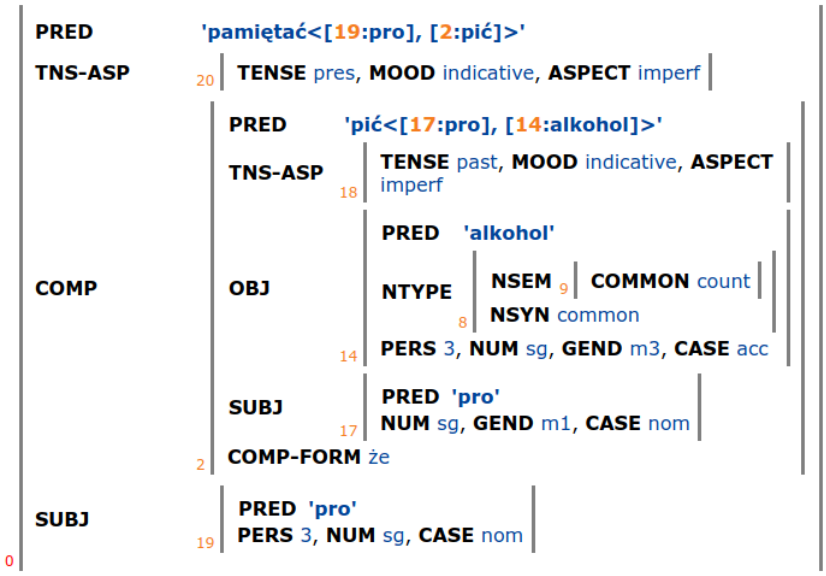


Figure 2.37: F-structure of (2.34)

The verb *domyśla się* 'suspects, guesses' in (2.35) takes the interrogative clause *co tkwi w środku* 'what lies inside' as the closed clausal complement, where *co* 'what' acts as the interrogative item. The f-structure in Figure 2.38 shows that the predicate DOMYŚLAĆ_SIĘ 'suspect, guess', 0, contains a COMP attribute, 74, filled by the predicate TKWIĆ 'lie, be stuck' whose SUBJ attribute, 32, is filled by the interrogative predicate CO 'what' (its TYPE is INT). Note that within the value of COMP, 74, there is no COMP-FORM attribute (as there is no complementiser), but there is a CLAUSE-TYPE attribute, whose value is INT (i.e., interrogative).

(2.35) Nikt nie domyśla się, co tkwi w środku.
nobody.NOM.SG.M NEG suspects.3SG INH what.NOM.SG.N lies.3SG in middle
'Nobody suspects what lies inside.'

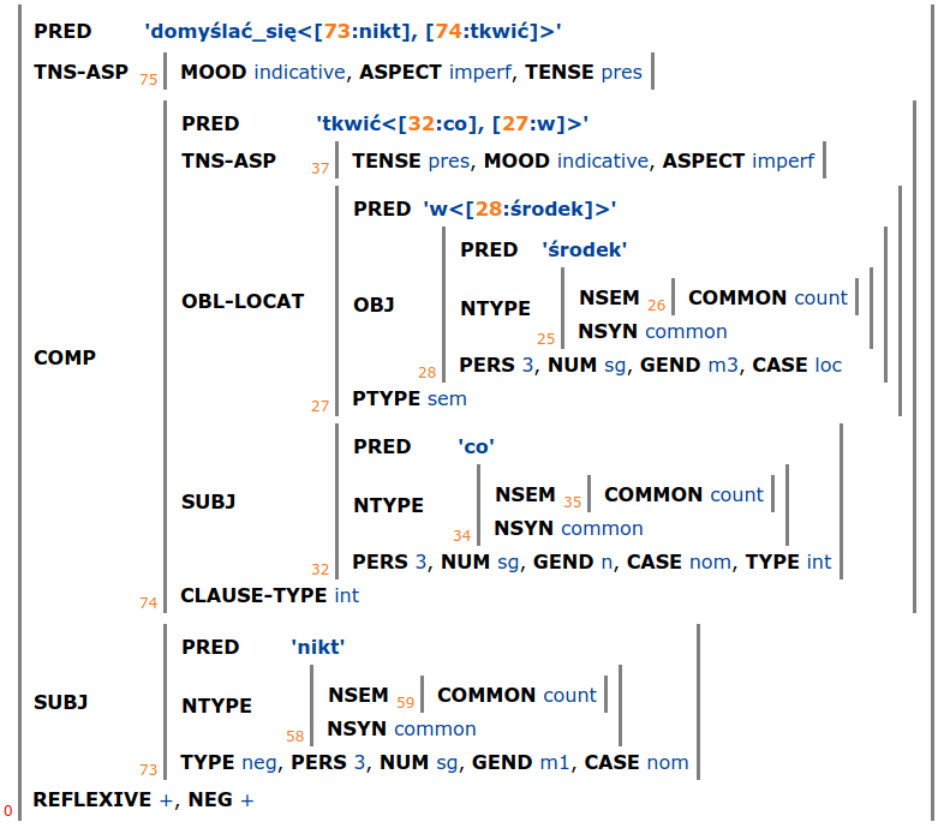


Figure 2.38: F-structure of (2.35)

Another example of a subordinate clause not introduced by a semantic complementiser is given in (2.36). There, the verb *liczyłem* '(I) counted, (I) hoped' takes an oblique dependent which contains the correlative pronoun TO 'this' (see the '+'-valued CORRELATIVE attribute in substructure 85), which in turn takes the subordinate clause *że przyjdziecie z tym do mnie* 'that you will come with this to me' as the closed clausal complement, 86, where *że* 'that' is a non-semantic complementiser. The f-structure in Figure 2.39 shows that the predicate LICZYĆ 'count, hope', 0, contains an OBL attribute, 85, filled by the predicate TO 'this', whose COMP attribute, 86, is filled by the predicate PRZYJŚĆ 'come', which contains the COMP-FORM attribute contributed by the complementiser.

(2.36) Liczyłem na to, że przyjdziecie z tym do mnie.
counted.1SG.M on this that come.2PL with this to me
'I hoped that you will come with this matter to me.'

On the other hand, the verb *zatrzymaj się* 'stop' in (2.37) has an adjunct which is the subordinate clause *bo strzelę* 'or (I will) shoot', where *bo* 'because' is the semantic complementiser. The f-structure in Figure 2.40 shows that the predicate ZATRZYMAĆ_SIĘ 'stop', 0, has an ADJUNCT attribute, 1, whose value contains the predicate BO 'because', 2, whose COMP attribute, 3, is filled by the predicate STRZELIĆ 'shoot'.

(2.37) - Zatrzymaj się, bo strzelę!
stop.2SG.IMP INH because shoot.1SG
'— Stop, or I will shoot!'

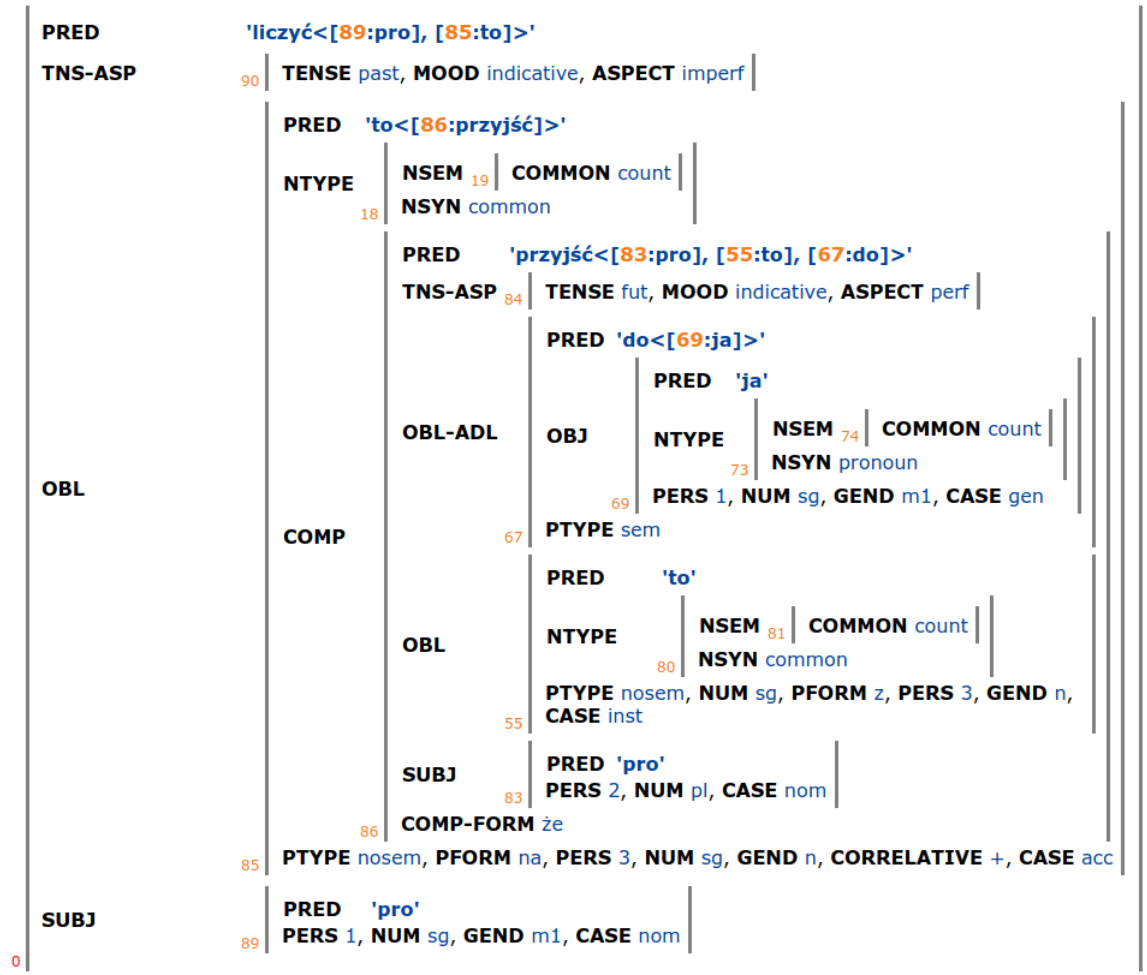


Figure 2.39: F-structure of (2.36)

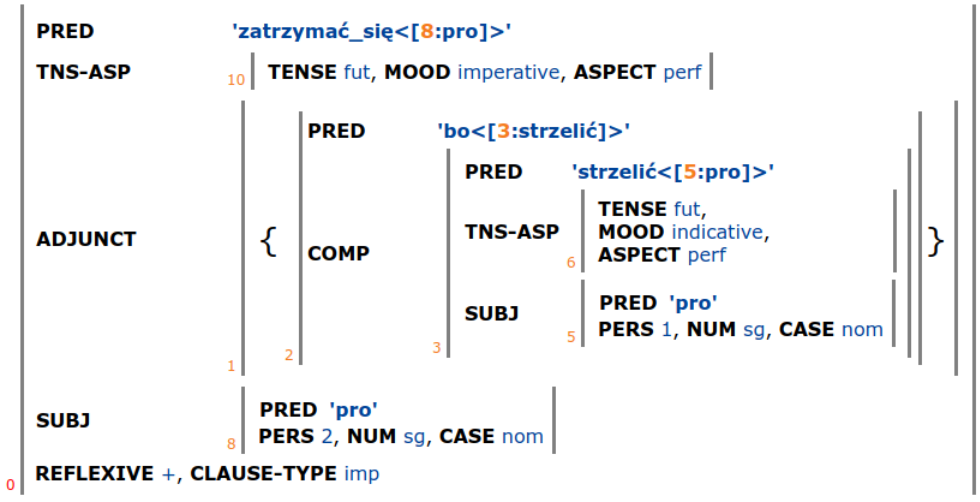


Figure 2.40: F-structure of (2.37)

The adverb *tak* 'so, in such a way' in (2.38) takes the subordinate clause *że prawie przestał oddychać* 'that (he) almost stopped breathing' as the closed clausal complement, which features the non-semantic complementiser *że* 'that'. The f-structure in Figure 2.41 shows that the predicate TAK 'so', 2, has a COMP attribute, 4, filled by the predicate PRZESTAĆ 'stop' which contains the COMP-FORM attribute contributed by the complementiser.

(2.38) Bogumił zamilkł tak, że prawie przestał oddychać.
Bogumił.NOM.SG.M fell silent.3SG.M so that almost stopped.3SG.M breathe.INF
'Bogumił fell so silent that he almost stopped breathing.'

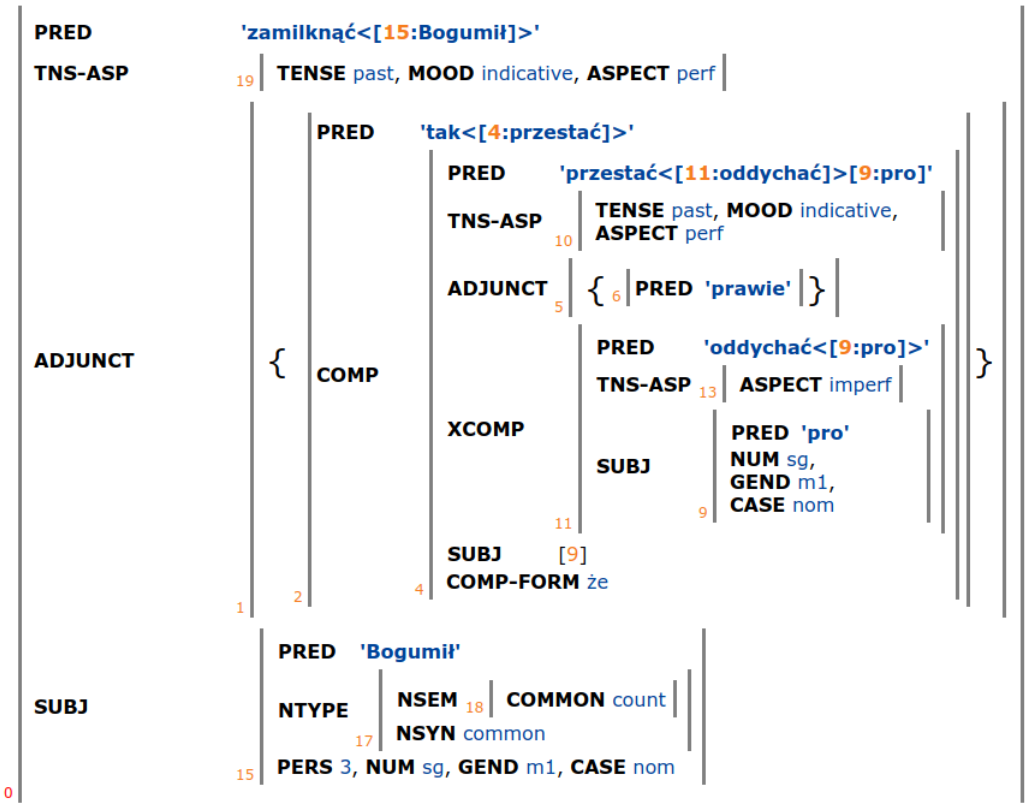


Figure 2.41: F-structure of (2.38)

Finally, as shown below, sentences may start with a complementiser. In these cases, if the complementiser is semantic, it acts as the head; if the complementiser is not semantic, the main verb of the subordinate clause acts as the main predicate.

In (2.39) the subordinate clause introduced by the non-semantic complementiser *że* 'that' acts as the main clause, because the verb taking it is a complement is not present in the sentence. The f-structure in Figure 2.42 shows that the main predicate is POTRAFIĆ 'know how, be able to', 0, and that it contains the COMP-FORM attribute contributed by the complementiser.

(2.39) Że nie potrafi kupić żadnego piłkarza do Legii.
that NEG can.3SG buy.INF no.GEN.SG.M footballer.GEN.SG.M to Legia
'That he is not able to buy any footballer for Legia.'

In (2.40) the subordinate clause introduced by the semantic complementiser *bo* 'because' acts as the main clause, because the verb which it modifies is not present in the sentence. The f-structure in Figure 2.43 shows that the main predicate is BO 'because', 0, whose COMP attribute, 1, is filled by the predicate MÓC 'may, be able to'.

(2.40) Bo teraz mogę swoją ideę realizować wszędzie.
because now can.1SG own.ACC.SG.F idea.ACC.SG.F realise.INF everywhere
'Because now I can realise my idea everywhere.'

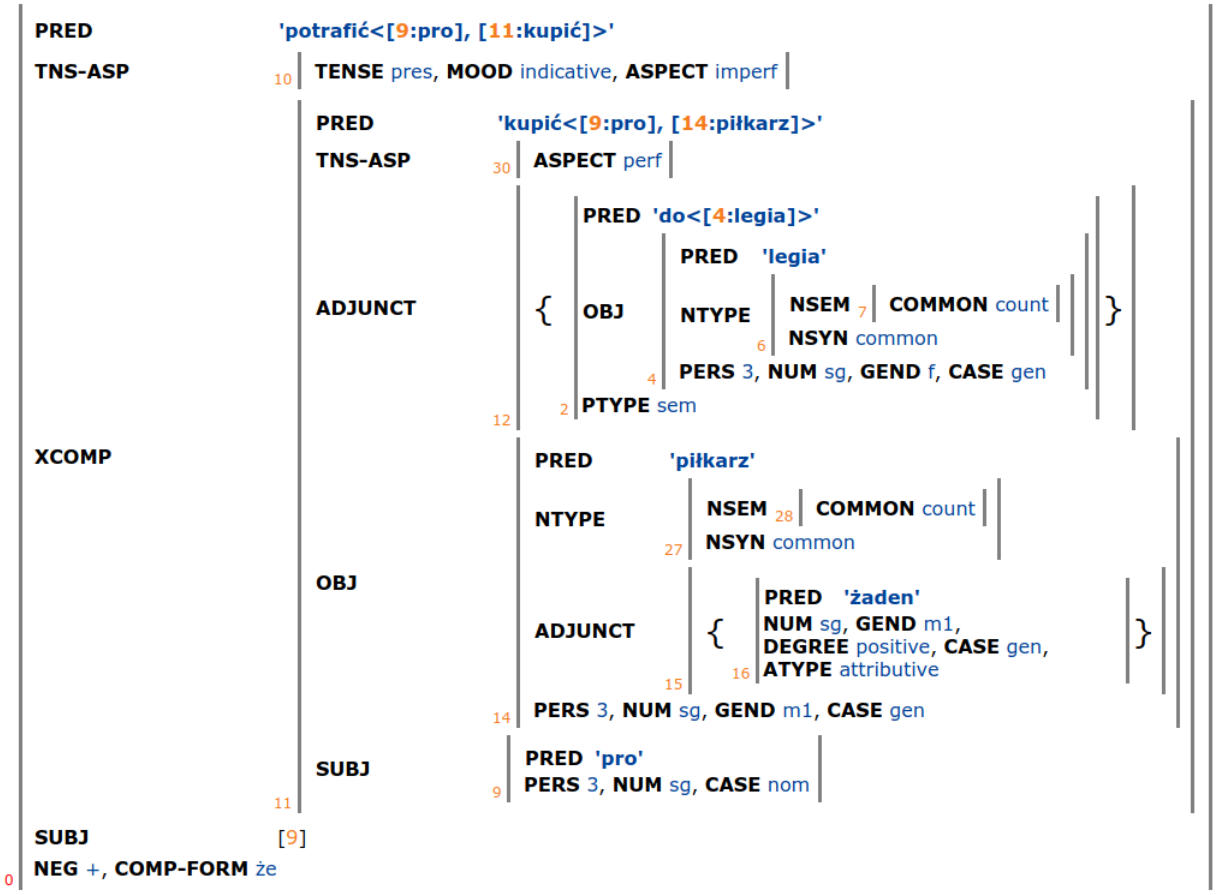


Figure 2.42: F-structure of (2.39)

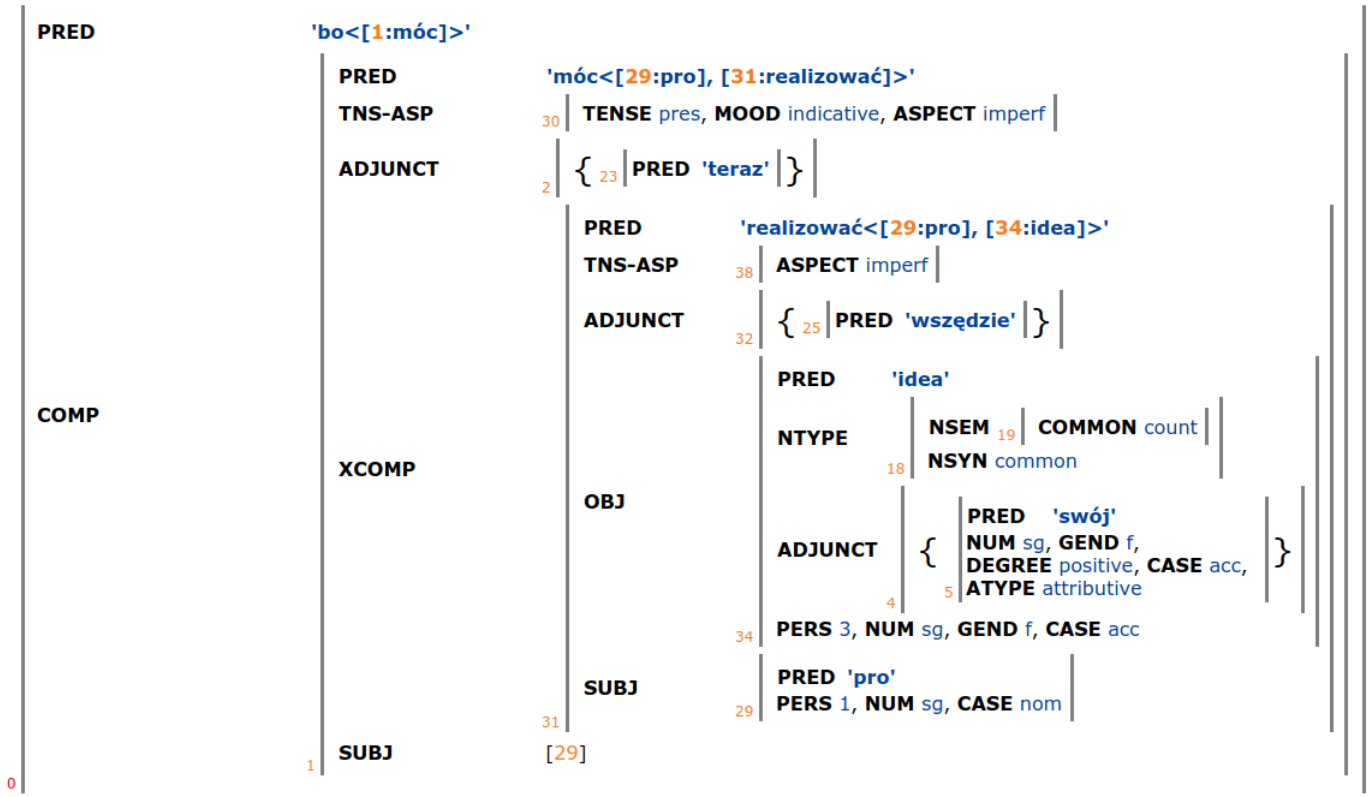


Figure 2.43: F-structure of (2.40)

2.12 Open (controlled) clausal complement (XCOMP)

The verb *chce* 'wants' in (2.41) takes the infinitival *połknąć tajemniczą kartkę* 'swallow (the) mysterious sheet' as the open clausal complement. The f-structure in Figure 2.44 shows that the predicate CHCIEĆ 'want', 0, contains an XCOMP attribute, 28, filled by the predicate POŁKNĄĆ 'swallow'. Since CHCIEĆ is a subject control verb, the subject of CHCIEĆ, filled by the predicate OSIEŁEK (a proper name), 23, is the same as the SUBJ of POŁKNĄĆ.

(2.41) Osiełek chce połknąć tajemniczą kartkę.
Osiełek.NOM.SG.M wants.3SG swallow.INF mysterious.ACC.SG.F sheet.ACC.SG.F
'Osiełek wants to swallow the mysterious sheet of paper.'

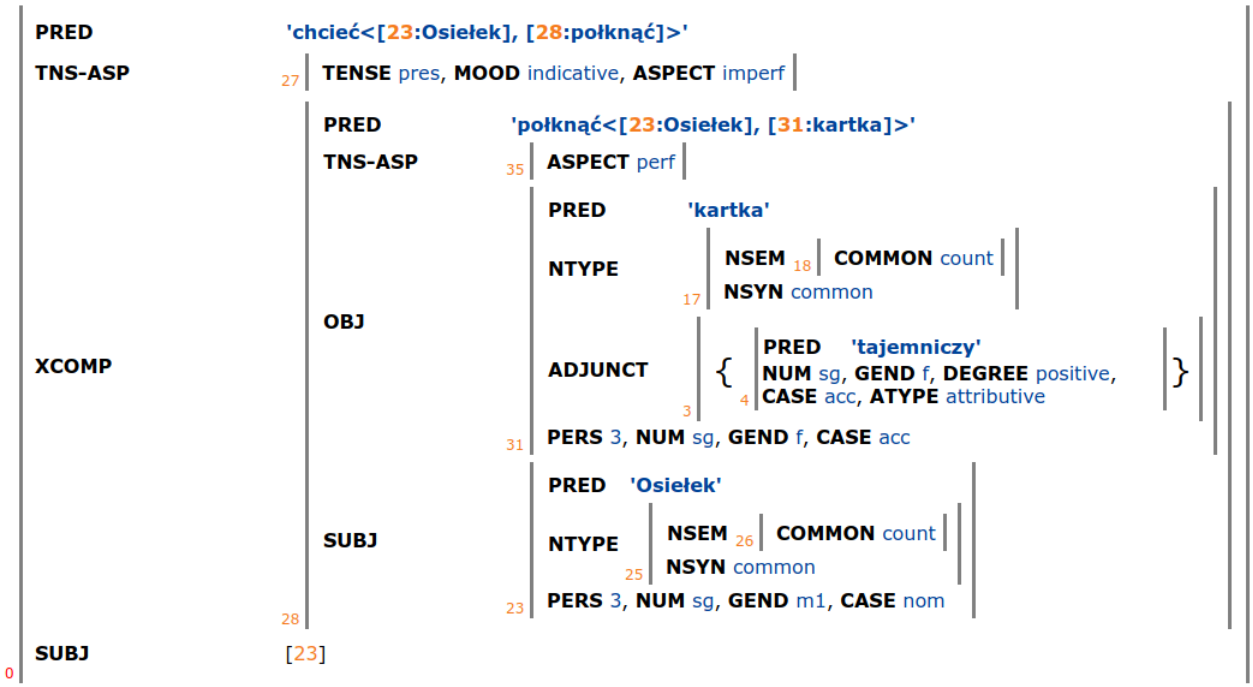


Figure 2.44: F-structure of (2.41)

The verb *kazali* 'ordered, asked' in (2.42) takes the infinitival *odejść* 'leave, go away' as the open clausal complement. The f-structure in Figure 2.45 shows that the predicate KAZAĆ 'order, ask', 1, contains an XCOMP attribute, 21, filled by the predicate ODEJŚĆ 'leave, go away'. Since KAZAĆ is an object control verb, the dative indirect object of KAZAĆ, filled by the predicate ON 'he', 17, is the same as the SUBJ of ODEJŚĆ.

(2.42) Ale później kazali mu odejść.
but later ordered.3PL.M he.DAT.SG.M leave.INF
'But later they ordered him to leave.'

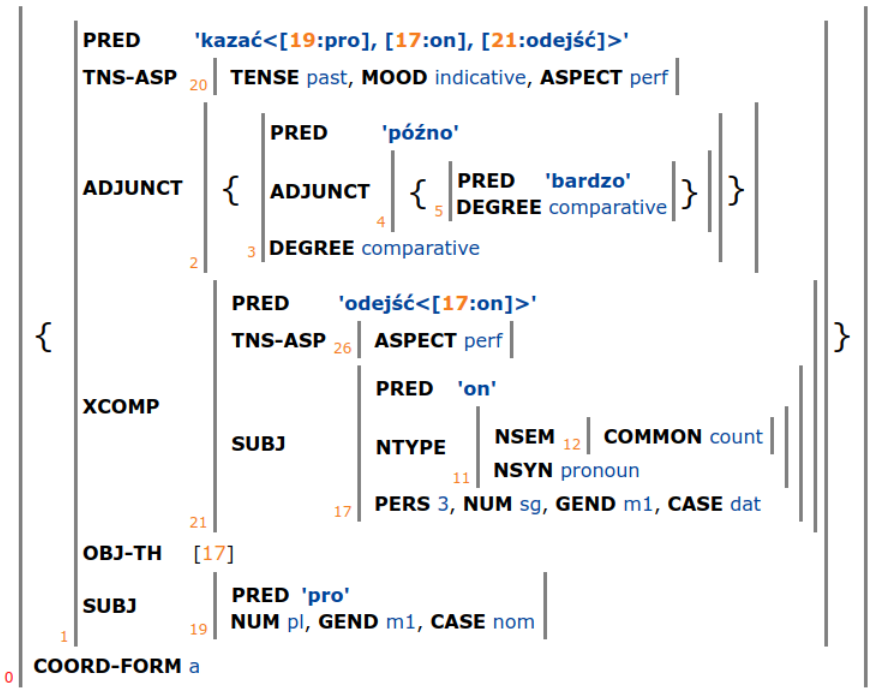


Figure 2.45: F-structure of (2.42)

2.13 Open (controlled) predicative complement (XCOMP-PRED)

The verb *staje się* 'becomes' in (2.43) takes the nominal phrase *wyleniałym tygrysem* 'shabby tiger' as the open predicative complement. The f-structure in Figure 2.46 shows that the predicate STAWAĆ_SIĘ 'become', 0, contains an XCOMP-PRED attribute, 22, filled by the predicate TYGRYS 'tiger'. Since the predicative complement of the lexeme STAWAĆ SIĘ applies to its subject, the subject of the predicate STAWAĆ_SIĘ, filled by the predicate CZŁOWIEK 'human', 17, is the same as the SUBJ of TYGRYS.

(2.43) Człowiek staje się wyleniałym tygrysem.
human.NOM.SG.M becomes.3SG INH shabby.INS.SG.M tiger.INS.SG.M
'One becomes a shabby tiger.'

The verb *wyglądać* 'look, seem' in (2.44), whose subject is *pro*-dropped, takes the prepositional phrase *na niezgorszego gwałciciela* 'for quite a rapist' as the open predicative complement featuring the non-semantic preposition *na* 'on, for'. The f-structure in Figure 2.47 shows that the predicate WYGLĄDAĆ 'look, seem', 40, contains an XCOMP-PRED attribute, 47, filled by the predicate GWAŁCICIEL 'rapist', which bears accusative case, as required by the non-semantic preposition NA 'on, for', which contributes its PFORM. Since the predicative complement of WYGLĄDAĆ applies to its subject, the implicit subject of WYGLĄDAĆ, filled by the predicate PRO, 38, is the same as the SUBJ of GWAŁCICIEL.

(2.44) Musiałem wyglądać na niezgorszego gwałciciela.
must.1SG.M look.INF for not bad.ACC.SG.M rapist.ACC.SG.M
'I must have looked like quite a rapist.'

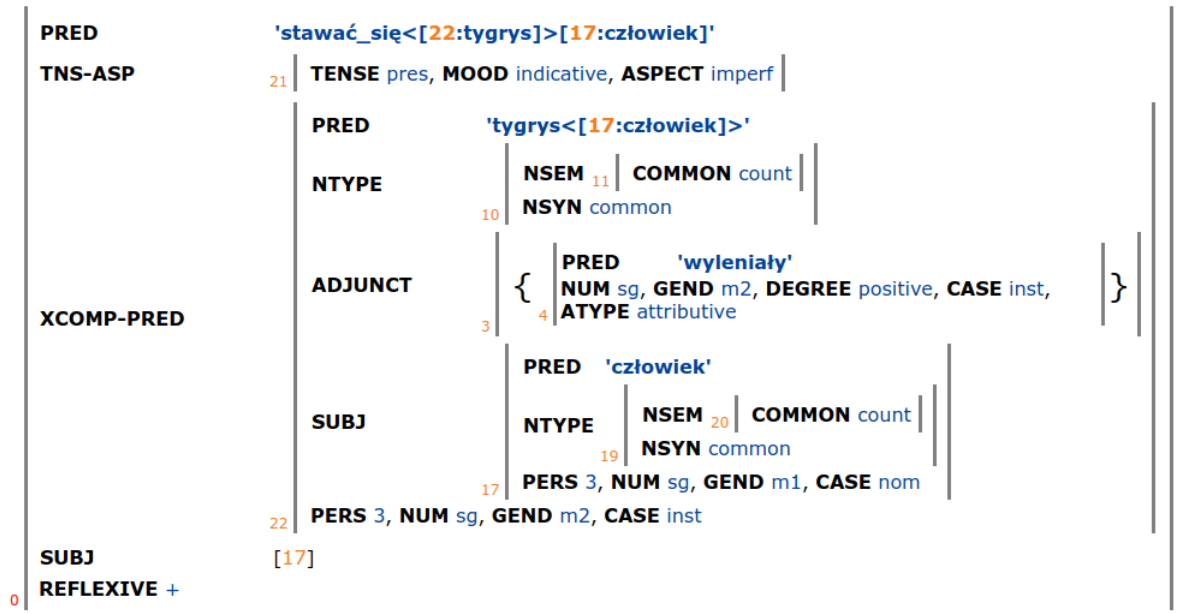


Figure 2.46: F-structure of (2.43)

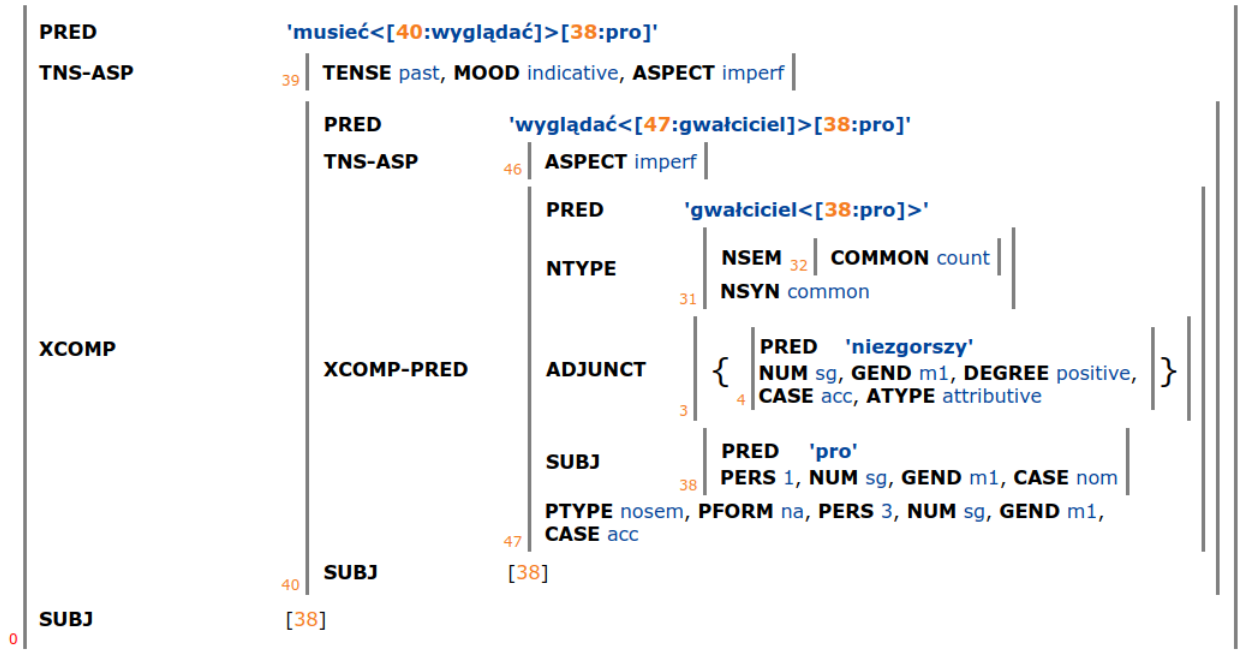


Figure 2.47: F-structure of (2.44)

The verb *czuję się* 'feel' in (2.45) takes the adjectival phrase *przegrany* 'defeated' as the open predicative complement. The f-structure in Figure 2.48 shows that the predicate CZUĆ_SIĘ 'feel', 0, contains an XCOMP-PRED attribute, 20, filled by the predicate PRZEGRANY 'defeated'. Since the predicative complement of the lexeme CZUĆ SIĘ applies to its subject, the subject of CZUĆ_SIĘ, filled by the predicate JA 'I', 3, is the same as the SUBJ of PRZEGRANY.

(2.45) Ja się nie czuję przegrany.
I.NOM.SG.M INH NEG feel.1SG defeated.NOM.SG.M
'I do not feel defeated.'

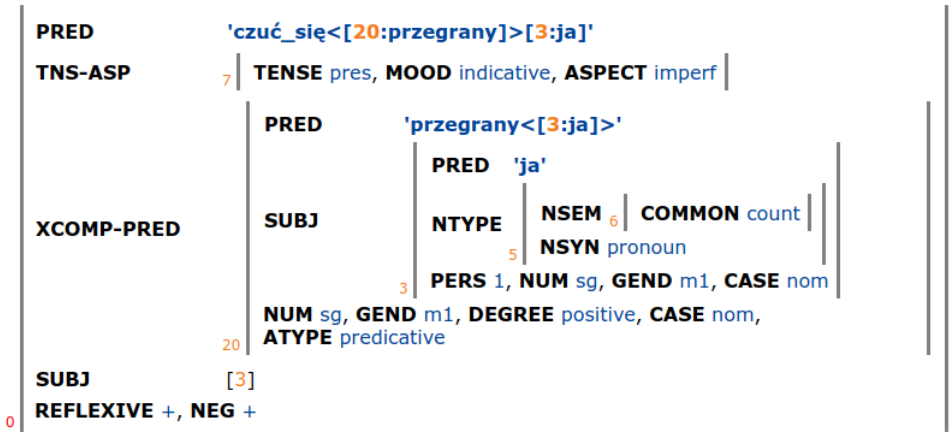


Figure 2.48: F-structure of (2.45)

Finally, the verb *uznał* 'considered, judged' in (2.46) takes the prepositional adjectival phrase *za mało taktowny* 'for not very tactful' as the open predicative complement featuring the non-semantic preposition *za* 'for, as'. The f-structure in Figure 2.49 shows that the predicate UZNAĆ 'consider, judge', 0, contains an XCOMP-PRED attribute, 41, filled by the predicate TAKTOWNY 'tactful', which bears accusative case, as required by the non-semantic preposition ZA 'for, as' which contributes its PFORM. Since the predicative complement of UZNAĆ applies to its object, the object of UZNAĆ, filled by the predicate PODPIS 'caption, signature', 34, is the same as the SUBJ of TAKTOWNY.

(2.46) Uznał ten podpis za mało taktowny.
considered.3SG.M this.ACC.SG.M caption.ACC.SG.M for little tactful.ACC.SG.M
'He considered this caption to have little tact.'

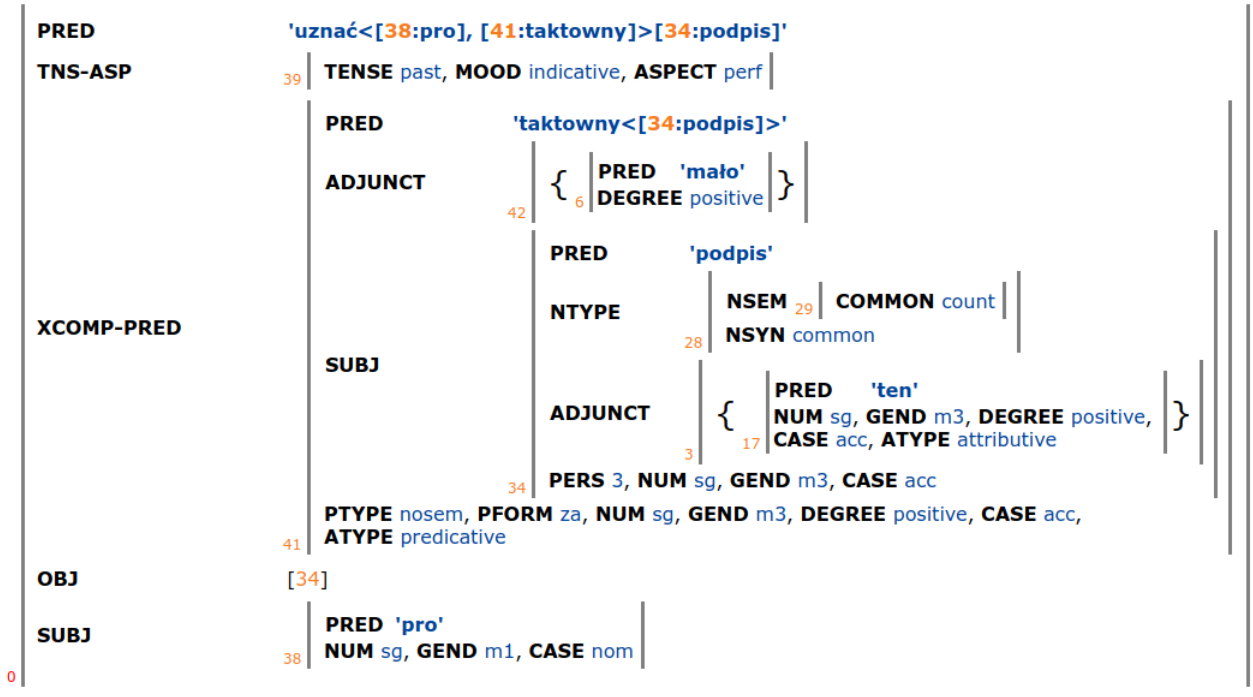


Figure 2.49: F-structure of (2.46)

2.14 Closed adjunct (ADJUNCT)

The ADJUNCT attribute occurs in many f-structures above, but let us illustrate it here with example (2.47), which includes two closed adjuncts: the adverb *nigdy* 'never' modifying the main verb *uważał* 'considered' and the adjective *tych* 'these' modifying the noun *kolegów* 'colleagues'. Note that the f-structure in Figure 2.50 contains three instances of ADJUNCT: the predicate UWAŻAĆ 'consider', 0, has an ADJUNCT attribute, 1, containing the predicate NIGDY 'never', 120; the predicate KOLEGA 'colleague', 170, has an ADJUNCT attribute, 27, containing the predicate TEN 'this', 28; finally, the predicate DOBRY 'good', 135, has an ADJUNCT attribute, 136, containing the predicate BARDZO 'very', 154 – even though there is no word *bardzo* 'very' in (2.47), the synthetic comparative degree is represented in the same way as analytic degree formed using BARDZO (cf. (2.26) on page 42 and Figure 2.29 there).

(2.47) Nie uważał siebie nigdy za lepszego od obu tych kolegów.
NEG considered.3SG.M self.GEN never for better.ACC.SG.M from both these colleagues
'He never considered himself to be better than both these colleagues.'

2.15 Open (controlled) adjunct (XADJUNCT)

Apart from typical adjuncts illustrated above, LFG distinguishes a class of open, controlled adjuncts: it includes secondary predicates (controlled by various dependents) and adverbial participles (controlled by the main clause subject).

The adjective *zmęczeni* 'tired' in (2.48) is a predicative open modifier of the main verb *wysiadaliśmy* '(we) were getting out' controlled by its subject. The f-structure in Figure 2.51 shows that the predicate WYSIADAĆ 'get out', 0, has an XADJUNCT attribute, 31, containing the predicate ZMĘCZONY 'tired', 32. Since the predicative open modifier of WYSIADAĆ applies to its subject, the implicit subject of WYSIADAĆ, filled by the predicate PRO, 26, is the same as the SUBJ of ZMĘCZONY.

(2.48) Zmęczeni wysiadaliśmy z ciasnej szoferki.
tired.NOM.PL.M get out.1PL.M from cramped.GEN.SG.F driver's cab.GEN.SG.F
'Tired, we were getting out of the cramped driver's cab.'

The contemporary adverbial participle *patrząc* 'looking' in (2.49) is an open modifier of the main verb *odpływamy* '(we) sail away' controlled by its subject. The f-structure in Figure 2.52 shows that the predicate ODPŁYWAĆ 'sail away', 0, has an XADJUNCT attribute, 46, containing the predicate PATRZEĆ 'look', 47. Since the adverbial participle is controlled by the subject, the implicit (*pro*-dropped) subject of ODPŁYWAĆ, filled by the predicate PRO, 44, is the same as the SUBJ of PATRZEĆ.

(2.49) Odpływamy patrząc na skaliste brzegi.
sail.1PL looking at rocky.ACC.PL.M shore.ACC.PL.M
'We sail away looking at the rocky shores.'

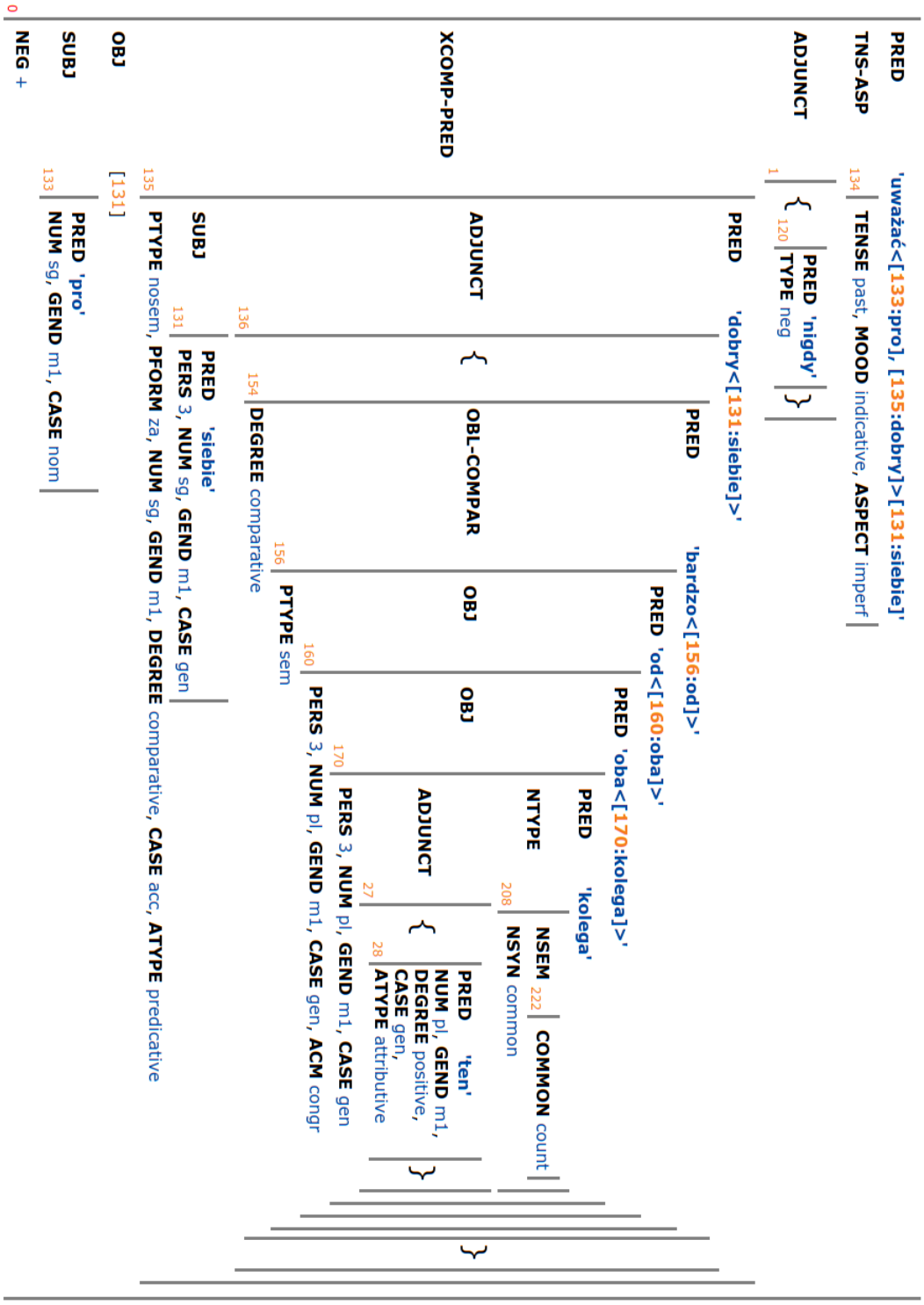


Figure 2.50: F-structure of (2.47)

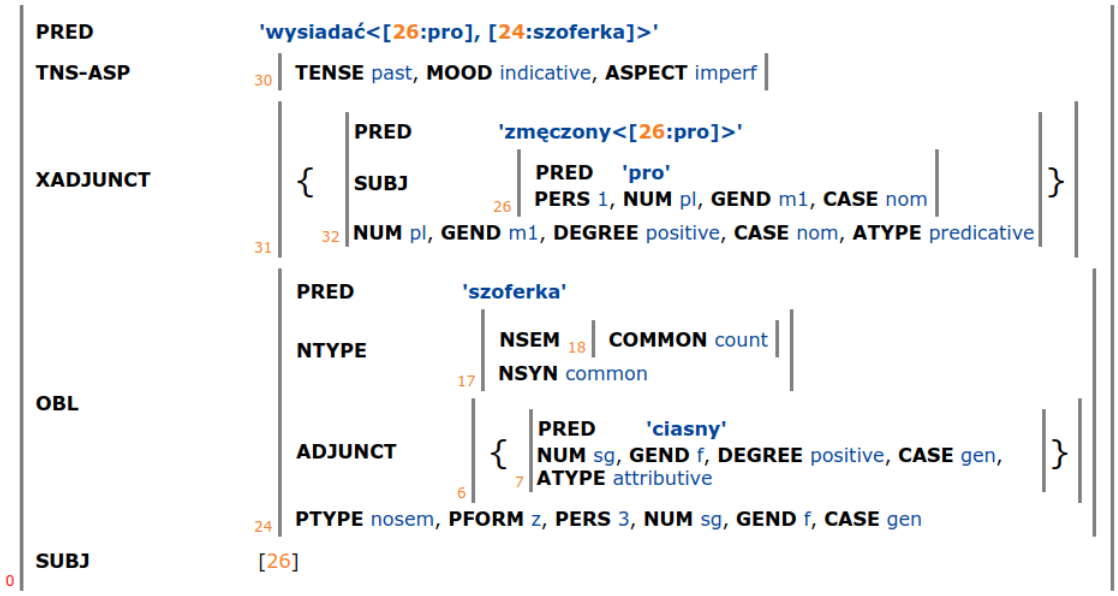


Figure 2.51: F-structure of (2.48)

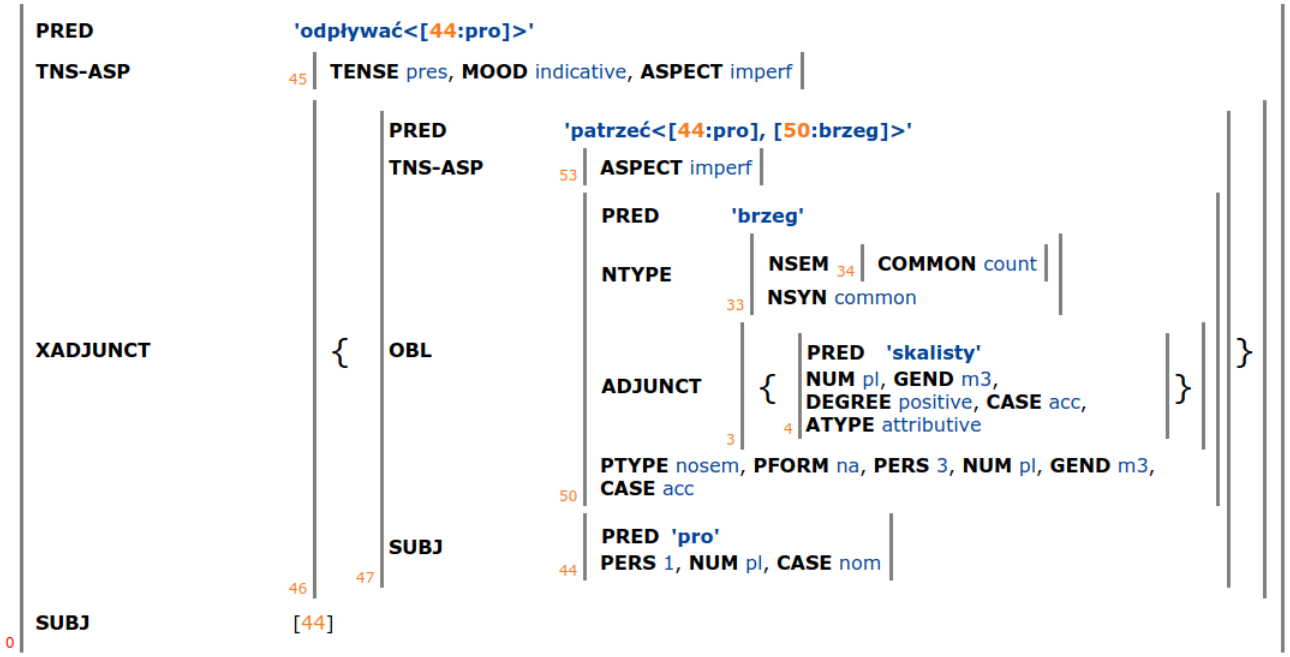


Figure 2.52: F-structure of (2.49)

The anterior adverbial participle *zmarszczywszy* 'having frowned' in (2.50) is an open modifier of the main verb *zastanawiał się* 'pondered' controlled by its subject. The f-structure in Figure 2.53 shows that the predicate ZASTANAWIAĆ_SIĘ 'ponder', 0, has an XADJUNCT attribute, 18, containing the predicate ZMARSZCZYĆ 'frown', 19. Since the adverbial participle is controlled by the subject, the implicit (*pro*-dropped, again) subject of ZASTANAWIAĆ_SIĘ, filled by the predicate PRO, 16, is the same as the SUBJ of ZMARSZCZYĆ.

(2.50) Zmarszczywszy brwi zastanawiał się chwilę.
frowned eyebrow.ACC.PL.F pondered.3SG.M INH moment.ACC.SG.F
'Having frowned, he pondered for a while.'

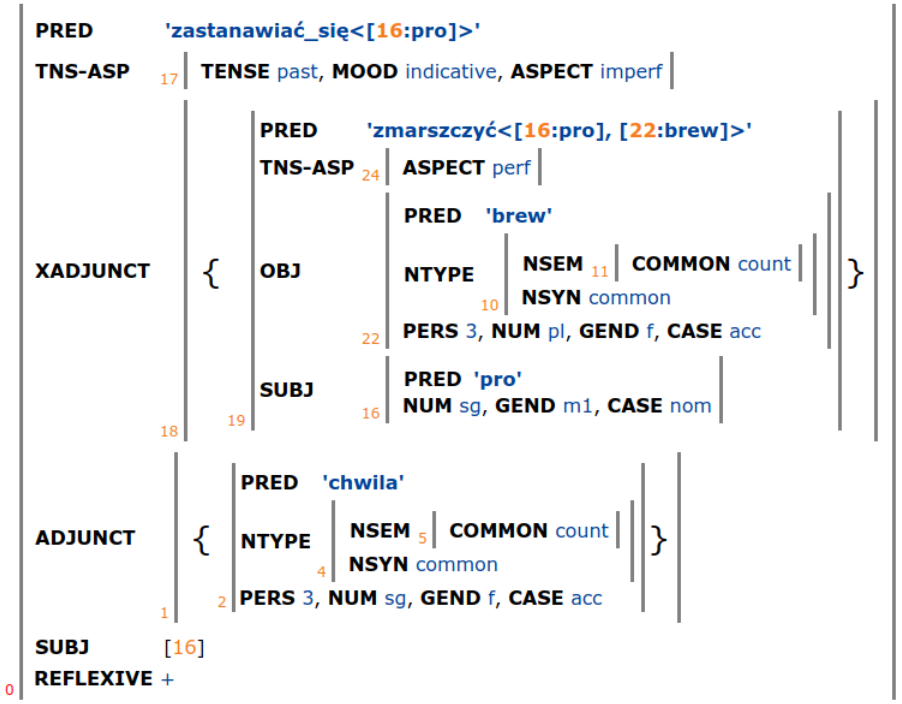


Figure 2.53: F-structure of (2.50)

2.16 Possessive dependent (POSS)

The noun *Warszawa* 'Warsaw' in (2.51) has the pronoun *ich* 'their' as its genitive possessive dependent. The f-structure in Figure 2.54 shows that the predicate WARSZAWA, 24, contains a POSS attribute, 28, filled by the predicate ON 'he' marked for genitive case.

(2.51) W podwórzu stała ich Warszawa.
in backyard.LOC.SG.N stood.3SG.F they.GEN.PL.M Warszawa.NOM.SG.F
'Their Warszawa (car) stood in the backyard.'

2.17 Appositive dependent (APP)

The noun *Chilijczyk* 'Chilean' in (2.52) is followed by two more (proper) nouns: *Ariel* and *Dorfman*, forming an appositive construction. The f-structure in Figure 2.55 shows that the predicate CHILIJCZYK 'Chilean', 14, has an APP attribute, 15, filled by the predicate ARIEL, which in turn has an APP attribute, 16, filled by the predicate DORFMAN – the appositives form a chain.

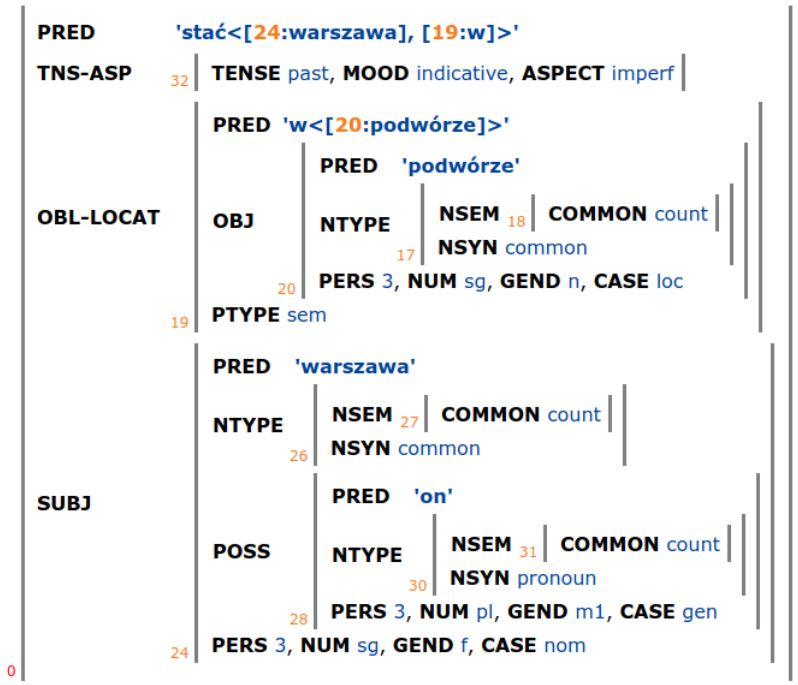


Figure 2.54: F-structure of (2.51)

(2.52) Napisał ją Chilijczyk Ariel Dorfman.
wrote.3SG.M she.ACC.SG.F Chilean.NOM.SG.M Ariel.NOM.SG.M Dorfman.NOM.SG.M
'The Chilean Ariel Dorfman wrote it (the book).'

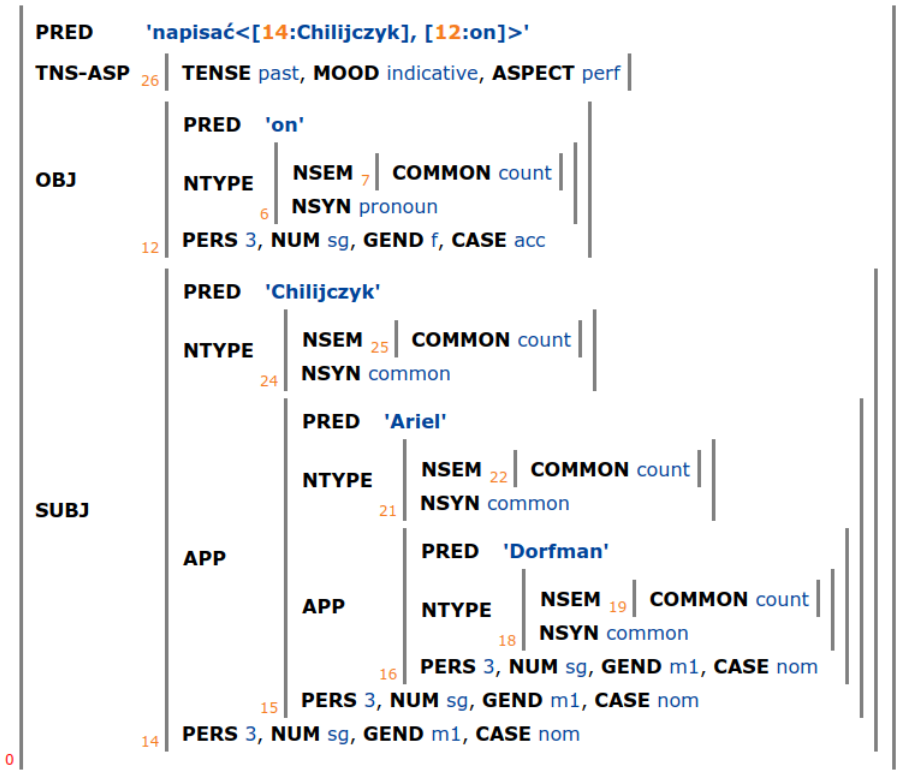


Figure 2.55: F-structure of (2.52)

Chapter 3

C-structure

As mentioned in the previous chapter, f-structure is in some ways more important of the two syntactic structures assumed in LFG. Nevertheless, it is useful to understand the intended scope of labels used to mark nonterminals in c-structures (see Section 3.1). In particular, the procedure – described in Part II of this monograph – of converting the LFG structure bank to the Universal Dependencies standard relies to a large extent on preterminal labels. Such preterminal labels often correspond directly to morphosyntactic classes of the tagset of the National Corpus of Polish (see Appendix A); for example, the preterminal `DEPR` (introduced in Section 3.1.5) is named after – and has the same scope – as the `depr` class in the NKJP tagset.

Apart from presenting the repertoire of nonterminal labels, another aim of this chapter is to discuss two issues related to the mapping between c-structures and f-structures. One is concerned with so-called co-heads, i.e., with nonterminal c-structure nodes mapped to the same functional substructure (see Section 3.2). The other is concerned with a certain type of discrepancies between c-structures and f-structures, namely, when a dependent of a predicate occurs outside the immediate vicinity of this predicate in the c-structure, but should still be represented locally to this predicate in the f-structure (see Section 3.3).

Again, c-structures are displayed here as screenshots from the INESS system. For example, in the case of sentence (2.20), repeated below, whose f-structure is repeated from the previous chapter in Figure 3.1, the c-structure is displayed as Figure 3.2.

(2.20) Nie mają wyboru.
NEG have.3PL choice.GEN.SG.M
'They have no choice.'

Note the two kinds of edges between c-structure nonterminals: solid and dashed. (The edges between preterminals and terminals – i.e., text tokens – are always solid.) Solid lines mean that the two nonterminals map to the same functional structure. For example, the node with label `FIN` and its mother `IP` map to the same f-structure, namely, the one bearing index 0 in Figure 3.1. By the same token, also the nodes `NEG`, `S`, `ROOT` and `PERIOD` map to the same f-structure: they are all connected with solid edges. This in particular means that the preterminals `NEG` and `FIN` are co-heads of the `IP`. On the other hand, nodes `NP`, `N` and `SUBST` map to a different functional structure (there is no solid path between them and the other nodes), namely, to the one with index 2.

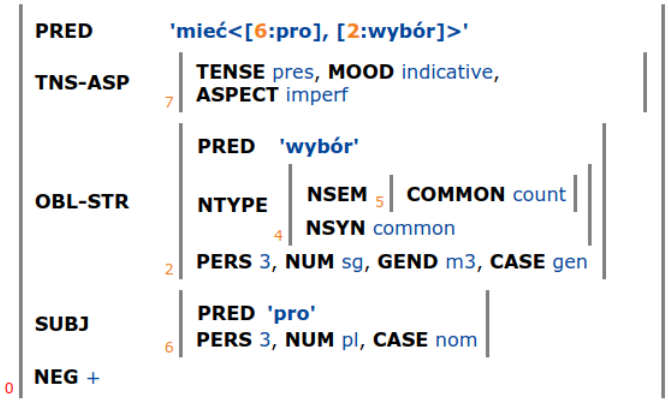


Figure 3.1: F-structure of (2.20)

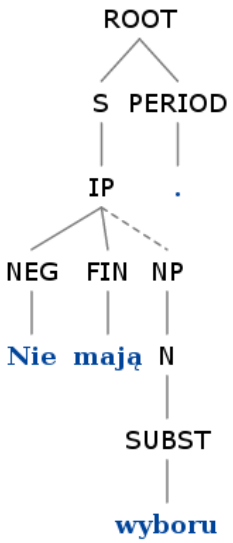


Figure 3.2: C-structure of (2.20)

3.1 Category breakdown

3.1.1 ROOT, HEADER and punctuation

The root of each constituency tree corresponding to a sentence has the label ROOT – such trees are a majority in the structure bank. There is also a recent addition, namely constituency trees whose root does not correspond to a sentence and so it has the label HEADER. Since such utterances are not included in the converted dependency structures, we will not discuss HEADER below. Apart from the more linguistically motivated nonterminal labels presented in the ensuing sections, there are also various labels for punctuation marks, including the following ones occurring in c-structures displayed in this chapter (see Section 3.2.2 for a more complete list):

- PERIOD: period (.)
- COMMA: comma (,)
- DASH: various dashes
- EXCL-POINT: exclamation mark (!)
- INT-MARK: question mark (?)

3.1.2 Sentences and subordinate clauses

As mentioned above, mostly sentences, i.e., utterances with a verbal head, are represented in the LFG structure bank of Polish.

Every `ROOT` node corresponds to a sentence, i.e., an utterance with a verbal head, hence it has an immediate constituent of one of the following two kinds:

- `S`: sentence built around a verbal predicate (not necessarily finite)
- `S-INK`: as `S`, but with an incorporating conjunction

All c-structures displayed in this chapter involve `S` constituents immediately dominated by `ROOT`. Incorporating conjunctions are a recent addition to the LFG grammar and the structure bank, there are very few sentences illustrating this phenomenon in the corpus, so we will ignore `S-INK` here; see Patejuk 2018 for details.

`S` constituents are also parts of subordinate clauses. The topmost labels of such clauses are:

- `CP[int]`: interrogative subordinate clause together with the surrounding punctuation; cf., e.g., Figure 3.4
- `CP[rel]`: relative subordinate clause together with the surrounding punctuation (see also `CPres` below for a special type of relative clauses); cf., e.g., Figure 3.5
- `CP[sub]`: subordinate clause introduced by a complementiser (hence, CP = complementiser phrase), together with the surrounding punctuation; cf., e.g., Figure 3.3, where the `CP[sub]` is a dependent of a finite verb, Figure 3.7, where it is a dependent of the correlative pronoun *to* 'this', Figure 3.8, where it is a dependent of the adverb *tak* 'so, in such a way', and Figure 3.9, where – unlike the non-semantic complementiser *że* 'that' in the other figures – the complementiser *bo* 'because, or else' is meaningful (it is 'semantic')

According to Polish punctuation rules, subordinate clauses are surrounded by commas. When such a comma coincides with sentence boundary or with another punctuation mark, e.g., sentence-final punctuation, it is omitted. For example, in the case of (2.34), repeated below, only the comma introducing the subordinate clause *że pił alkohol* 'that he drank alcohol' is present, as the comma ending it would have to be placed next to the period.

(2.34) Pamięta, że pił alkohol.
remembers.3SG that drank.3SG alcohol.ACC.SG.M
'He remembers that he drank alcohol.'

Nevertheless, constituency trees always contain both commas, as illustrated in Figure 3.3. Such commas are the 'frontier' daughters of `CP[…]` constituents, with another – 'median' – daughter, `CPbare[…]`, dominating the subordinate clause proper:

- `CPbare[int]`: interrogative subordinate clause without the surrounding punctuation; dominates `XPextr[int]` (interrogative phrase) followed by `S`; cf., e.g., Figure 3.4
- `CPbare[rel]`: relative subordinate clause without the surrounding punctuation; dominates `XPextr[rel]` (relative phrase) followed by `S`; cf., e.g., Figure 3.5
- `CPbare[sub]`: subordinate clause with a complementiser, without the surrounding punctuation; dominates `COMP` (a complementiser) followed by `S`; cf., e.g., Figures 3.3 and 3.7–3.9

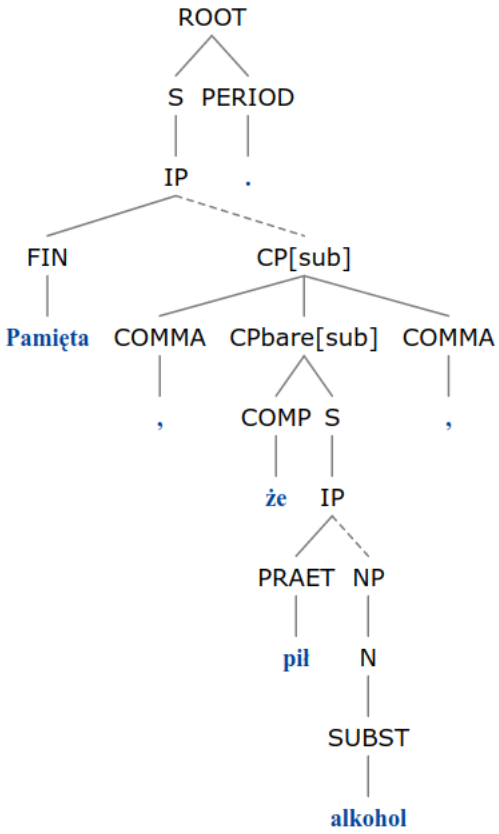


Figure 3.3: C-structure of (2.34)

As mentioned above, the constituents preceding `S` in such subordinate clauses are, respectively:

- `XPextr[int]`: topmost extracted interrogative (int) phrase; cf., e.g., Figure 3.4
- `XPextr[rel]`: topmost extracted relative (rel) phrase; cf., e.g., Figure 3.5
- `COMP`: preterminal for complementisers (`comp` in NKJP); cf., e.g., Figures 3.3 and 3.7–3.9

The two `XPextr[…]` labels are described in more detail in Sections 3.1.12 and 3.3 below. `XPextr[int]` is illustrated in Figure 3.4, corresponding to example (2.35), repeated below.

(2.35) Nikt nie domyśla się, co tkwi w środku.
nobody.NOM.SG.M NEG suspects.3SG INH what.NOM.SG.N lies.3SG in middle
'Nobody suspects what lies inside.'

As `XPextr[int]` (and similarly for `XPextr[rel]`) represents an interrogative (respectively, relative) phrase of any category, it dominates in Figure 3.4 a more specific nominal interrogative node `NP[int]`. As described in the ensuing subsections, not only nominal phrases, but also other categories are parameterised for the subtypes `int`, `rel` and – for other reasons (cf. Section 3.3) – `neg` (negative).

Similarly, `XPextr[rel]` is illustrated in Figure 3.5, corresponding to example (3.1).

(3.1) To utwór, od którego wszystko się zaczęło.
is work.NOM.SG.M from which.GEN.SG.M everything.NOM.SG.N INH started.3SG.N
'This is the piece from which everything began.'

(Note the relative prepositional phrase, `PP[rel]`, dominated by `XPextr[rel]`.)

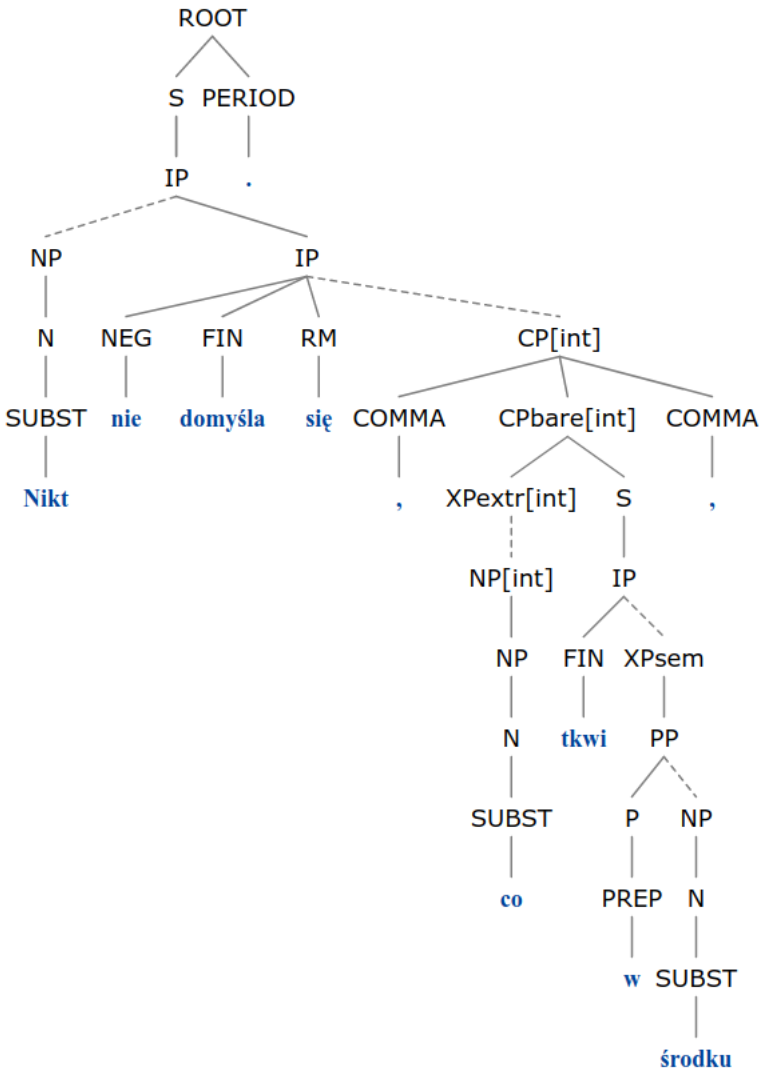


Figure 3.4: C-structure of (2.35)

Apart from typical relative clauses, marked as `CP[rel]`, there is a rather special type of relative clauses, involving the word *co* homonymous with the pronoun meaning 'what', but acting here as a kind of complementiser; see example (3.2) and the corresponding c-structure in Figure 3.6.

(3.2) - Byli tacy, co całą noc tam siedzieli i widzieli.
were.3.PL.M1 such.NOM.PL.M1 RSM all night there sat.3.PL.M1 and saw.3.PL.M1
'– There were such (people) who sat there all night and saw (this).'

Since such relative clauses may involve resumptive pronouns (the one exemplified here does not), *co* has the preterminal `RSM` (resumptive marker) and the whole clause is marked as `CPres` ('resumptive' relative clause):

- `RSM`: the word *co* introducing 'resumptive' relative clauses
- `CPres`: 'resumptive' relative clause

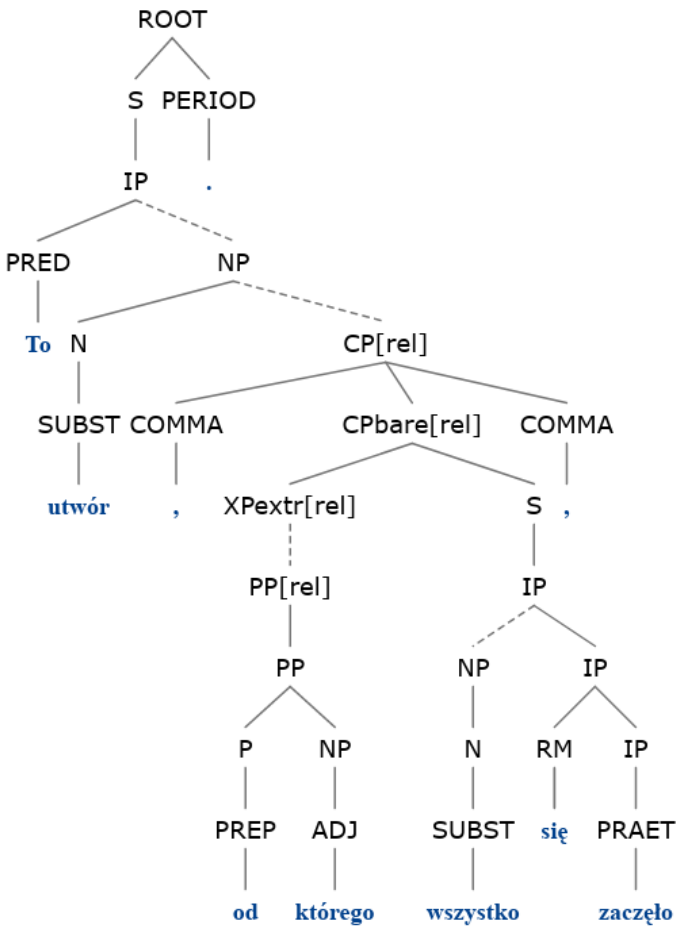


Figure 3.5: C-structure of (3.1)

3.1.3 Verbal constituents

In the typical case, illustrated by almost all c-structures in this chapter, the only daughter of S is IP, the maximal verbal projection. (See Section 3.3 for constructions in which there is also another constituent immediately dominated by S.) In some, relatively rare (and, hence, not illustrated here) constructions involving auxiliaries, another verbal projection occurs, VP, which cannot however contain negation – if such a sentence is negated, the negation must be hosted by the auxiliary outside of the VP. Only in such cases is the verbal element dominated by a V node. Otherwise, IP is usually headed by another IP, as in the case of the highest IP in Figure 3.4, or directly by a verbal preterminal whose name corresponds to the morphosyntactic class according to the NKJP tagset: FIN, as in the case of the other two IPs in Figure 3.4, PRAET, as in the case of the lowest IP in Figure 3.5, etc.:[1]

- IP: topmost verbal category, may host negation

[1]Apart from those listed below, two additional – rather ephemeral – nonterminal verbal categories occur in some constituent structures:

- ILEX: immediately dominating category for lexical verb preterminals: BEDZIE, FIN, IMPS, IMPT, INF, PRAET, PRED, WINIEN
- IAUX: immediately dominating category for the auxiliary preterminal: AUX

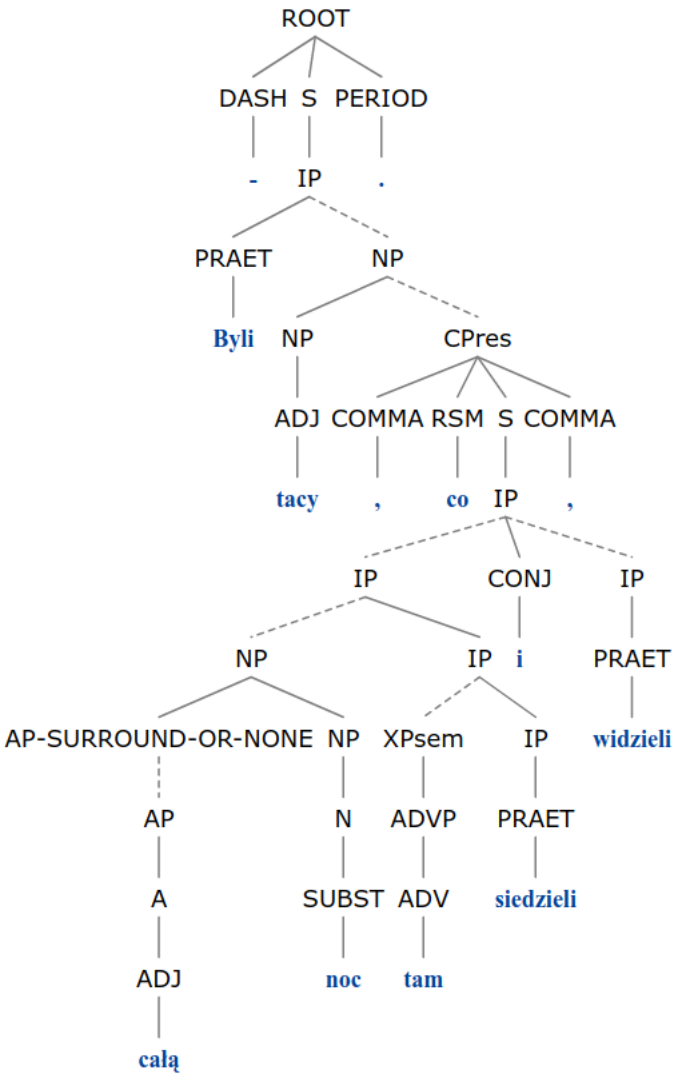


Figure 3.6: C-structure of (3.2)

- VP: topmost verbal category in some constructions involving auxiliary verbs, cannot host negation
- V: immediately dominating category for PRAET, INF and PRED preterminals within a VP
- FIN: preterminal for a lexical verb, a non-past finite form (fin in NKJP); cf., e.g., Figures 3.2, 3.3 (the higher IP) or 3.4 (both lower IPs)
- PRAET: preterminal for a lexical verb, a past form (praet in NKJP); cf., e.g., Figure 3.3 (the lower IP)
- INF: preterminal for a lexical verb, an infinitival form (inf in NKJP); cf., e.g., Figure 3.8 on page 69 (the lowest IP) and Figure 3.22 on page 82 (the lowest IP)
- IMPS: preterminal for a lexical verb, an impersonal *-no*/*-to* form (imps in NKJP); cf., e.g., Figure 3.17 on page 76
- IMPT: preterminal for a lexical verb, an imperative form (impt in NKJP); cf., e.g, Figure 3.9 on page 69 (the higher IP)
- PRED: preterminal for a 'quasi-verb' not inflecting for person, including the predicative copula TO (pred in NKJP); cf., e.g., Figure 3.5 (the highest IP)
- WINIEN: preterminal for a form of a small class of verbs of the WINIEN 'should' type (winien in NKJP)

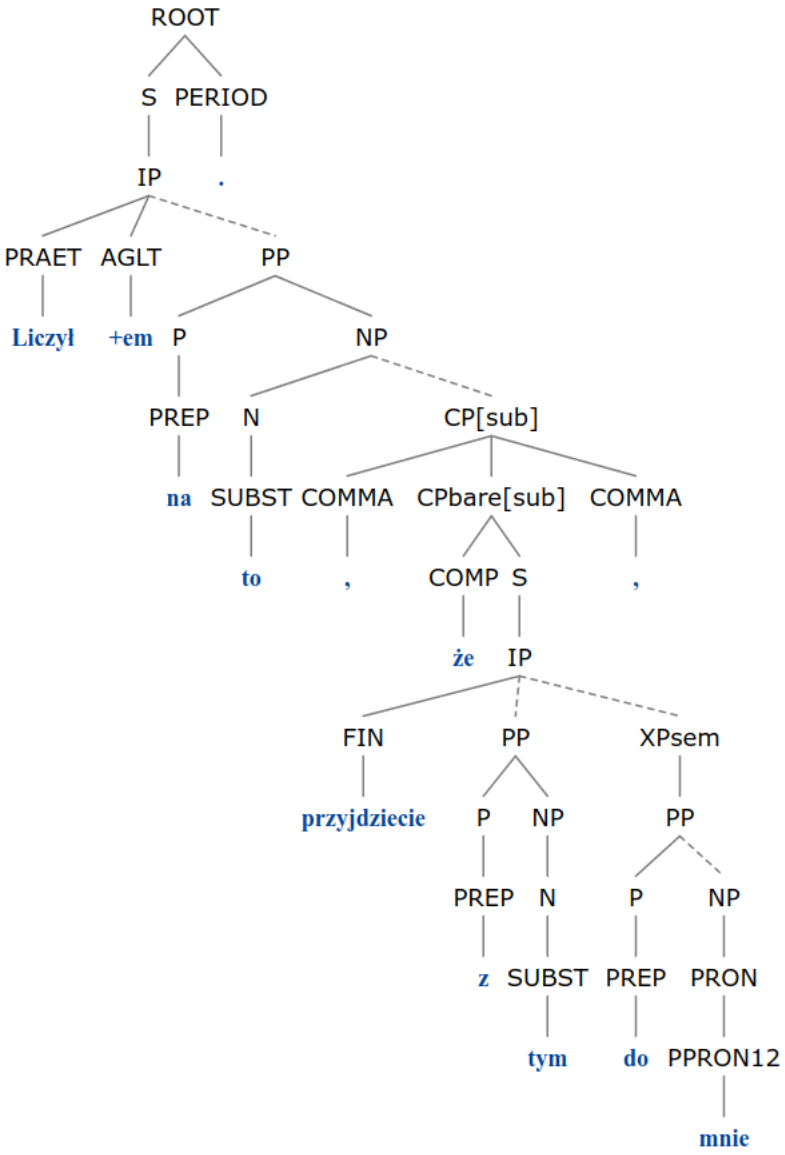


Figure 3.7: C-structure of (2.36)

- `BEDZIE`: preterminal for a future form of BYĆ 'be' (`bedzie` in NKJP)
- `AUX`: preterminal for an auxiliary (form of BYĆ)

3.1.4 Mobile inflection and markers

Example (2.36), repeated below, involves a verbal form, *liczyłem* '(I) counted', which consists of two segments: the masculine past form *liczył* 'counted' – which on its own may express third person (hence '(3)' in the glosses) – and the so-called mobile inflection, *em*, expressing first person and singular number.

(2.36) Liczyłem na to, że przyjdziecie z tym do mnie.
counted.(3)SG.M-1SG on this that come.2PL with this to me
'I hoped that you will come with this matter to me.'

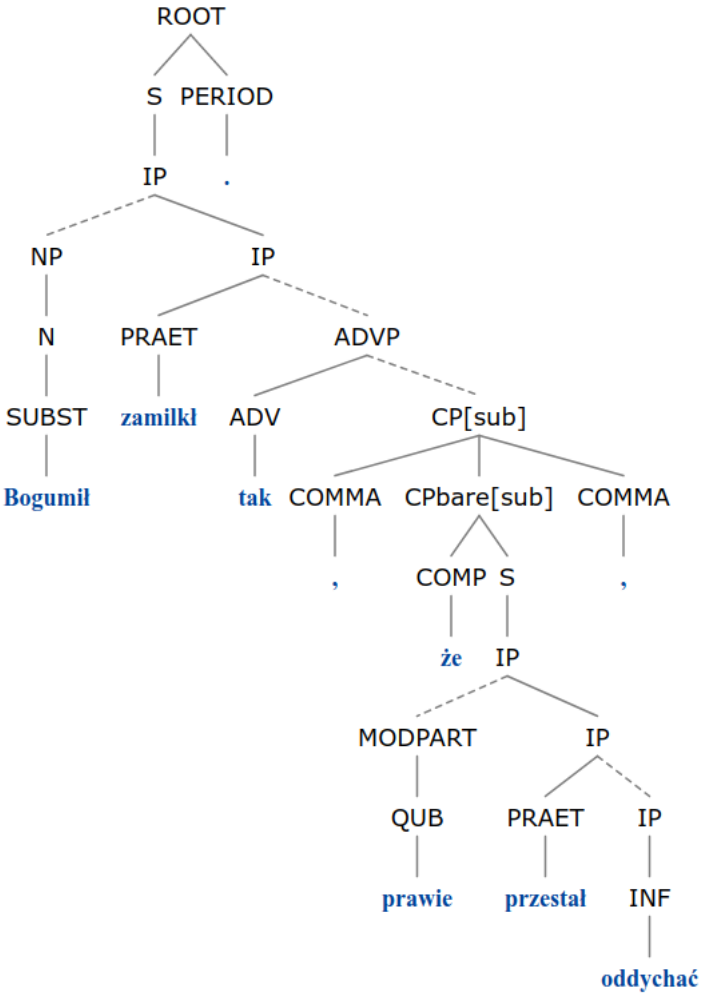


Figure 3.8: C-structure of (2.38) on page 50

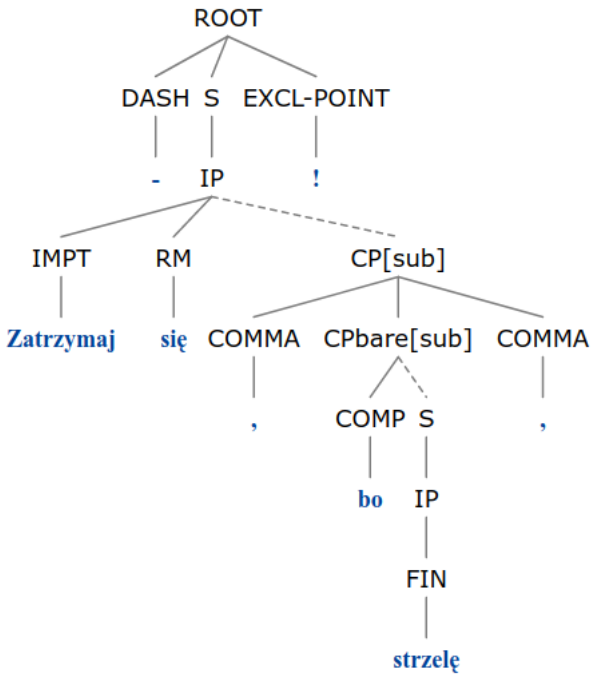


Figure 3.9: C-structure of (2.37) on page 48

In the corresponding c-structure in Figure 3.7, the two segments are represented as separate leaves in the tree (with the mobile inflection preceded by a plus).

Also the conditional mood marker, *by*, may occur in such forms, as in (3.3), where the form *czułbym* '(I'd) feel' consists of three segments: the past form *czuł* 'felt', the conditional mood marker *by* and the mobile inflection *m* signalling first person and singular number.

(3.3) Czułbym się tam źle.
feel.(3)SG.M-COND-1SG INH there badly
'I would feel there bad.'

As shown in Figure 3.10, in such cases the mood marker and the mobile inflection form a constituent, `MOODAGLT`.

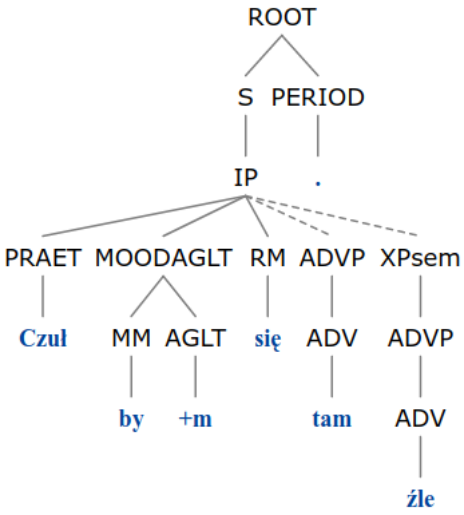


Figure 3.10: C-structure of (3.3)

The reason such person and number markers are called mobile inflections is that they may 'detach' from the verb and 'attach' to another – preceding – constituent, as in (3.4):

(3.4) Coś uczynił?...
what.ACC.N-2SG did.(3)SG.M
'What have you done?'

Here, the second person singular mobile inflection *ś* attaches to the interrogative pronoun *co* 'what'. As shown in Figure 3.11, such sequences involving 'reattached' mobile inflections are marked in the structure bank as `MOODAGLTP`.

The following list summarises these categories:

- `AGLT`: preterminal for a mobile inflection (`aglt` in NKJP)
- `MM`: mood marker – conditional (*by*) or imperative (*niech*)
- `MOODAGLT`: category which must contain a mood marker `MM`, optionally followed by `AGLT`
- `MOODAGLTP`: topmost category which rewrites either to any sequence of dependents (possibly zero) followed by either the mobile inflection `AGLT` or `MOODAGLT`

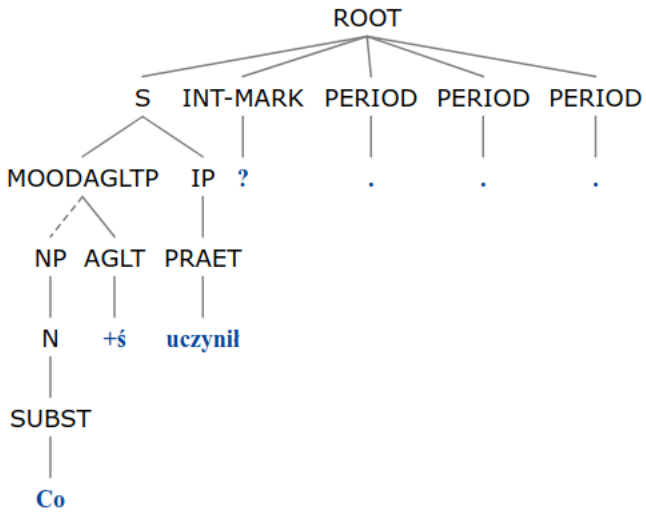


Figure 3.11: C-structure of (3.4)

Note that not only the conditional *by*, but also the imperative *niech* (not illustrated here) bears the MM preterminal.

Many of the trees in this chapter illustrate another kind of marker, the so-called 'reflexive marker': RM is used as the preterminal of all occurrences of *się*, regardless of its function. In most cases, this small word is an inherent (meaningless) part of the verb, as in Figures 3.4–3.5 and 3.10, but it may also mark the impersonal construction, as in Figure 3.15 on page 74.

Moreover, two markers indicate two kinds of negation: the usual (sentential, eventuality) negation witnessed in several trees in this chapter, e.g., in Figures 3.2 and 3.4, and the less frequent constituent negation (Przepiórkowski and Patejuk 2015). In very rare cases, the two negations may be dependents of the same head, as in (3.5), whose c-structure is shown in Figure 3.12.

(3.5)

Władza	ustawodawcza	nie	nie	posiadała	legitymacji
authority.NOM.SG.F	legislative.NOM.SG.F	NEG	NEG	had.3SG.F	legitimacy.GEN.SG.F
demokratycznej.					
democratic.GEN.SG.F					

'Legislature did not not have democratic legitimacy.'

Summarising:

- RM: the word *się*, regardless of its function
- NEG: sentential (eventuality) negation
- CNEG: constituent negation

3.1.5 Nominal constituents

Usually, there are three levels of nominal projections: the maximal nominal phrase, NP, intermediate N and a preterminal corresponding to an NKJP tag, e.g., SUBST ('substantive') for a typical noun. Additionally, when such a nominal phrase is fronted, it may be marked as interrogative, relative or negative:

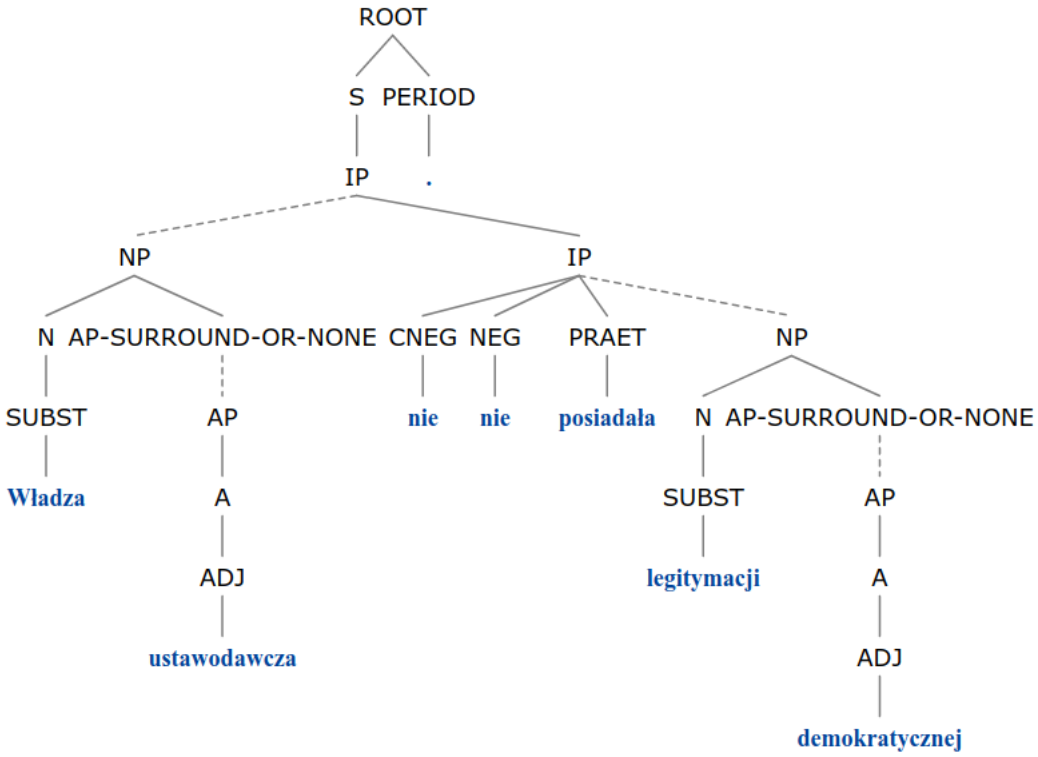


Figure 3.12: C-structure of (3.5)

- NP: topmost nominal phrase category (dominates N, PRON, NUMP); cf., e.g., Figures 3.2, 3.13 and 3.14
- NP[int], NP[neg], NP[rel]: extra nominal phrase category (on top of NP) marking that it is interrogative, negative or relative; cf., e.g., Figure 3.4 on page 65 for NP[int] or Figure 3.24 on page 84 for NP[neg]
- N: immediately dominating category for SUBST, DEPR, SIEBIE and GER (gerund; see Section 3.1.9) preterminals; cf., e.g., Figures 3.2, 3.13 and 3.14
- SUBST: preterminal for nouns (subst in NKJP); cf., e.g., Figures 3.2, 3.13 and 3.14
- DEPR: preterminal for depreciative nominals (depr in NKJP)
- SIEBIE: preterminal for the SIEBIE 'oneself' lexeme (siebie in NKJP)

NP may also dominate a numeral phrase:

- NUMP: topmost numeral phrase category (dominated by NP); cf., e.g., Figure 3.13
- NUMbare: immediately dominating category for NUM preterminal; cf., e.g., Figure 3.13
- NUM: preterminal for numerals (num in NKJP); cf., e.g., Figure 3.13

Finally, an NP may be realised as a personal pronoun (other nominal pronouns are treated as nouns):

- PRON: immediately dominating category for PPRON12 and PPRON3 preterminals; cf., e.g., Figure 3.14
- PPRON12: preterminal for first and second person pronouns (ppron12 in NKJP); cf., e.g., Figure 3.7 on page 68 or Figure 3.22 on page 82
- PPRON3: preterminal for third person pronouns (ppron3 in NKJP); cf., e.g., Figure 3.14

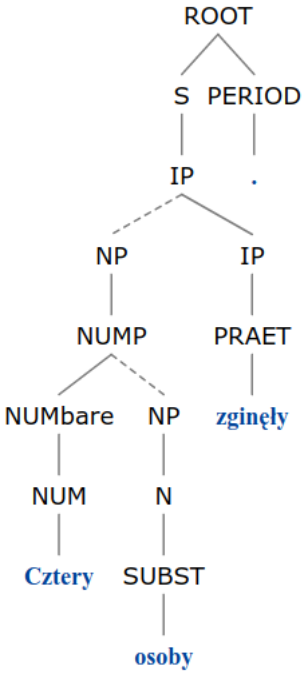


Figure 3.13: C-structure of (2.3) on page 18

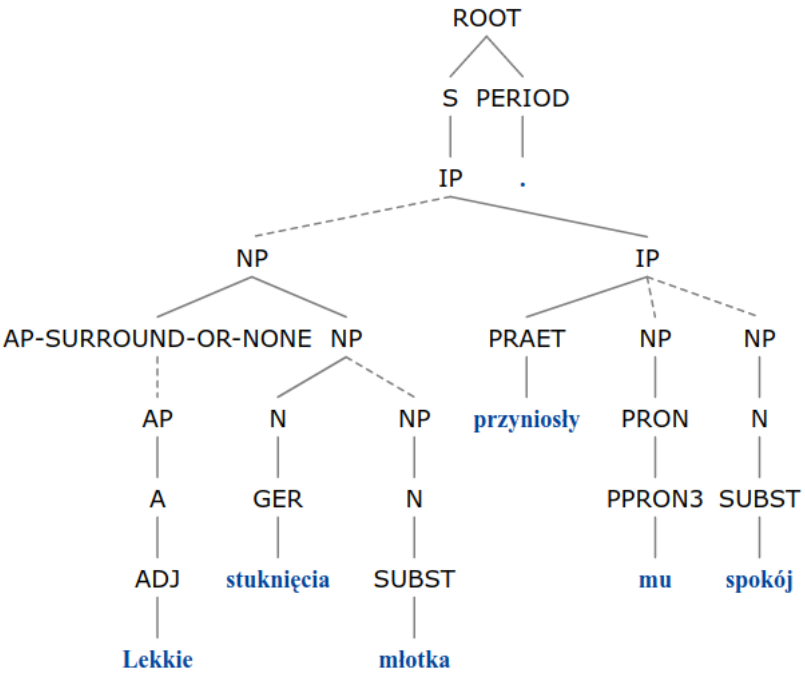


Figure 3.14: C-structure of (2.5) on page 19

3.1.6 Prepositional constituents

There are two types of prepositional phrases: PP, with a nominal constituent, and PAP, with an adjectival constituent:

- PP: topmost prepositional phrase category (dominates P and NP); cf., e.g., Figure 3.15

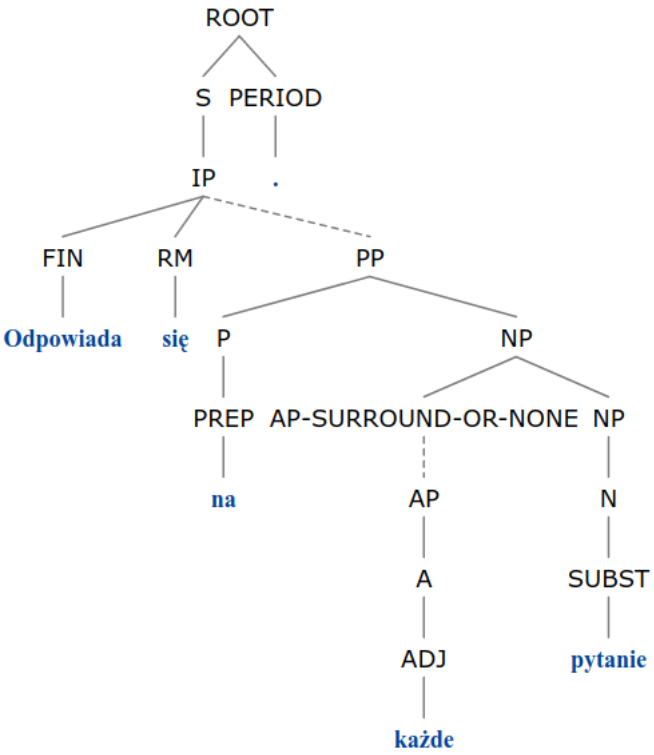


Figure 3.15: C-structure of (2.9) on page 24

- PP[int], PP[neg], PP[rel]: extra prepositional phrase category (on top of PP) marking that it is interrogative (int), negative (neg) or relative (rel)
- PAP: topmost prepositional-adjectival phrase category (dominates P and AP); cf., e.g., Figure 3.16
- P: immediately dominating category for PREP preterminal; cf., e.g., Figures 3.15 and 3.16
- PREP: preterminal for prepositions (prep in NKJP); cf., e.g., Figures 3.15 and 3.16

3.1.7 Adjectival constituents

Adjectival phrases may – but do not have to – act as modifiers within nominal phrases; both possibilities are illustrated in Figure 3.16. When they do, they are dominated – for technical reasons concerned with the proper handling of punctuation – by a perhaps too verbosely named node AP-SURROUND-OR-NONE:

- AP: topmost adjectival phrase category; cf., e.g., Figures 3.14–3.16
- AP[int]: extra adjectival phrase category (on top of AP) marking that it is interrogative (int)

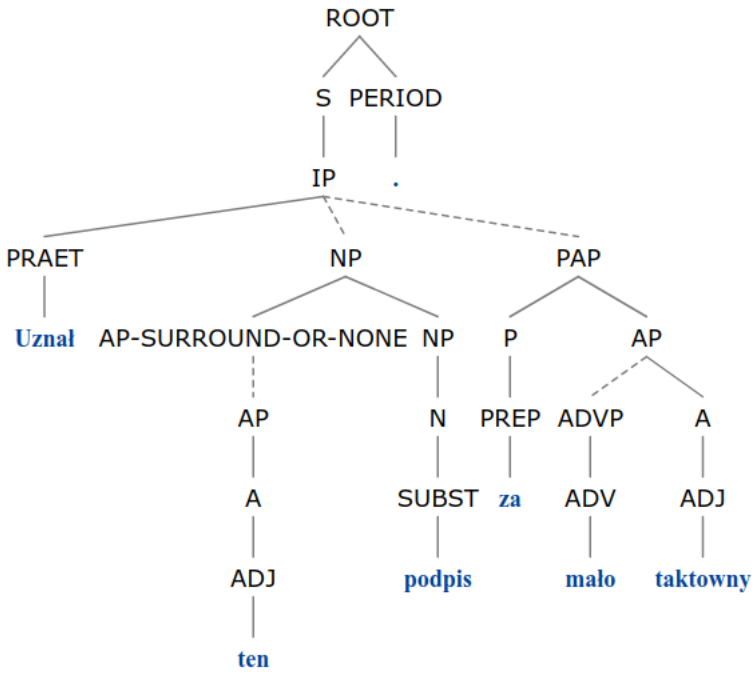


Figure 3.16: C-structure of (2.46) on page 55

- `AP-SURROUND-OR-NONE`: extra adjectival phrase category (on top of `AP`) used when `AP` is part of `NP` (to ensure that it may be optionally surrounded by commas, but only once); cf., e.g., Figures 3.14–3.16
- `A`: immediately dominating category for `ADJ`, `ADJA`, `ADJC`, `ADJP`, `PACT` and `PPAS` preterminals; cf., e.g., Figures 3.14–3.16
- `MODJAKI`: immediately dominating category for selected `ADJ` preterminals (TAKI 'such', JAKI 'what kind of', JAKIŚ 'some kind of') which can modify adjectives
- `ADJ`: preterminal for adjectives (`adj` in NKJP); cf., e.g., Figures 3.14–3.16
- `ADJA`: preterminal for ad-adjectival adjectives (`adja` in NKJP)
- `ADJC`: preterminal for exclusively predicative adjectives (`adjc` in NKJP)
- `ADJP`: preterminal for post-prepositional adjectives (`adjp` in NKJP)

3.1.8 Adverbial constituents

Adverbial phrases normally have few projections:

- `ADVP`: topmost adverbial phrase category; cf., e.g., Figure 3.17
- `ADVP[int]`, `ADVP[rel]`: extra adverbial phrase category (on top of `ADVP`) marking that it is interrogative (int) or relative (rel)
- `ADV`: preterminal for adverbs (`adv` in NKJP); cf., e.g., Figure 3.17

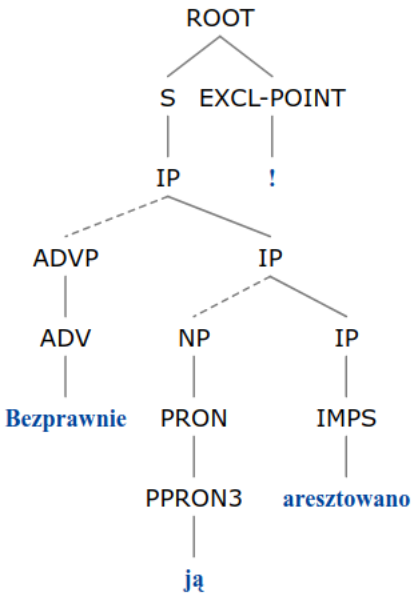


Figure 3.17: C-structure of (2.1) on page 13

3.1.9 Mixed categories

As in the legacy tagset, some preterminals mark mixed categories – gerunds, adjectival participles and adverbial participles:

- `GER`: preterminal for gerund (`ger` in NKJP), a verbal-nominal category immediately dominated by `N`; cf., e.g., Figure 3.14
- `PACT`: preterminal for active adjectival participle (`pact` in NKJP), a verbal-adjectival category immediately dominated by `A`
- `PPAS`: preterminal for passive adjectival participle (`ppas` in NKJP), a verbal-adjectival category immediately dominated by `A`
- `PCON`: preterminal for contemporary adverbial participle (`pcon` in NKJP), a verbal-adverbial category immediately dominated by `IP` (where negation may attach)
- `PANT`: preterminal for anterior adverbial participle (`pant` in NKJP), a verbal-adverbial category immediately dominated by `IP` (where negation may attach)

3.1.10 Modifying particles

Unlike adverbs, which normally modify verbs and adjectives, particles may modify a wider range of constituents:

- `MODPART`: immediately dominating category for `QUB` preterminal; cf., e.g., Figure 3.18 and Figure 3.8 on page 69
- `QUB`: preterminal for particles, adnumeral operators and intensifiers (`qub` in NKJP); cf., e.g., Figure 3.18 or Figure 3.8 on page 69
- `QUB[int]`: preterminal for interrogative (int) particles such as CZY (`qub` in NKJP)

3.1.11 Interjections

- `INTERJ`: preterminal for interjections (`interj` in NKJP); cf. Figure 3.18

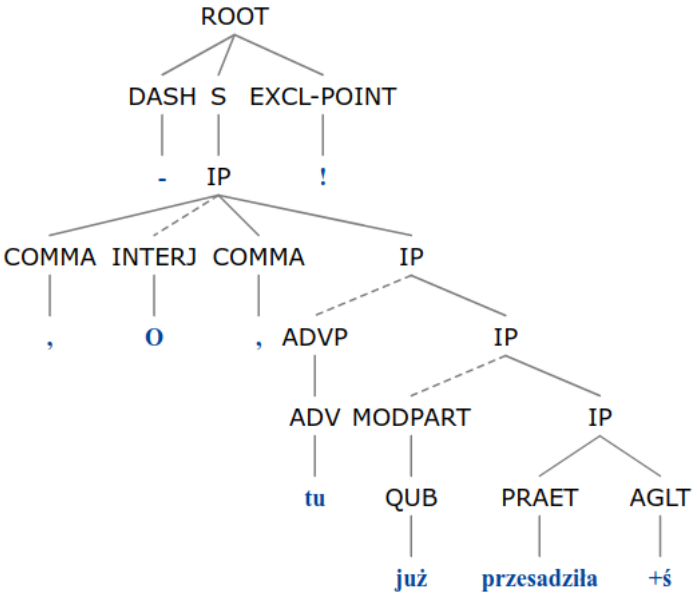


Figure 3.18: C-structure of (3.6)

Just as subordinate clauses, interjections are normally introduced by surrounding commas, which do not have to appear in the sentence if they coincide with other punctuation (or with sentence boundary), as in (3.6). Nevertheless, they are still present in the c-structure, as shown in Figure 3.18.

(3.6) - O, tu już przesadziłaś!
oh here already overstepped.2.SG.F
'– Oh, here you have overstepped the mark!'

3.1.12 Special phrases (not based on a specific category): XP…

- `XPsem`: topmost semantically defined XP category
- `XPscr[int]`: topmost scrambled interrogative (int) phrase
- `XPscr[neg]`: topmost scrambled negative (neg) phrase
- `XPextr[int]`: topmost extracted interrogative (int) phrase
- `XPextr[rel]`: topmost extracted relative (rel) phrase

Nodes of type `XPsem` may head constituents of diverse syntactic categories, depending on the semantic role of such constituents in f-structures. For example, a typical realisation of those `XPsem` c-structure nodes which map to the values of the f-structure OBL-ADL attribute (cf. Section 2.9.3) is by a prepositional phrase (`PP`), especially with a preposition such as DO 'to', as in the c-structure in Figure 3.7 on page 68. There, the `PP` *do mnie* 'to me', is an adlative argument of the verb *przyjdziecie* '(you will) come', as made explicit in the f-structure in Figure 2.39 on page 49 (see the functional substructure with index 67 there). However, such an adlative

XPsem may also be realised by an adverbial phrase headed by *dokąd* 'where to', *tam* 'there', etc. Similarly, XPsem constituents corresponding to OBL-LOCAT arguments are often realised by prepositional phrases with prepositions such as *w* 'in', as in Figure 3.4 on page 65, but may also be realised for example by the adverbial *tam* 'there', as in Figure 3.6 on page 67.

Also the other XP… nodes mentioned above may have different categorial realisations, but they also indicate a non-local occurrence of a constituent; see Section 3.3 for details.

3.1.13 Coordination: (PRE)CONJ

In coordinate structures, the conjunction, CONJ, is the head; if a preconjunction, PRECONJ, also occurs in the sentence, as in (3.7) and the corresponding Figure 3.19, it is a co-head (a term to be defined in the ensuing section).

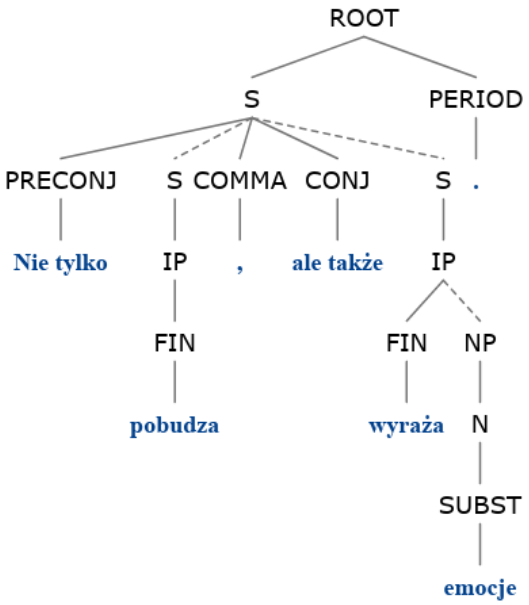


Figure 3.19: C-structure of (3.7)

(3.7) Nie tylko pobudza, ale także wyraża emocje.
NEG only invigorates.3SG but also expresses.3SG emotion.ACC.PL.F
'Not only does it invigorate, but it also expresses emotions.'

No other nonterminals are specific to coordination:

- CONJ: preterminal for conjunctions (conj in NKJP)
- PRECONJ: preterminal for preconjunctions (conj in NKJP)

3.2 Co-heads

Each nonterminal node in a constituency structure has at least one head, i.e., a daughter which maps to the same functional structure. Such head daughters are marked via solid edges. For example, in Figure 3.19, the head daughter of the IP node is FIN (rather than the NP constituent).

However, sometimes more than one daughter maps to the same f-structure as the mother node. Consider again sentence (2.20), repeated again below, and its syntactic structures in Figures 3.1 and 3.2, repeated below as Figures 3.20 and 3.21.

(2.20) Nie mają wyboru.
NEG have.3PL choice.GEN.SG.M
'They have no choice.'

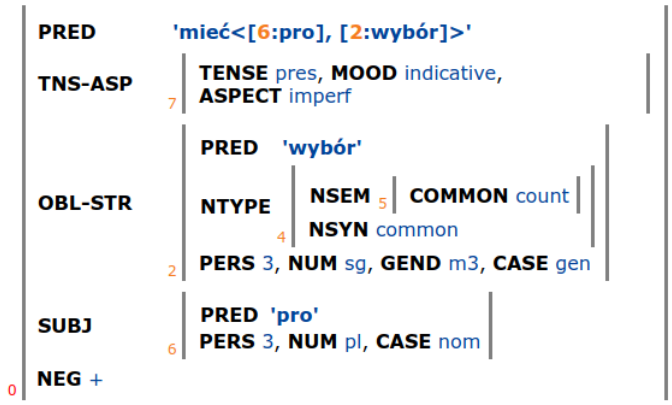


Figure 3.20: F-structure of (2.20)

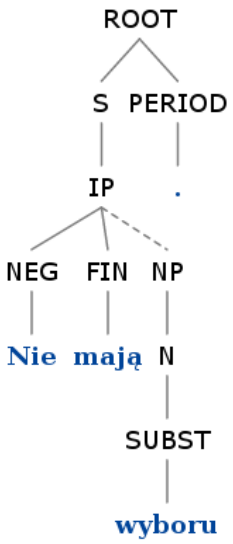


Figure 3.21: C-structure of (2.20)

Here, the IP node in Figure 3.21 has two heads – hence, *co-heads* – namely, the preterminals NEG and FIN. All three nodes map to the same functional structure – the topmost f-structure in Figure 3.20 (with index 0). This is an example of one of two typical situations giving rise to

co-heads: when some of the heads introduce functional information, such as the presence of negation (as in this example), or a particular tense introduced by an auxiliary; see Section 3.2.1. Another typical situation – much less interesting linguistically – is illustrated by the two co-heads of the ROOT constituent in Figure 3.21: one of them is simply a punctuation mark; see Section 3.2.2.

3.2.1 Functional co-heads

One type of co-heads are categories which introduce f-structure annotation: if more than one co-head introduces functional annotation, at most one of them – the one headed by a content word – can introduce a PRED attribute. This is because this attribute is instantiated, which means that its value can be introduced only once. The other co-heads introduce additional – typically categorial or morphosyntactic – information.

Typical functional co-heads include:

- non-semantic prepositions (P): contribute PFORM instead of PRED, which is contributed by the nominal inside the PP or by the adjective inside the PAP
- non-semantic complementisers (COMP): contribute COMP-FORM instead of PRED, which is contributed by the verbal predicate inside the CP
- verbs:
  - AUX: an auxiliary contributing person and number information
  - AGLT: mobile inflection contributing person and number information
- markers:
  - negation:
    - NEG: sentential negation (also called 'eventuality negation')
    - CNEG: constituent negation
  - RM: the word *się* (in various uses)
  - RSM: the resumptive pronoun *co*
  - MM: mood (imperative, conditional)
- QUB[int]: yes/no interrogative particle
- CONJ, PRECONJ: (pre)conjunctions

3.2.2 Punctuation co-heads

The other type of co-heads are categories which correspond to punctuation marks – since they do not introduce any f-structure attributes, they can only be co-heads. If there is any f-structure annotation corresponding to a c-structure node to which a punctuation category projects, it can only be contributed by heads (introducing PRED attribute) and by functional co-heads.

Typical punctuation co-heads include:

- sentence ending marks (note that multiple marks may be used together):
  - PERIOD: period (.)

  - `ELLIPSIS`: ellipsis as one character (…)
  - `INT-MARK`: question mark (?)
  - `EXCL-POINT`: exclamation mark (!)
- `COMMA`: comma (,):
  - as a conjunction (in asyndetic coordination)
  - as pure punctuation (e.g., surrounding subordinate phrases)
- `DASH`: dash (-, –, —):
  - at the start of dialogue turn
  - as a list item
- brackets (not necessarily balanced):
  - `L-PRN`, `R-PRN`: left (() and right ()) round bracket
  - `L-SQR`, `R-SQR`: left ([) and right (]) square bracket
- quotes (not necessarily balanced):
  - `LD-QT`, `RD-QT`: left (”) and right (”) double quotes
  - `LE-QT`, `RE-QT`: left (“) and right (”) English quotes
  - `LP-QT`, `RP-QT`: left („) and right (”) Polish quotes

3.3 Non-local dependencies

The LFG grammar of Polish underlying the structure bank distinguishes – together with some generative literature – between scrambling and extraction. Both terms refer to possibly non-local realisations of some constituents. Scrambling is typical of languages with so-called free word order, such as Polish, and consists in the freedom of various constituents to appear in diverse positions within a tensed clause.

For example, in (3.8), the interrogative word *co* ‘what’, though it is a dependent of *wiedzieć* ‘know’, is placed outside this phrase in terms of c-structure, as shown in Figure 3.22 – *co* is fronted and at the level of c-structure it belongs to the phrase headed by *możesz* ‘(you) may, (you) can’. (Note the sequence of solid edges between `XPscr[int]` and the `IP` headed by *możesz* and note that this sequence does not extend to the `IP` headed by *wiedzieć*.) Despite this fact, its f-structure representation in Figure 3.23 shows that CO ‘what’ is a dependent (OBL-STR, 124) of WIEDZIEĆ ‘know’, 102, rather than MÓC ‘may, can’, 0. Hence, at the level of c-structure, *co* is identified as an interrogative scrambled element (`XPscr[int]`).

(3.8) Co ty możesz wiedzieć o głodzie, chłodzie i
what.ACC.SG.N you.2SG can.2SG know.INF about hunger.LOC.SG.M cold.LOC.SG.M and
bezdomności?
homelessness.LOC.SG.F
‘What can you know about hunger, being cold and homelessness?’

Similarly, in (3.9), the negative phrase *nic mądrzejszego* ‘nothing smarter’, though it is a dependent of *wymyślić* ‘invent’, is placed outside this phrase in terms of c-structure, as shown in Figure 3.24 – it is fronted and at the level of c-structure it belongs to the phrase headed by *potrafi* ‘may, be capable of’. Despite this fact, its f-structure representation in Figure 3.25 shows that NIC ‘nothing’ is a dependent (OBJ, 32) of WYMYŚLIĆ ‘invent’, 30, rather than of

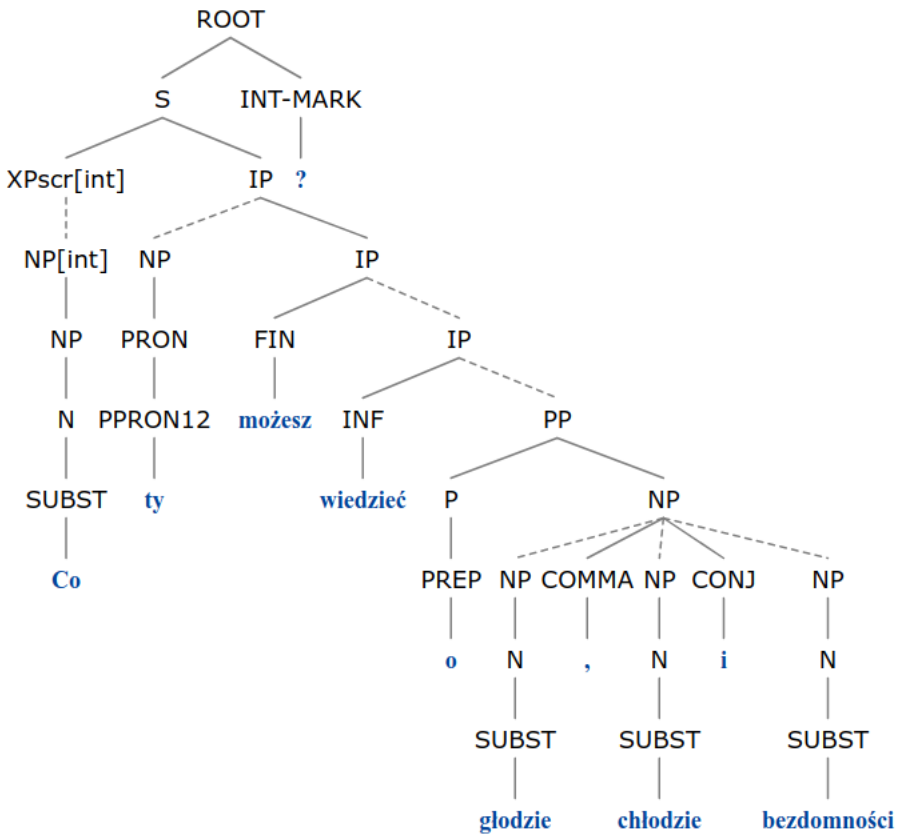


Figure 3.22: C-structure of (3.8)

POTRAFIĆ 'may, be capable of', 0. Hence, at the level of c-structure, *nic* is identified as a negative scrambled element (`XPscr[neg]`).

(3.9) Nic mądrzejszego ten naród nie potrafi
nothing.GEN.SG.N smarter.GEN.SG.N this.NOM.SG.M nation.NOM.SG.M NEG can.3SG
wymyślić.
invent.INF
'This nation is not capable of inventing something smarter.'

In the current version of the structure bank, scrambling is limited to these two kinds of 'dislocated' constituents: interrogative and negative. Under scrambling, the scrambled item must be placed at the level of c-structure outside the phrase to which it belongs in terms of f-structure, but it may not cross clause boundaries (`CP` in c-structure, COMP in f-structure) – it may only cross the boundaries of infinitival verb phrases.

On the other hand, extraction – which is understood here as applying to obligatorily fronted interrogative constituents in subordinate questions and relative constituents in relative clauses – is less constrained, as the dislocated element may in principle cross such a clause boundary (cf., e.g., Witkoś 1993, 1995 and references therein). However, in the LFG structure bank of Polish, there seem to be no sentences illustrating this truly non-local potential. Hence, the c-structure in Figure 3.4 on page 65 is typical: ignoring punctuation, the subordinate interrogative clause `CP[int]` consists of an obligatorily fronted interrogative element `XPextr[int]`, which is the interrogative nominal phrase `NP[int]` dominating the sole noun *co* 'what', and of the rest of the sentence `S`, with no real extraction across clause boundary taking place.

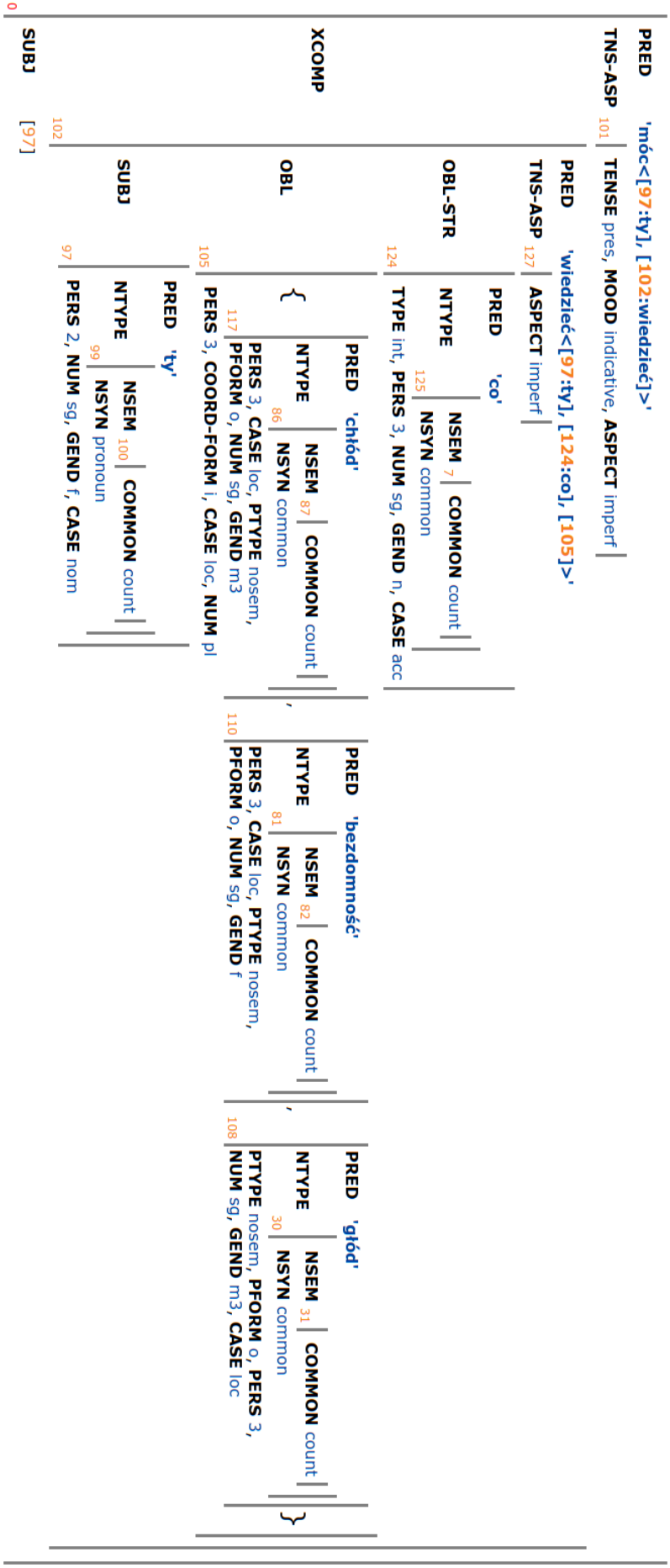


Figure 3.23: F-structure of (3.8)

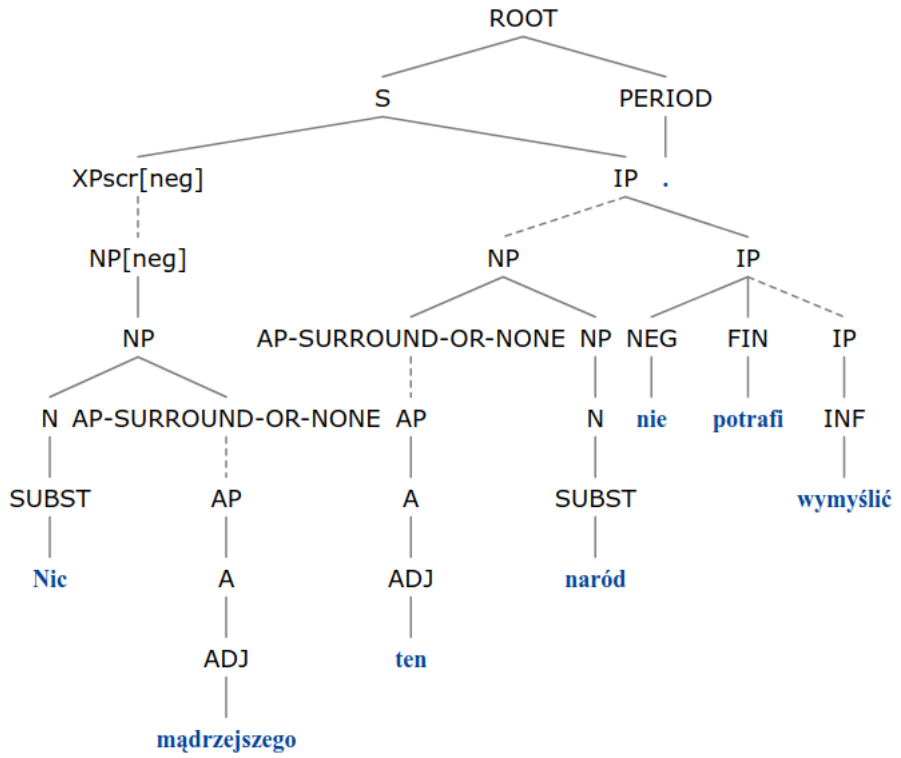


Figure 3.24: C-structure of (3.9)

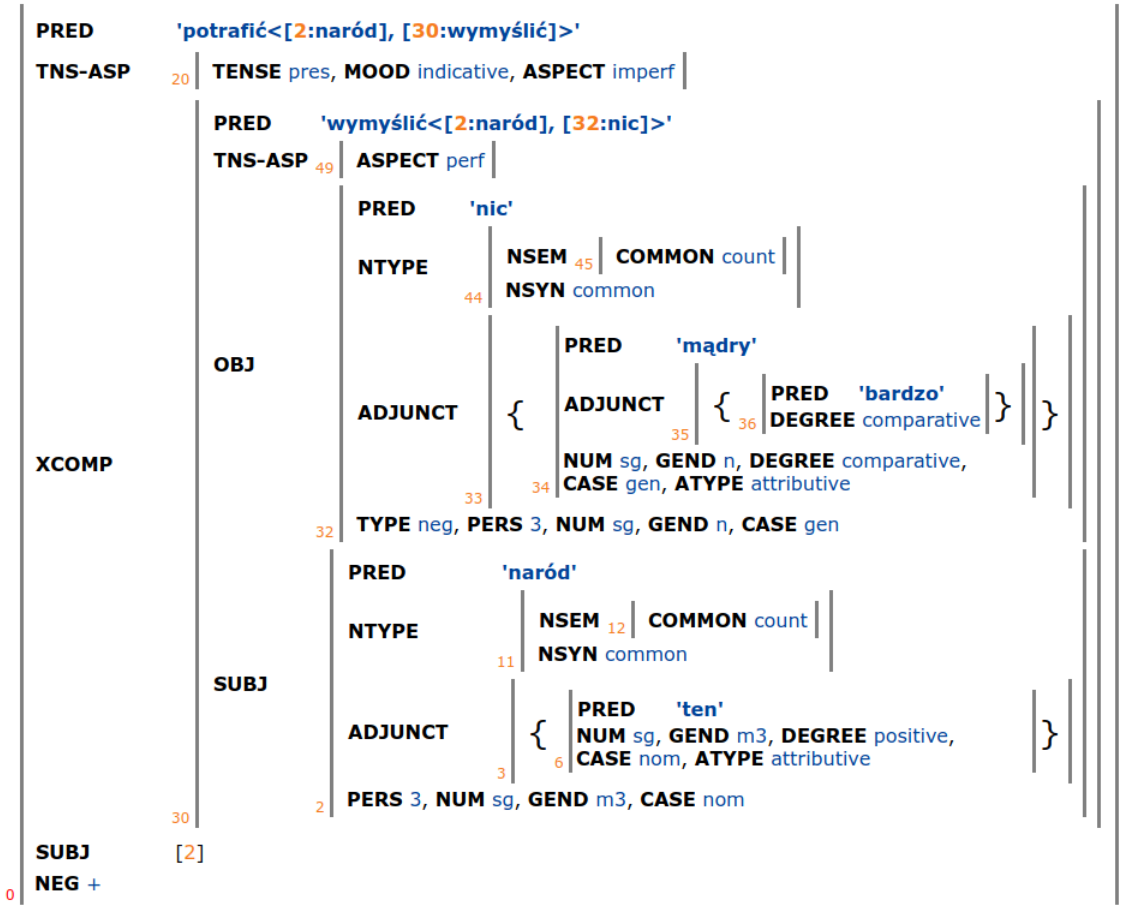


Figure 3.25: F-structure of (3.9)

Part II

From LFG to Enhanced UD

Chapter 4

Input, intermediate representation, output

Conversion of LFG structures to dependency structures is not a new task (cf., e.g., Øvrelid et al. 2009, Çetinoğlu et al. 2010 and, more recently, Meurer 2017), but – with the notable exception of Meurer 2017 – previous attempts are only mentioned or very roughly outlined in the literature. Moreover, previous work has been limited to *dependency trees* as the output format. As is well known, simple dependency trees cannot straightforwardly represent many kinds of linguistic information, so the conversion from representations such as those assumed in LFG invariably resulted in considerable loss of information.

There is some disagreement about which syntactic level of representation – constituency structure or functional structure – is the most natural basis for constructing dependency representations. While f-structure seems to be a natural candidate, Meurer 2017 sketches a conversion procedure based mainly on c-structure and consisting in step-wise transformations of the constituency tree into a dependency tree.

The approach presented here follows the more standard observation that f-structures provide a good basis for dependency relations. Of course, c-structures cannot be completely ignored, as only they contain the actual tokens in the sentence. We show, however, that – apart from f-structures – information encoded in terminal and pre-terminal nodes of the constituency tree, together with the standard correspondence between c-structure preterminals and f-structure components, is sufficient to perform the conversion, i.e., that the actual constituency information may be completely ignored.

4.1 LFG input

Let us illustrate the input to – and output of – the conversion procedure on the basis of the following example:[1]

(4.1) Mężczyzna nie zdążył ich otworzyć.
man.NOM NEG managed them.GEN open.INF
'The man didn't manage to open them on time.'

[1] Concerning morphosyntactic information in glosses, see footnote 1 on page 3.

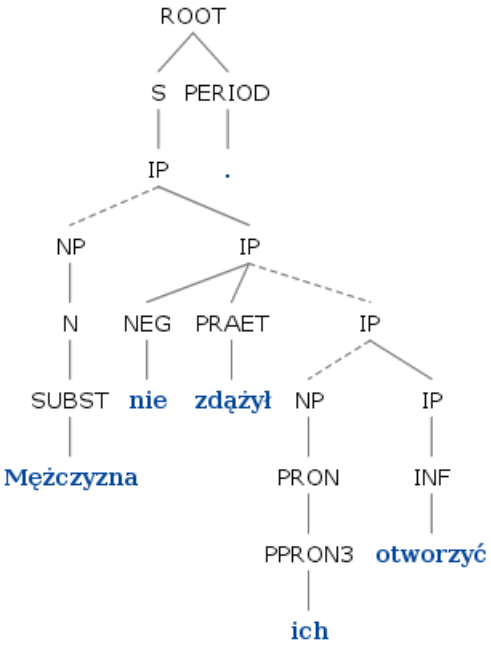


Figure 4.1: C-structure of (4.1)

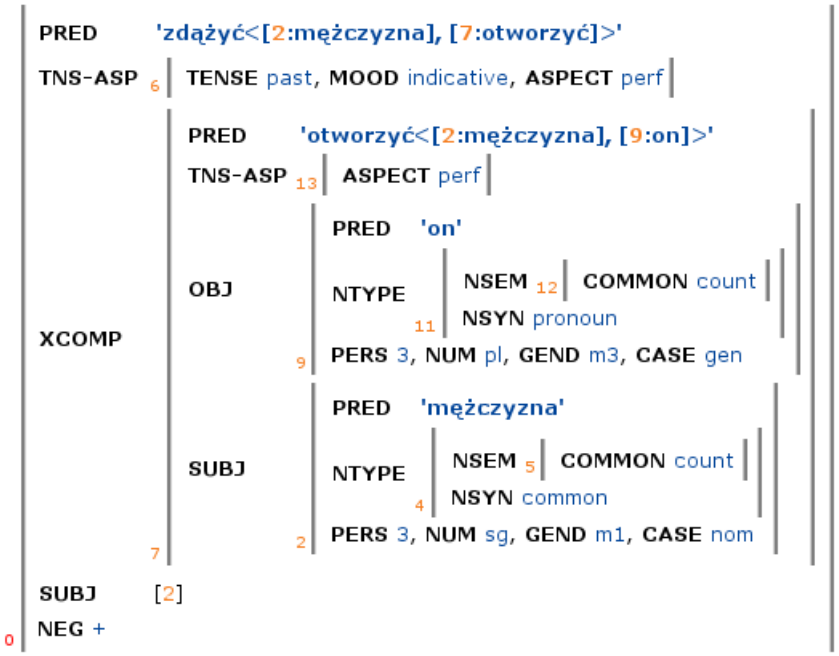


Figure 4.2: F-structure of (4.1)

Constituency and functional LFG representations of (4.1) are shown in Figures 4.1–4.2. According to the c-structure in Figure 4.1, the negated main verb *nie zdążył* ‘didn’t manage (on time)’ combines with an IP, *ich otworzyć* ‘open them’, and an NP, *mężczyzna* ‘man’. Additionally, the f-structure in Figure 4.2 shows that this is a subject control construction: *mężczyzna* ‘man’, which is the overt subject of the main verb, is also understood as the subject of the infinitival *otworzyć* ‘open’. The f-structure also contains information about various morphosyntactic features of particular constituents. Below, we will mostly ignore such information and we will concentrate on grammatical functions, simplifying the presentation of such f-structures as in Figure 4.3.

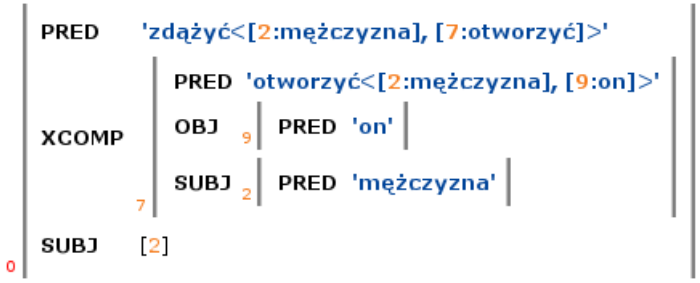


Figure 4.3: Schematic f-structure of (4.1)

Such LFG analyses are, within the INESS search and visualisation platform, represented in a rather opaque Prolog format, which is the legacy format of the XLE system for running LFG grammars. So the first step was to convert such Prolog-based representations into a more standard XML format,[2] namely, TigerXML (Brants et al. 2002; König et al. 2003).[3] The complete XML representation of the running example is given in Appendix B.

4.2 Intermediate dependency representation

As described in detail in Chapter 7, conversion from LFG to UD is performed in two stages, with an internal intermediate dependency representation as the result of the first stage. This dependency representation is maximally close to the input LFG representation and, in particular, it retains the f-structure grammatical functions as names of dependency relations. Such an initial LFG-like dependency representation for the running example is given in Figure 4.4.

Note that this representation correctly models control, i.e., the fact that *mężczyzna* ‘man’ is the subject of both the main verb and the embedded infinitival verb. As a result, there are two incoming SUBJ edges to this noun, so this representation is not a dependency tree. For this reason, a simpler initial dependency representation is also constructed at this stage, which is a tree. In this particular case, this is achieved by removing the control information, i.e., by deleting the SUBJ edge from the controlled infinitival verb – see Figure 4.5.

[2]This conversion step was performed by Michał Kućko, a student at the Cognitive Science Programme of the University of Warsaw.

[3]See also `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html`, accessed on 21 February 2018.

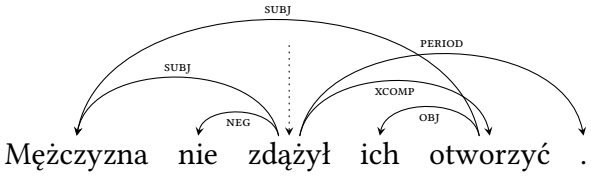


Figure 4.4: Initial dependency representation of (4.1)

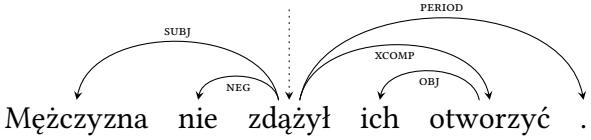


Figure 4.5: Initial dependency representation of (4.1) – basic tree

In the second stage of conversion, these two representations – basic tree and full dependency representation – are converted into the two Universal Dependencies representations: basic and enhanced.

4.3 UD output

The output of the conversion is a list of basic and enhanced Universal Dependencies representations in the CoNLL-U format[4], derived from the earlier CoNLL-X format (Buchholz and Marsi 2006). This is a textual format, representations of particular sentences are separated by an empty line, and each representation consists of a number of comment lines (starting with the hash character, #) followed by the actual encoding of the dependency representation of a sentence, as in Figure 4.6.

A single line in the representation proper corresponds to a single token in the sentence,[5] and their order reflects the order of the tokens in the sentence. Each line consists of 10 columns separated by the tab character (represented by a single space in Figure 4.6):

1. ID: the consecutive number of the token in the sentence,
2. FORM: the token,
3. LEMMA: the lemma of this token,
4. UPOS: the coarse-grained part of speech drawn from the repertoire of 17 universal parts of speech assumed in UD,
5. XPOS: the legacy tag of the token (see Appendix A),
6. FEATS: morphosyntactic features in the `Feature=Value` format, separated by the vertical bar, i.e. `|`,
7. HEAD: the ID of the governor of the current token (or `0`, if it is the root) in the basic dependency tree,
8. DEPREL: the label of the dependency relation from the governor in the basic tree,

[4] See `http://universaldependencies.org/format.html`, accessed on 21 February 2018.

[5] In Figure 4.6 and the following CoNLL-U representations some lines are broken for typographic reasons.

```
# sent_id = train-5386
# text = Mężczyzna nie zdążył ich otworzyć.
# converted_from_file = NKJP1M_1305000000506_morph_1-p_morph_1.40-s-dis@1.xml
# genre = news
1 Mężczyzna mężczyzna NOUN subst:sg:nom:m1 Case=Nom|Gender=Masc|Number=Sing|SubGender=Masc1 3 nsubj 3:nsubj|5:nsubj _
2 nie nie PART qub Polarity=Neg 3 advmod 3:advmod _
3 zdążył zdążyć VERB praet:sg:m1:perf
  Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|SubGender=Masc1|Tense=Past|VerbForm=Fin|Voice=Act 0 root 0:root _
4 ich on PRON ppron3:pl:gen:m3:ter:akc:npraep
  Case=Gen|Gender=Masc|Number=Plur|Person=3|PrepCase=Npr|PronType=Prs|SubGender=Masc3|Variant=Long 5 obj 5:obj _
5 otworzyć otworzyć VERB inf:perf Aspect=Perf|VerbForm=Inf|Voice=Act 3 xcomp 3:xcomp SpaceAfter=No
6 . . PUNCT interp PunctType=Peri 3 punct 3:punct _
```

Figure 4.6: CoNLL-U representation of (4.1)

9. DEPS: a |-separated list of incoming dependency relations – each represented as `head:deprel`, e.g., `3:nsubj` – in the enhanced dependency structure,
10. MISC: any other – non-morphosyntactic – features in the same format as in the FEATS field.

As such textual representations are rather hard to read, in the following chapters we will visualise them – but without the information in the LEMMA, XPOS, FEATS and MISC fields – as in Figure 4.7. There, the tokens are adorned with their UPOS value, the basic dependency tree

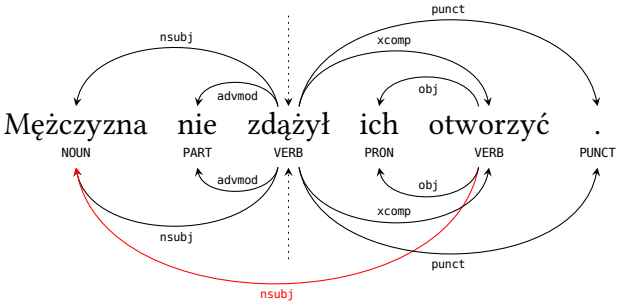


Figure 4.7: Final UD representation of (4.1)

(encoded in HEAD and DEPREL) is presented above them, the enhanced dependency structure (encoded in DEPS) – below them, and those dependency relations which are not identical in both representations are shown in red. Whenever the enhanced dependency tree is identical with the basic dependency tree, it will only be drawn once (above the tokens).

Chapter 5

Tokenisation

Tokens for the UD representation are generally read off the LFG c-structure. There are three exceptions, to be illustrated with the example in (5.1), whose c-structure is presented in Figure 5.1, the initial dependency representation – in Figure 5.2, and final UD representation – in Figure 5.3.

(5.1) Teraz już wiem na pewno, że nas oszukałyście.
now already know.1SG for sure COMP us.ACC cheated.2PL.F
'Now I know for sure that you have cheated us.'

Note that the nodes in the initial dependency representation correspond directly to the leaves in the LFG c-structure, and that there are some tokenisation differences between these two representations and the final UD representation. The three relevant differences between the input and the output of conversion procedure are discussed in the three sections below.

5.1 Mobile inflections

For technical reasons, mobile inflections expressing number and person, e.g., *-ście* '2PL', are marked in the LFG structure bank with an initial '+', which needs to be removed during conversion. Information about the special status of such elements is preserved – not only in the original morphosyntactic tag (the value of the XPOS in CoNLL-U representations, e.g., `aglt:pl:sec:imperf:nwok` in the case of *-ście*, where `aglt` stands for the Polish term for such a mobile inflection, *aglutynant*), but also in the subtype of the dependency relation (see `aux:aglt` in Figure 5.3). Since such mobile inflections attach to the preceding word, the previous token is marked as `SpaceAfter=No` in the CoNLL-U representation (in the MISC field), cf. line 9 in Figure 5.4.

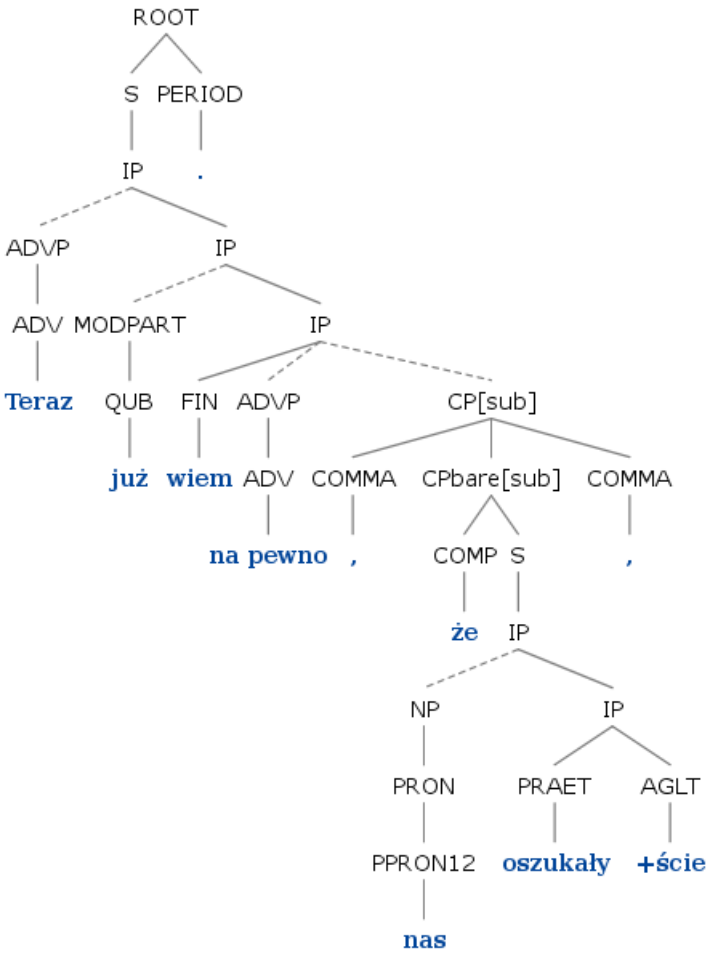


Figure 5.1: C-structure of (5.1)

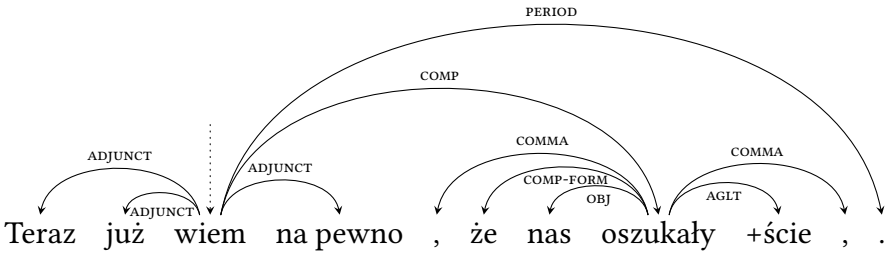


Figure 5.2: Initial dependency representation of (5.1)

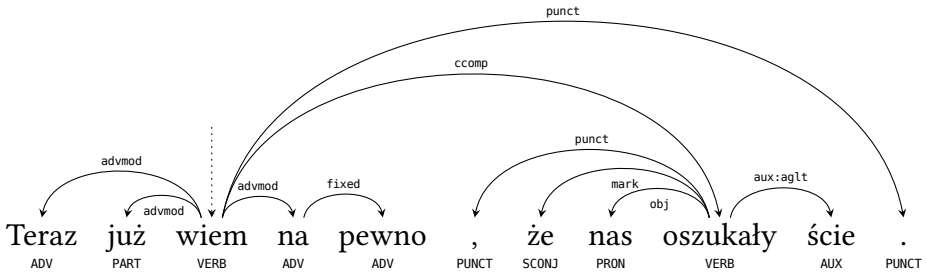


Figure 5.3: Final UD representation of (5.1)

```
1 Teraz teraz ADV adv _ 3 advmod 3:advmod _
2 już już PART qub _ 3 advmod 3:advmod _
3 wiem wiedzieć VERB fin:sg:pri:imperf
  Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 0 root 0:root _
4 na na ADV adv _ 3 advmod 3:advmod _
5 pewno pewno ADV adv _ 4 fixed 4:fixed SpaceAfter=No
6 , , PUNCT interp PunctType=Comm 9 punct 9:punct _
7 że że SCONJ comp _ 9 mark 9:mark _
8 nas my PRON ppron12:pl:acc:m1:pri Case=Acc|Gender=Masc|Number=Plur|Person=1|PronType=Prs|SubGender=Masc1 9 obj 9:obj _
9 oszukały oszukać VERB praet:pl:f:perf
  Aspect=Perf|Gender=Fem|Mood=Ind|Number=Plur|Tense=Past|VerbForm=Fin|Voice=Act 3 ccomp 3:ccomp SpaceAfter=No
10 ście być AUX aglt:pl:sec:imperf:nwok Aspect=Imp|Number=Plur|Person=2|Variant=Short 9 aux:aglt 9:aux:aglt SpaceAfter=No
11 . . PUNCT interp PunctType=Peri 3 punct 3:punct _
```

Figure 5.4: CoNLL-U representation of (5.1)

5.2 Spurious punctuation

Another feature of LFG representations is the presence of occasional commas in the LFG tree which were not present in the input text. Such spurious commas result from the interaction of the POLFIE grammar, which includes rules requiring subordinate clauses, etc., to be surrounded by commas, and the tokeniser, which optionally adds such commas at certain places of the input (roughly, near the beginning and the end of a sentence). Spurious punctuation may also appear in LFG trees in the cases of those abbreviations ending with a period which occur at the end of a sentence. In both cases orthographic rules of Polish require that the two logical punctuation marks – a period which is an integral part of the abbreviation and a period marking the end of the sentence, or a comma signalling the boundary of a subordinate clause, etc., and another adjacent punctuation (or beginning of a sentence) – be contracted to one punctuation mark. Hence, the "spurious" commas or periods in LFG representations simply reflect the underlying "logical" punctuation structure of the sentence.

Nevertheless, such added punctuation marks need to be removed in the conversion. In the case of the example sentence (5.1), this means removing the penultimate token from the representation in Figure 5.2. Fortunately, such added punctuation tokens do not have any dependents, so removing them is straightforward.

5.3 Words with spaces

The final exception to the principle that tokens in UD representations correspond directly to tokens in LFG representaitons is concerned with "words with spaces", e.g., *na pewno* 'for sure, certainly' in (5.1). Other cases of such "multi-token words" include certain conjunctions (e.g., *a także* 'and also', *jak i* 'as also', *ale nie* 'but not'), certain prepositions (e.g., *z powodu* 'because of', *na temat* 'on the topic of', *w czasie* 'during'), certain complementisers (e.g., *mimo że* 'although' or *podczas gdy* 'while'), and the adnumeral modifier *co najmniej* 'at least'.

UD guidelines on tokenisation explicitly state that such multi-token expressions should be treated as sequences of separate tokens,[1] related via the `fixed` relation,[2] as illustrated in Figure 5.3. If such a "word with spaces" has dependents, they are inherited by the first token, which acts as the head of the `fixed` dependency. This is illustrated with example (5.2), whose initial (LFG-like) and final (UD) dependency representations are given in Figures 5.5 and 5.6.[3] As discussed in Chapter 6 (Section 6.1), all tokens related with the `fixed` dependency have the same morphosyntactic information, which pertains to the "word with spaces" as a whole rather than to single tokens that constitute it.

(5.2) Policja wszczęła śledztwo w sprawie wybuchu.
police.NOM started investigation.ACC in matter explosion.GEN
'The Police started an investigation in the matter of the explosion.'

[1] `http://universaldependencies.org/u/overview/tokenization.html`

[2] `http://universaldependencies.org/u/dep/all.html#al-u-dep/fixed`

[3] Note that in the final UD representation the direction of the relation between *w sprawie* 'in matter' and *wybuchu* 'explosion' is reversed with respect to the initial representation; see Chapter 7 for details.

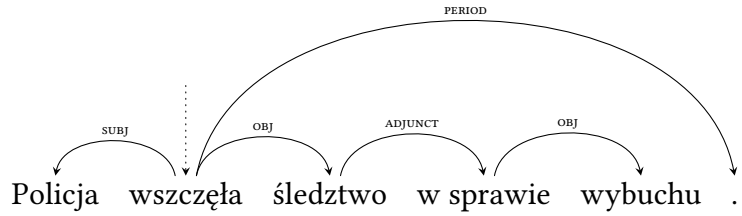


Figure 5.5: Initial dependency representation of (5.2)

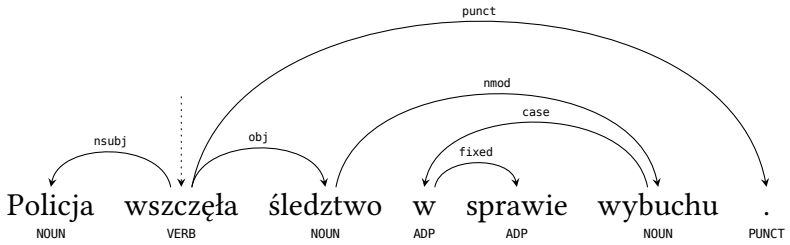


Figure 5.6: Final UD representation of (5.2)

Chapter 6

Morphosyntax

There are three fields in the CoNLL-U representation devoted to morphosyntactic information: UPOS, whose values are drawn from the set of 17 coarse universal part-of-speech categories, XPOS, whose values may be language- and treebank-specific, and FEATS, a list of `Feature=Value` pairs, where features and values should be drawn from the universal feature inventory, but may also be language-specific. Additionally, the MISC field may contain various information that goes beyond pure morphosyntax. The use of these four fields in $\text{UD}^{\text{PL}}_{\text{LFG}}$ – the Polish UD treebank resulting from the conversion – is described in the four sections below.

6.1 XPOS

The value of XPOS is a tag conforming to the NKJP tagset (Przepiórkowski 2009), which is a slightly modified version of the tagset designed for the IPI PAN Corpus of Polish (Przepiórkowski and Woliński 2003a). This tagset is amply documented,[1] and summarised in Appendix A. Each tag is a colon-separated list of atoms, e.g., `subst:sg:nom:f` for the nominal form *książka* 'book', where the first atom (here: `subst`) is the detailed part-of-speech, and the other atoms are values of morphosyntactic features appropriate for this part-of-speech (here: `sg` for singular number, `nom` for nominative case, `f` for feminine gender).

As sentences in $\text{UD}^{\text{PL}}_{\text{LFG}}$ are drawn from corpora manually annotated at the morphosyntactic level, values of XPOS are normally taken literally from the annotation in these corpora. The only exception concerns "words with spaces": once they are split into separate tokens, each token receives the XPOS value which pertains to the whole multi-token word. For example, in the case of the complex preposition *w sprawie* 'in (the) matter (of)' discussed in Section 5.3 above, both tokens are annotated as `prep:gen`, i.e., a preposition combining with a genitive nominal, even though the preposition *w* alone never combines with the genitive case (it combines with locative – or accusative – nominals), and the token *sprawie* alone should be interpreted as a singular feminine noun (in the locative – or dative – case).[2]

[1] See, e.g., `http://nkjp.pl/poliqarp/help/en.html`.

[2] This is similar to the treatment of some multi-token expressions in the most recent version 2.1 of the English UD treebank, where, e.g., both tokens in *of course* are marked as adverbs.

6.2 UPOS

The coarse parts-of-speech are determined mostly on the basis of the preterminal in the LFG c-structure, rather than on the basis of the detailed morphosyntactic tag in XPOS. The reason for this is that such preterminals make some important distinctions which are not available at the level of the tag. For example, while *jest* 'is' will always be assigned the tag `fin:sg:ter:imperf` (finite imperfective verb in the singular number and third person), the LFG tree will contain information whether it is used as the main verb (the copula or the existential verb) or as the auxiliary (together with certain predicates which do not inflect for person, but analytically inflect for tense).

In some cases also the lemma must be consulted. For example, a noun (i.e., a token with the preterminal SUBST) will normally receive the UPOS value NOUN, but not when the lemma starts with a capital letter (it is a proper noun then, i.e., PROPN)[3] or when it is a nominal pronoun such as KTO 'who', NIC 'nothing', etc. (the right value is PRON). Specific conditions for assigning particular UPOS values are given below.

NOUN

- either the preterminal is SUBST and the lemma does not satisfy conditions for PROPN, PRON or DET (see below),
- or the preterminal is DEPR or GER.

Comments:

- SUBST is the usual preterminal for nouns,
- DEPR stands for derogatory forms of some human-masculine nouns,
- GER stands for gerunds; they are mixed verbal-nominal categories, here marked as nouns to preserve uniformity with other Slavic UD treebanks (as recommended by Dan Zeman, p.c.).

PROPN

- the preterminal is SUBST and the lemma starts with a capital letter.

PRON

- either the preterminal is PPRON12, PPRON3, SIEBIE, or RM,
- or the preterminal is SUBST and the lemma is one of: KTO, KTÓŻ 'who', KTOŚ 'somebody', KTOKOLWIEK 'whoever', NIKT 'nobody', CO, CÓŻ 'what', COŚ 'something', COKOLWIEK 'whatever', NIC 'nothing', TO 'this', TAMTO 'that', WSZYSCY 'all (human)', WSZYSTKO 'all (non-human)'.

[3]Note that the capitalisation of the lemma is independent of the capitalisation of the form of this lemma as it occurs in the text. In particular, even if a common noun is capitalised at the beginning of a sentence (or in an all-caps headline), the lemma of such a capitalised common noun is in the lower case. Hence, a capital letter at the beginning of a lemma is a reasonable indicator of a proper noun.

Comments:

- preterminals PPRON12 and PPRON3 indicate personal pronouns (of the 1st/2nd or 3rd person),
- SIEBIE and RM indicate so-called reflexive pronouns: SIEBIE, which inflects for case, and SIĘ, which does not inflect; while most – but certainly not all – occurrences of SIEBIE are indeed reflexive or reciprocal, most occurrences of the reflexive marker SIĘ are not, and it is not clear whether this word can ever act as an anaphoric pronoun;[4] so, strictly speaking, marking SIĘ as PRON (and assigning it the Reflex=Yes and PronType=Prs features, see below) is simply wrong; the only reason SIĘ is marked as a reflexive personal pronoun in $\text{UD}^{\text{PL}}_{\text{LFG}}$ is to make its annotation uniform with the previous UD treebank of Polish, $\text{UD}^{\text{PL}}_{\text{SZ}}$, and with other UD treebanks of Slavic languages (as recommended by Dan Zeman, p.c.).

NUM

- the preterminal is NUM and the lemma does not satisfy conditions for DET (see below).

ADJ

- either the preterminal is ADJ and the lemma does not satisfy conditions for DET (see below),
- or the preterminal is ADJC, ADJA or ADJP,
- or the preterminal is PPAS or PACT.

Comments:

- ADJC, ADJA and ADJP are special forms of adjectives (distinguished on the basis of the same distinction made in the legacy tagset): those marked as ADJC are used only predicatively (e.g., *zdrów* 'healthy', apart from the regular form *zdrowy*), those marked as ADJA occur in certain adjective—adjective constructions (e.g., *biało* in *biało-czerwony* 'white-and-red'), and those marked as ADJP are only used in certain prepositional constructions (e.g., *po polsku* 'in Polish', as in speaking Polish),
- PPAS and PACT mark passive and active adjectival participles, i.e., mixed verbal-adjectival categories, here marked as adjectives to preserve uniformity with other Slavic UD treebanks (as recommended by Dan Zeman, p.c.).

DET

- either the preterminal is NUM and the lemma is one of: ILE 'how many', ILEŻ 'how many (non-human)', ILUŻ 'how many (human)', TYLE 'so many', MAŁO 'little, few', NIEMAŁO 'not little, not few', MNIEJ 'fewer, less', NAJMNIEJ 'fewest, least', DUŻO 'much, many', NIEDUŻO 'not much, not many', WIELE 'many', NIEWIELE 'not many', WIĘCEJ 'more', NAJWIĘCEJ 'most', KILKA 'several', KILKANAŚCIE 'dozen or so', KILKADZIESIĄT 'several tens', KILKASET 'several hundred', PARĘ 'a few', PARĘNAŚCIE 'dozen or so', PARĘDZIESIĄT 'several tens', NIECO 'some',

[4] See, e.g., Reinhart and Reuland 1991, 1993 on a cross-linguistic analysis of 'reflexive markers' as morphemes reducing the argument structure of the verb, as well as Kupść 1999, Patejuk and Przepiórkowski 2015a and references therein on various functions of SIĘ in Polish.

SPORO ‘considerably many, much’, TROCHĘ ‘some’, ILEŚ ‘some number’, ILEKOLWIEK ‘however much, many’, MNÓSTWO ‘great quantity’,
- or the preterminal is SUBST and the lemma is MNÓSTWO ‘great quantity’,
- or the preterminal is ADJ and the lemma is one of: ów ‘this, that’, TAKI ‘such’, TEN ‘this’, TAMTEN ‘that’, TAKIŻ ‘such’, TENŻE ‘this’, KAŻDY, WSZELKI, WSZYSTEK ‘each, all’, ŻADEN ‘none’, KTÓRY, KTÓRYŻ ‘which’, CZYJ, CZYJŻE ‘whose’, CZYJŚ ‘somebody’s’, CZYJKOLWIEK ‘whosever’, NICZYJ ‘nobody’s’, SWÓJ ‘oneself’s’, MÓJ ‘my’, TWÓJ ‘your.SG’, NASZ ‘our’, WASZ ‘your.PL’, JAKI, JAKIŻ ‘what kind’, PEWIEN ‘certain’, JAKIŚ ‘some’, JAKIKOLWIEK ‘whatever like’, KTÓRYŚ ‘one of which’, KTÓRYKOLWIEK ‘whichever’, NIEJAKI ‘certain’, NIEKTÓRY ‘some’, NIEJEDEN ‘not one’.

ADP

- the preterminal is PREP.

VERB

- the preterminal is FIN, PRAET, INF, IMPS, IMPT, PCON, PANT, BEDZIE, WINIEN or PRED,
- *and* syntactic conversion does not determine that coarse part-of-speech should be AUX.

Comments:

- note that the two conditions should be understood conjunctively, not disjunctively (as in other cases),
- preterminal names in the first condition correspond directly to the grammatical classes (fine-grained parts-of-speech) in the legacy tagset, and they indicate the following verbal forms: finite forms (PRAET – past and FIN – non-past), infinitival (INF), impersonal (IMPS), imperative (IMPT), adverbial participles (PCON and PANT), future forms of BYĆ ‘be’ (BEDZIE), forms of the couple of lexemes behaving like WINIEN ‘ought’ (WINIEN), and predicates which do not inflect for person, but analytically inflect for tense and may act as the head of the sentence (PRED),
- see AUX below and Chapter 7 on the second condition.

AUX

- either the preterminal is AUX, AGLT or MM,
- or the preterminal is one of the verbal classes mentioned in VERB above, but syntactic conversion determines that the coarse part-of-speech should be AUX.

Comments:

- the preterminal AUX is used for the forms of BYĆ ‘be’ which indicate tense – most often in periphrastic future tense, but also in the combination with some quasi-verbal predicates (see PRED in the comment to VERB above) and in the past tense conditional construction, e.g., *Byłby upadł* ‘He would have fallen’, tokenised as *Był by upadł*, lit. ‘was COND fell’,
- AGLT is the preterminal of mobile inflections (see Section 5.1 above),

- MM stands for 'mood markers', i.e., particles expressing the conditional (BY) or the imperative (NIECH, NIECHAJ) mood,
- the coarse part-of-speech AUX is also assigned to various forms of BYĆ, BYWAĆ (two lexemes for 'be' differing in habituality), ZOSTAĆ, ZOSTAWAĆ (two lexemes for 'become' differing in aspect, used in passive constructions) and TO (used as a copula) at the stage where aux:pass and cop dependency relations are established (see Chapter 7, Section 7.2.4), i.e., when:
  - either the token is a form of BYĆ, BYWAĆ, ZOSTAĆ or ZOSTAWAĆ, and it has an outgoing initial (LFG) relation XCOMP-PRED to a token which is a passive participle (in which case XCOMP-PRED is replaced with aux:pass and the direction of the dependency is reversed),
  - or the token is a form of BYĆ, BYWAĆ or TO, and it has an outgoing initial relation XCOMP-PRED or OBL-LOCAT (in which case the relation is replaced with cop and the direction of the dependency is reversed).

ADV

- the preterminal is ADV.

SCONJ

- the preterminal is COMP.

CCONJ

- the preterminal is CONJ or PRECONJ.

Comment:

- the distinction between conjunctions and preconjunctions is preserved at the level of dependency relations (cc vs. cc:preconj).

PART

- the preterminal is QUB, QUB[int], NEG or CNEG.

Comments:

- QUB corresponds to the qub tag in the legacy tagset and indicates a particle,
- two kinds of particles are singled out: the question particle CZY (and its variants CZYŻ and CZYŻBY) – their preterminal is QUB[int] – and the negative particle NIE – with preterminals NEG and CNEG; both kinds are mapped to PART, but distinguished by values of PartType and Polarity in FEATS (see below),
- NEG is the preterminal of the usual sentential (verbal, eventuality) negation, and CNEG – of constituent negation (cf. Przepiórkowski and Patejuk 2015); this distinction is lost in translation.

INTJ

- the preterminal is INTERJ.

PUNCT

- the preterminal is one of: COMMA, PERIOD, POINT, ELLIPSIS, DASH, HYPHEN, LD-QT, LE-QT, LP-QT, (left quote), RD-QT, RE-QT, RP-QT (right quote), L-PRN, L-SQR (left paren), R-PRN, R-SQR (right paren), EXCL-POINT, INT-MARK.

Comment:

- all punctuation marks are mapped into PUNCT, but they are distinguished by values of PunctType and PunctSide in FEATS (see below).

The following universal parts-of-speech are not used in UD^{PL}_{LFG}: SYM, X.

6.3 FEATS

Morphosyntactic features are determined not only on the basis of the preterminal in the LFG c-structure and the lemma, but also to a large extent on the basis of the legacy tag, i.e., the value of XPOS.

6.3.1 Universal features with universal values

Case It is read directly off the legacy tag. Possible values: Nom, Acc, Gen, Dat, Ins, Loc, Voc. Case in the FEATS field indicates the value of case as an *inflectional* feature of the current token. Some treebanks use this feature also as a valency feature, to indicate the case *governed* by a given adposition. In UD^{PL}_{LFG}, this use of Case is relegated to the MISC field – see Section 6.4 below.

Number It is read directly off the legacy tag. Possible values: Sing, Plur.

Number[psor] The same possible values as in the case of Number. The value Sing is assigned in the case of forms of the adjectival possessive pronouns MÓJ 'my' and TWÓJ 'your.SG', and the value Plur – in the case of NASZ 'our' and WASZ 'your.PL'.

Gender Five gender values are assumed for Polish according to the legacy tagset (after Mańczak 1956): three masculine genders, feminine and neuter. In UD^{PL}_{SZ} the three masculine genders were distinguished with the use of the Animacy, but – as discussed below – this solution is untenable. In UD^{PL}_{LFG} we adopt the solution suggested to us by Dan Zeman (p.c.), namely, to retain the standard values of Gender – here: Masc, Fem, Neut – and to distinguish the three

masculine genders via a new language-specific feature, SubGender, described in Section 6.3.3 below.

Degree It is read directly off the legacy tag. Possible values: Pos, Cmp, Sup. In the case of those adjectival forms which are translated to the DET UPOS (see Section 6.2 above) – they are somewhat arbitrarily assigned the positive degree in the legacy tagset – this feature is removed (such words do not inflect for degree anyway).

Person The usual values of the first, second and third person of various verbal and pronominal forms are read off the legacy tag. Additionally, impersonal forms (bearing the preterminal IMPS) are marked as 'zero person'. Hence, possible values: 0, 1, 2, 3.

Aspect It is read directly off the legacy tag. Possible values: Imp, Perf.

Voice This feature has the value Pass in the case of tokens with the preterminal PPAS (i.e., passive participles) and Act in the case of tokens with the following preterminals: PACT (i.e., active participles), FIN, PRAET, INF, IMPS, IMPT, WINIEN, PCON and PANT.

Tense This feature is not explicitly present in the legacy tag, on the assumption that tense is a feature of constructions larger than single tokens. However, it can be inferred from the preterminal (which corresponds to the fine-grained part-of-speech present in the tag) and from the value of aspect, thus:

- if the preterminal is PRAET or IMPS, then Tense=Past,
- if the preterminal is FIN, then:
  - in the case of imperfective aspect, Tense=Pres,
  - in the case of perfective aspect, Tense=Fut,
- if the preterminal is BEDZIE (a future form of BYĆ 'be'), then Tense=Fut,
- if the preterminal is AUX, then the value of Tense is assigned on the basis of the fine-grained part of speech in XPOS:
  - Fut in the case of bedzie,
  - Pres in the case of fin,
  - Past in the case of praet,
- if the preterminal is PRED or WINIEN, then Tense=Pres,
- if the token is an adverbial participle:
  - in the case of PCON (the contemporary participle), Tense=Pres,
  - in the case of PANT (the anterior participle), Tense=Past.

Hence, the possible values of Tense are: Past, Pres, Fut. Note that these values pertain to tokens rather than sentences. For example, the periphrastic future tense is created in Polish with the use of a future form of BYĆ 'be', marked as Tense=Fut, and either the infinitival (preterminal INF) or the preterite (PRAET) form of the verb. In the latter case, the verb is marked as Tense=Past, even though the whole sentence is unequivocally in the future tense.

Mood This feature is also not present in the legacy tag, but all finite forms, as well as some mood markers, are marked with it:

- the mood marker NIECH (and its variant NIECHAJ) is marked as Mood=Imp,
- the mood marker BY is marked as Mood=Cnd,
- verbs bearing the preterminal IMPT, i.e., imperative forms, are marked as Mood=Imp,
- all other finite verbs (i.e., verbs with the VerbForm=Fin feature, see immediately below) are marked as Mood=Ind.

Hence, the possible values of Mood are: Ind, Cnd, Imp. As in the case of Tense, Mood should be understood as a feature of tokens, and not (necessarily) clauses.

VerbForm Possible values are:

- Fin – in the case of tokens bearing the following preterminals: FIN (present or future forms, depending on aspects), PRAET (past forms), IMPS (impersonal forms), IMPT (imperative forms), AUX (auxiliaries), BEDZIE (future forms of BYĆ 'be'), WINIEN and PRED,
- Inf – in the case of tokens with the preterminal INF,
- Vnoun – in the case of gerunds, i.e., tokens with the preterminal GER; note that their UPOS is NOUN,
- Part – in the case of adjectival participles: passive (preterminal PPAS) and active (preterminal PACT); note that their UPOS is ADJ,
- Conv – in the case of adverbial participles: contemporary (preterminal PCON) and anterior (preterminal PANT); their UPOS is VERB.

PronType The value of PronType is determined on the basis of the lemma and – to a lesser extent – the preterminal in the LFG tree. Note that tokens with PronType values are not limited to those with the UPOS value of PRON; this feature may also be present on tokens marked as DET or ADV, and – in one particular case – on SCONJ:

- PronType=Prs occurs with tokens bearing one of the following preterminals in the LFG tree: PPRON12, PPRON3, SIEBIE and (only for reasons of cross-linguistic consistency) RM (all get the PRON UPOS), as well as in the case of ADJ with the following lemmata: MÓJ 'my', TWÓJ 'your.SG', NASZ 'our', WASZ 'your.PL', SWÓJ 'oneself's' (they get the DET UPOS),
- PronType=Dem occurs with the following lemmata: TO 'this', TAMTO 'that' (both with the PRON UPOS), ÓW 'this/that', TEN, TENŻE 'this', TAMTEN 'that', TAKI, TAKIŻ 'such', TYLE 'so many' (all DET), TAK 'so', TU, TUTAJ 'here', TAM 'there', ÓWDZIE 'in that place', STĄD 'from here', STAMTĄD 'from there', TĘDY 'through there', TAMTĘDY 'through there', WTEDY, WÓWCZAS, WTENCZAS 'then', ODTĄD 'from now/then', DOTĄD 'until now/then', DLATEGO 'for this reason, therefore' (all ADV),
- PronType=Ind occurs with the following lemmata: COŚ 'something', COKOLWIEK 'whatever', KTOŚ 'somebody', KTOKOLWIEK 'whoever' (all PRON), MAŁO 'little, few', NIEMAŁO 'not little, not few', MNIEJ 'fewer, less', NAJMNIEJ 'fewest, least', DUŻO 'much, many', NIEDUŻO 'not much, not many', WIELE 'many', NIEWIELE 'not many', WIĘCEJ 'more', NAJWIĘCEJ 'most', KILKA 'several', KILKANAŚCIE 'dozen or so', KILKADZIESIĄT 'several tens', KILKASET 'several hundred', PARĘ 'a few', PARĘNAŚCIE 'dozen or so', PARĘDZIESIĄT 'several ten', NIECO 'some',

SPORO ‘considerably many, much’, TROCHĘ ‘some’, ILEŚ ‘some number’, ILEKOLWIEK ‘however much, many’, MNÓSTWO ‘great quantity’, PEWIEN ‘certain’, JAKIŚ ‘some’, JAKIKOLWIEK ‘whatever like’, KTÓRYŚ ‘one of which’, KTÓRYKOLWIEK ‘whichever’, NIEJAKI ‘certain’, NIEKTÓRY ‘some’, NIEJEDEN ‘not one’, CZYJŚ ‘somebody’s’, CZYJKOLWIEK ‘whosever’ (all DET), DOKĄDŚ ‘to somewhere’, DOKĄDKOLWIEK ‘to whatever place’, SKĄDŚ ‘from somewhere’, SKĄDKOLWIEK ‘from whatever place’, GDZIEŚ ‘somewhere’, GDZIEKOLWIEK ‘wherever’, JAKOŚ ‘in some way’, JAKKOLWIEK ‘in whatever way’, KIEDYŚ ‘sometime’, KIEDYKOLWIEK ‘whenever’, KTÓRĘDYŚ ‘some way’, KTÓRĘDYKOLWIEK ‘whichever way’, NIEKIEDY ‘sometimes’, GDZIENIEGDZIE ‘in some places’ (all ADV),

- PronType=Neg occurs with the following lemmata: NIKT ‘nobody’, NIC ‘nothing’ (both PRON), ŻADEN ‘none’, NICZYJ ‘nobody’s’ (both DET), NIGDY ‘never’, NIGDZIE ‘nowhere’ (both ADV),
- PronType=Tot occurs with the following lemmata: WSZYSCY ‘all (human)’, WSZYSTKO ‘all (non-human)’ (all PRON), KAŻDY ‘each’, WSZELKI, WSZYSTEK ‘each, all’ (all DET), ZAWSZE ‘always’, WSZĘDZIE ‘everywhere’, ZEWSZĄD ‘from everywhere’ (all ADV),
- the following lemmata are marked either as PronType=Int or PronType=Rel, depending on the value of the TYPE attribute (INT or REL) in the LFG f-structure corresponding to (the preterminal of) the token: KTO ‘who’, CO ‘what’ (both PRON), ILE ‘how many’, JAKI ‘what kind’, KTÓRY ‘which’ (all DET), GDZIE ‘where’, KIEDY ‘when’ (both ADV),
- PronType=Int also occurs with the following lemmata: KTÓŻ ‘who’, CÓŻ ‘what’ (both PRON), ILEŻ ‘how many (non-human)’, ILUŻ ‘how-many (human)’, JAKIŻ ‘what kind’, KTÓRYŻ ‘which’, CZYJ, CZYJŻE ‘whose’ (all DET), DLACZEGO, DLACZEGOŻ, DLACZEGÓŻ, CZEMU, CZEMUŻ ‘why’, DOKĄD, DOKĄDŻE ‘where to’, SKĄD, SKĄDŻE ‘where from’, ODKĄD ‘since when’, JAK, JAKŻE ‘how’, KTÓRĘDY, KTÓRĘDYŻ ‘which way’, GDZIEŻ ‘where’, KIEDYŻ ‘when’ (all ADV),
- additionally PronType=Rel is also assigned to tokens whose preterminal is RSM, i.e., to the complementiser CO used in so-called resumptive relative clauses; as a complementiser, this token is marked with the SCONJ UPOS.

Hence, the possible values of PronType are: Prs, Dem, Ind, Int, Rel, Neg, Tot. Note that the previous UD treebank of Polish, $\text{UD}^{\text{PL}}_{\text{SZ}}$, did not disambiguate between Int and Rel – tokens which could be either bore the PronType=Int,Rel annotation. In the case of the current $\text{UD}^{\text{PL}}_{\text{LFG}}$, the input LFG structures disambiguate such pronouns, so no tokens bear the disjunctive PronType=Int,Rel specification.

(See also Section 6.3.3 below on the language-specific feature Emphatic used to distinguish emphatic forms such as KTÓŻ ‘who’ from neutral forms such as KTO ‘who’.)

NumType All tokens with the NUM UPOS are marked as NumType=Card, i.e., only run-of-the-mill cardinal numerals bear the NUM UPOS. Other tokens treated as numerals in the LFG tree (i.e., with the NUM preterminal) get the DET UPOS and the following values of NumType:

- also Card – tokens with lemmata ILE ‘how many’, ILEŻ ‘how many (non-human)’, ILUŻ ‘how many (human)’, TYLE ‘so many’, MAŁO ‘little, few’, NIEMAŁO ‘not little, not few’, MNIEJ ‘fewer, less’, NAJMNIEJ ‘fewest, least’, DUŻO ‘much, many’, NIEDUŻO ‘not much, not many’, WIELE ‘many’, NIEWIELE ‘not many’, WIĘCEJ ‘more’, NAJWIĘCEJ ‘most’, KILKA ‘several’, KILKANAŚCIE ‘dozen or so’, KILKADZIESIĄT ‘several tens’, KILKASET ‘several hundred’, PARĘ ‘a few’, PARĘNAŚCIE ‘dozen or so’, PARĘDZIESIĄT ‘several tens’, NIECO ‘some’, SPORO ‘considerably many,

much', TROCHĘ 'some', ILEŚ 'some number', ILEKOLWIEK 'however much, many', MNÓSTWO 'great quantity',
- `Frac` – tokens with lemmata PÓŁ 'half' and ĆWIERĆ 'quarter'.

Possible values: `Card` and `Frac`.

`Poss` This feature has a single value, `Yes`, and is assigned to forms of the following lemmata: CZYJ, CZYJŻE 'whose', CZYJŚ 'somebody's', CZYJKOLWIEK 'whosever', NICZYJ 'nobody's', SWÓJ 'oneself's', MÓJ 'my', TWÓJ 'your.SG', NASZ 'our', WASZ 'your.PL' (all with the `DET` UPOS).

`Reflex` This is another feature with `Yes` as the single possible value, and it is assigned to forms of SIEBIE 'self' (their preterminal is `SIEBIE`), SIĘ (so-called 'reflexive marker', its preterminal is `RM`) and SWÓJ 'oneself's'. As discussed above (Section 6.2, `PRON`), the marking of SIĘ as `Reflex=Yes` is usually – perhaps always – linguistically wrong (cf. fn. 4 on page 101), and it is only motivated by consideration of uniformity.

`Variant` Possible values of this feature are: `Short` and `Long`. It is used in four situations:

- to mark those (usually shorter) forms of some adjectives which may be used only predicatively (i.e., which have the `ADJC` preterminal), e.g., *zdrowy* 'healthy' (attributive or predicative) vs. *zdrów* 'healthy' (only predicative): predicative-only forms are marked as `Variant=Short`, neutral forms are not marked with `Variant`,
- to distinguish between two forms of some pronouns: the shorter – usually not accentable, i.e., not accepting emphatic stress – form (`Variant=Short`, e.g., *go* 'him.ACC/GEN') and the longer – accentable – form (`Variant=Long`, e.g., *jego* 'him.ACC/GEN'),
- to distinguish between two forms of some prepositions: the shorter dictionary form (`Variant=Short`, e.g., *z* 'from, with') and the form with the additional final vowel (`Variant=Long`, e.g., *ze*); prepositions which have only one form do not have the `Variant` feature,
- to distinguish between the short form of a mobile inflection (see Section 5.1 above), e.g., *-m* '1SG' (`Variant=Short`), and the form with the preceding vowel, e.g., *-em* '1SG' (`Variant=Long`).

`Hyph` The single value of this feature is `Yes`, and it is used in the case of those forms of adjectives (bearing the preterminal `ADJA`) which only occur in certain adjective–adjective constructions (and are followed by a hyphen, cf. `ADJ` in Section 6.2).

`PrepCase` The two possible values of this feature are: `Pre` (for forms which must occur with an adposition) and `Npr` (for forms which cannot occur with an adposition). It is used in $\text{UD}^{\text{PL}}_{\text{LFG}}$ in two situations:

- to mark those adjectival forms which only occur with prepositions, i.e., tokens with the `ADJP` preterminal (cf. `ADJ` in Section 6.2); such forms are marked as `PrepCase=Pre`, and other forms of adjectives are not marked with `PrepCase` at all,

- to distinguish post-prepositional forms of some pronouns (`PrepCase=Pre`) from forms which cannot be arguments of prepositions (`PrepCase=Npr`); in this case the value of `PrepCase` is read directly off the legacy tag, where the value of post-prepositionality is marked as `praep` or `npraep`.

Note that while in Polish the locative case is governed by prepositions only, locative nouns are not marked with `PrepCase`.

`AdpType` As almost all Polish adpositions are prepositions, almost all tokens with the `ADP` UPOS (or, equivalently, with the `PREP` preterminal in the LFG tree) are marked as `AdpType=Prep`, with the only exception made for the postposition TEMU 'ago' (as in *dwa lata temu* 'two years ago'), marked as `AdpType=Post`.

`Polarity` Two possible values of this feature are: `Pos` (affirmative polarity) and `Neg` (negative polarity). In $\text{UD}^{\text{PL}}_{\text{LFG}}$ it is used in two situations:

- trivially: to mark the negative particle, NIE, as `Polarity=Neg`,
- less trivially: to mirror the negation feature of the legacy tagset, appropriate for gerunds and adjectival participles; in this case the feature is read directly off the legacy tag: `aff` is translated into `Polarity=Pos`, and `neg` – into `Polarity=Neg`.

`PunctType` and `PunctSide` The values of these features reflect different preterminals of punctuation marks in the LFG c-structure:

- in the case of `COMMA`, the value of `PunctType` is `Comm`,
- in the case of `PERIOD`, the value of `PunctType` is `Peri`,
- in the case of `EXCL-POINT`, the value of `PunctType` is `Excl`,
- in the case of `INT-MARK`, the value of `PunctType` is `Qest`,
- in the case of `DASH` and `HYPHEN`, the value of `PunctType` is `Dash`,
- in the case of `LD-QT`, `LE-QT` and `LP-QT`, the value of `PunctType` is `Quot` and the value of `PunctSide` is `Ini`,
- in the case of `RD-QT`, `RE-QT`, `RP-QT`, the value of `PunctType` is `Quot` and the value of `PunctSide` is `Fin`,
- in the case of `L-PRN` and `L-SQR`, the value of `PunctType` is `Brck` and the value of `PunctSide` is `Ini`,
- in the case of `R-PRN` and `R-SQR`, the value of `PunctType` is `Brck` and the value of `PunctSide` is `Fin`.

Punctuation marks with preterminals `ELLIPSIS` and `POINT` are not assigned any values of `PunctType` or `PunctSide`. Hence, the possible values of `PunctType` are: `Comm`, `Peri`, `Excl`, `Qest`, `Quot`, `Brck`, and the possible values of `PunctSide` are: `Ini`, `Fin`.

6.3.2 Universal features with language-specific values

VerbType This universal feature is used in $\text{UD}^{\text{PL}}_{\text{LFG}}$ with a single language-specific value, `Quasi`. This is a way of marking those tokens which are like verbs in constituting the centre of a clause and being able to conjugate analytically for tense, but which – unlike typical verbs – do not take nominative subjects, do not inflect for person, etc. In the Polish structuralist tradition they are called *czasowniki niewłaściwe* 'quasi-verbs' (Saloni 1974; Saloni and Świdziński 1985) and they bear the detailed part of speech `pred` in the legacy tagset, as well as the preterminal `PRED` in the LFG tree.

PartType The only – language-specific – value of this feature is `Int` (for 'interrogative'), and it is used to mark the question particle CZY 'if, whether' and its variants CZYŻ and CZYŻBY, as well as the archaic AZALIŻ.

Polite This universal feature is used in $\text{UD}^{\text{PL}}_{\text{LFG}}$ with a single language-specific value, `Depr`, used to mark derogatory forms for human-masculine nouns, as in: *profesorowie* 'professors' (neutral) vs. *profesory* 'professors' (derogatory). This feature is a direct translation of the `DEPR` preterminal (and the `depr` detailed part of speech in the legacy tagset).

6.3.3 Language-specific features

SubGender As mentioned above (cf. `Gender` in Section 6.3.1), at least five genders are distinguished in contemporary Polish linguistics since Mańczak 1956,[5] with three masculine genders, `m1`, `m2` and `m3`, sometimes called human-masculine (or virile), animate-masculine and inanimate-masculine. As there is some correlation between these three genders and the semantic feature of animacy, $\text{UD}^{\text{PL}}_{\text{SZ}}$ modelled this three-way distinction within the masculine gender via the `Animacy` feature and its three values: `Hum` (for `m1`), `Nhum` (for `m2`) and `Inan` (for `m3`). This solution is linguistically unsatisfactory, as there are many well-known cases of broken correlation between masculine subgender and animacy. For example, many inanimate nouns bear the `m2` gender (marked as `Animacy=Nhm`, i.e., non-human animate, in $\text{UD}^{\text{PL}}_{\text{SZ}}$), including masculine names of dances (e.g., WALC 'waltz' and FOKSTROT 'foxtrot'), but also, e.g., TRUP 'corpse' (which, rather than being non-human and animate, is human and non-animate!) and various derogatory terms for women (i.e., human animate entities), including BABSZTYL 'hag'.

As the existence of at least five gender values is widely accepted in Polish formal linguistics, in newer dictionaries (e.g., Bańko 2000) and in virtually all Polish corpora, we do not take the step back of approximating the three masculine genders with `Animacy`. The most straightforward representation would be to define language-specific values of the universal `Gender` feature: `Masc1`, `Masc2` and `Masc3` (all instead of the universal `Masc`), apart from the standard `Fem` and `Neut`. Instead, in order to maximise uniformity across UD treebanks, in $\text{UD}^{\text{PL}}_{\text{LFG}}$ we adopt the solution suggested to us by Dan Zeman (p.c.), namely, to retain the standard values of `Gender`

[5] With some proposals to further extend this repertoire of gender values, e.g., to nine (Saloni 1976; Saloni and Świdziński 1985).

– here: `Masc`, `Fem`, `Neut` – and to distinguish the three masculine genders via a new language-specific feature, `SubGender`. Hence, possible values of this feature are: `Masc1`, `Masc2` and `Masc3`.

`Emphatic` This language-specific feature marks those forms of pronominals (some of which receive the UPOS value of `DET` or `ADV`) which are made emphatic by the particle *-ż(e)*, e.g.: CÓŻ 'what' (vs. the neutral CO), ILEŻ 'how many' (vs. ILE), CZYJŻE 'whose' (vs. CZYJ), etc. The only value of this feature is `Yes`.

`Agglutination` Some preterite verbal forms have two variants: one for expressing 3rd person singular masculine, e.g., *mógł* 'could.PAST.3SG.M', and another for combining with the adjacent 1st or 2nd person singular masculine mobile inflection (see Section 5.1), e.g., *mogł* as in *mogłem* 'could.PAST.3SG.M'. The feature distinguishing such forms in the legacy tagset is called *aglutynacyjność* 'agglutination', and no attempt is made here to translate this name for the purpose of $\text{UD}^{\text{PL}}_{\text{LFG}}$. Possible values are: `Agl` (e.g., for *mogł*) and `Nagl` (e.g., for *mógł*).

6.4 MISC

`SpaceAfter` As discussed in Section 5.1 above, this feature indicates that there is no space between the current and the next token. It is a universal feature with the single possible value `No`.[6]

`Case` This feature is present on adpositions and indicates the case governed by the adposition. It is read directly off the legacy tag. Possible values: `Nom`, `Acc`, `Gen`, `Dat`, `Ins`, `Loc`, `Voc`.[7] This information might seem to be redundant (repeats `Case` value of the head noun); three constructions where it is not are:

- when two prepositions governing different cases are coordinated and combined with a single nominal dependent satisfying the requirement of the closest preposition, as in *przed i po śniadaniu* 'before and after breakfast.LOC', where the locative noun satisfies the case government of the closest preposition, *po* 'after', but not that of the further preposition, *przed* 'before', which governs the instrumental case here (such constructions do not currently occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$);
- when the preposition is stranded, as in *ten lek zażyj przed jedzeniem, a ten – po* 'take this medicine before the meal, and this – after' (again, such constructions do not occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$);
- when a preposition governing the accusative case combines with a numeral phrase consisting of an accusative numeral and a genitive noun; according to the UD guidelines – and contrary to morphosyntactic tests on headedness – the head of such a construction is the

[6] `http://universaldependencies.org/format.html`

[7] It may seem surprising that all Polish cases are listed here, including nominative and vocative, but the analysis of some Polish functional words as prepositions which may combine with the nominative case is well-established in Polish linguistics (Kallas 1986, 1995), and at least *per* may be treated as a preposition combining with the vocative case (apart from the nominative), as in the attested: *Robert Górski: ludzie na ulicy mówią do mnie per „Panie Premierze"* 'Robert Górski: people on (the) street talk to me *per* mister.VOC prime_minister.VOC'.

genitive noun and the preposition is its dependent, so a straightforward algorithm would determine that the preposition governs the genitive case, contrary to fact (such constructions are relatively numerous[8] in $\text{UD}^{\text{PL}}_{\text{LFG}}$).

DepType This feature corresponds to the legacy tagset feature *akomodacyjność* 'accommodability' (introduced originally in Bień and Saloni 1982), specific for numerals (some of which are re-analysed as determiners in $\text{UD}^{\text{PL}}_{\text{LFG}}$), which either agree with the noun they combine with (`DepType=Congr`) or require the noun to be in the genitive case (`DepType=Rec`). Hence, the two values of this feature are `Congr` and `Rec`.

[8] See, for example, (8.1) on p. 181, where the preposition *na* 'for' in *na 48 godzin* 'for 48.ACC hours.GEN' governs the accusative case, present on the numeral, but according to UD guidelines the `case` dependency targeting this preposition originates in the genitive noun, as shown in Figure 8.1 on p. 181.

Chapter 7

Syntax

The conversion of syntactic structures from LFG representations – i.e., constituent structures and functional structures – to UD v.2 representation is performed in two stages. First, LFG structures are converted to dependency structures in a maximally conservative way, i.e., respecting headedness information in c-structures and names of dependencies in f-structures; in this monograph, such LFG-like dependency structures are called 'initial dependency representations'. Second, such initial dependency representations are converted to 'final UD representations', i.e., dependency structures satisfying UD v.2 guidelines.

Let us consider the following sentence:

(7.1) Pracodawca musi też płacić wszelkie podatki i ubezpieczenia.
employer.NOM.SG must also pay.INF all.ACC.PL taxes.ACC.PL and insurance.ACC.PL
'The employer must also pay any taxes and insurance.'

Its c-structure and simplified (only PRED values and grammatical functions) f-structure are given in Figures 7.1–7.2. The form *musi* 'must' is the clear root of the sentence: it is the head of the syntactic tree in Figure 7.1 and also its preterminal, FIN, projects to the whole functional structure in Figure 7.2. According to the initial dependency representation in Figure 7.3, there are two dependents of *musi*, with dependency labels read directly off the f-structure: the SUBJ(ect) *pracodawca* 'employer' and the XCOMP headed by *płacić* 'pay'. As the substructure marked in Figure 7.2 with the index 57 shows, the SUBJect of *płacić* is the same as the SUBJect of *musi* – it is the substructure with the index 52. Hence, the initial dependency structure in Figure 7.3 is not a tree. The infinitival verb *płacić* has a simple ADJUNCT (the value of this attribute in the f-structure is a singleton set) and a more complex OBJect – the value of this attribute is a coordinate structure. In LFG, such coordinate structures are represented by sets. In this case this is a two-element set with the two elements corresponding to the two nouns, *podatki* 'taxes' and *ubezpieczenia* 'insurance policies'. The whole coordinate structure corresponds to the preterminal of the conjunction *i* 'and', so – on this analysis – coordination is headed by the conjunction. Hence, the OBJ dependency leaving the verb *płacić* goes to the conjunction *i* rather than to any of the conjuncts; the conjuncts are direct dependents of the conjunction, with the dependency name arbitrarily chosen as CONJ.[1] The f-structure in Figure 7.2 also

[1]There is no attribute CONJ in the f-structure; instead, the conjuncts are elements of the set corresponding to the conjunction.

contains the explicit information that the adjunct *wszelkie* 'all, any' is shared between the two conjuncts (see the substructure with the index 21). Hence, in the initial dependency structure in Figure 7.3 there are two incoming edges at *wszelkie*, from the two conjuncts, which again makes this structure a non-tree. Apart from these principled dependencies, there is also a 'technical' dependency edge from the root of the clause to the final period. In fact, the root of the sentence – the verb *musi* 'must' – and the final punctuation are co-heads of this sentence in the sense that their c-structure preterminals, `FIN` and `PERIOD`, both map to the whole f-structure. While in this case it is obvious which of these two co-heads is the true head of the sentence, the procedure of selecting the true head from a set of co-heads is non-trivial, as discussed in Section 7.1 below. The result of the first stage of the conversion process for the example sentence is given in Figure 7.3.

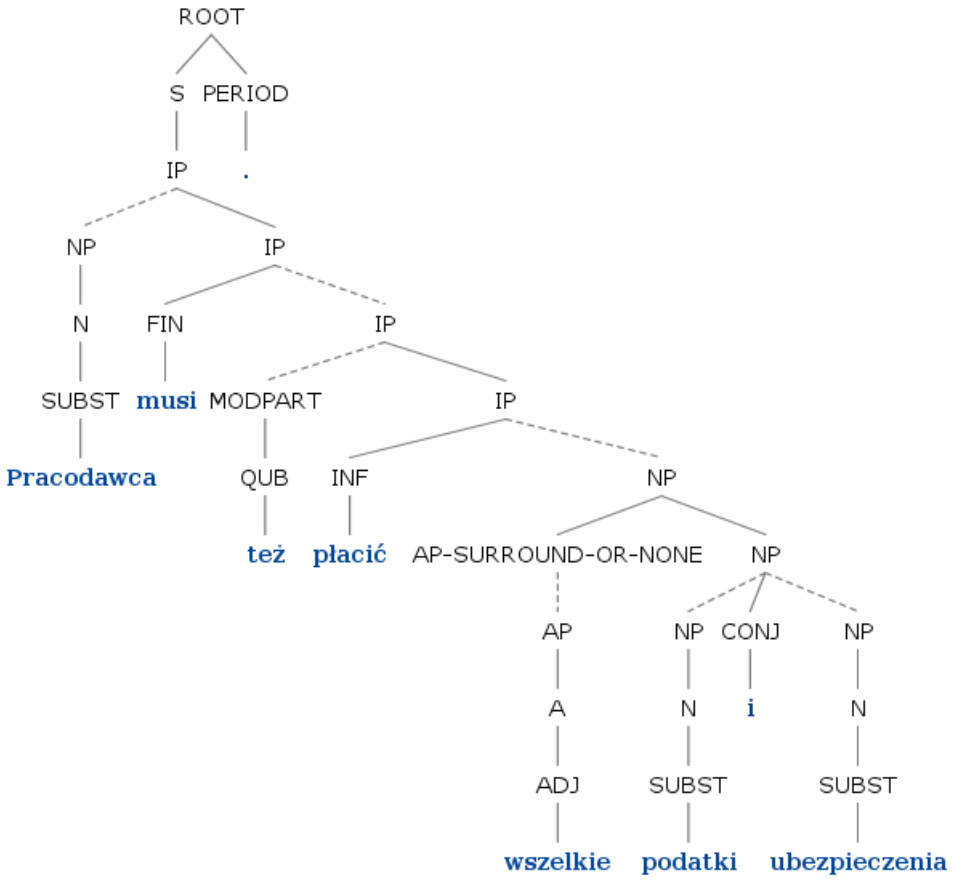


Figure 7.1: C-structure of (7.1)

At the second stage, this initial dependency representation is converted to the final UD representation in Figure 7.4. This final representation consists of a basic dependency tree, drawn above the input tokens, and the enhanced dependency graph, drawn below the sentence. (Whenever the enhanced dependency representation is identical to the basic tree, only the latter will be shown in such diagrams.) There are many differences between the initial and final structures. First of all, dependencies are renamed to those used in UD. In the case of OBJ and XCOMP, it is only a matter of changing capital letters to lower case letters, in the case of PERIOD, it is a matter of renaming it deterministically to `punct`, but in many other cases the change of the dependency name is much less trivial, and may depend on a number of factors.

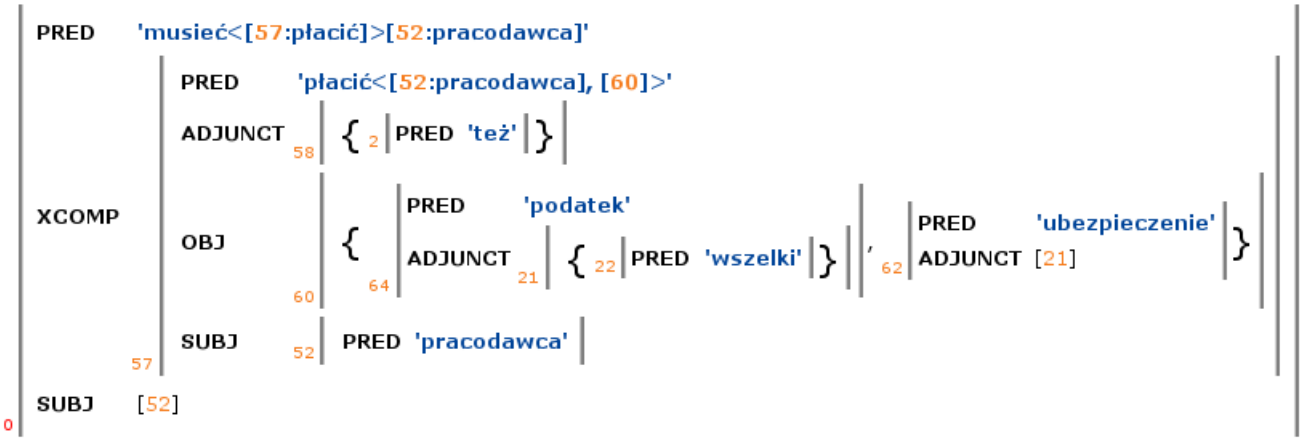


Figure 7.2: Schematic f-structure of (7.1)

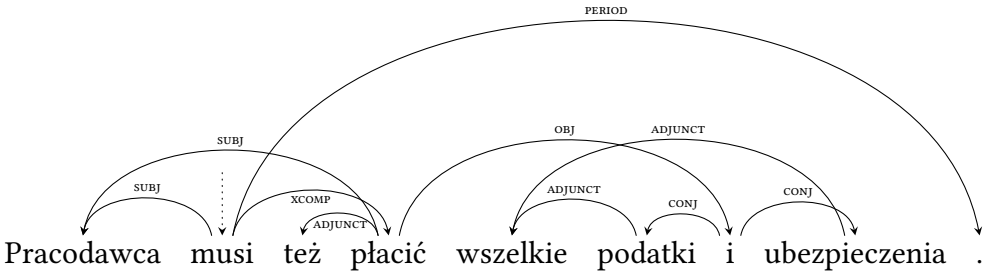


Figure 7.3: Initial dependency representation of (7.1)

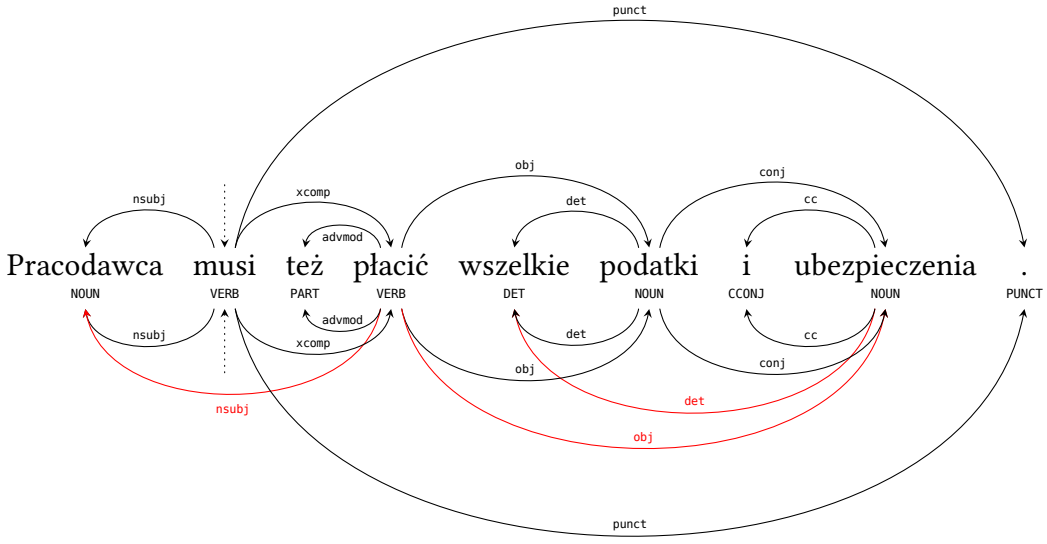


Figure 7.4: Final UD representation of (7.1)

Also the direction of some relations changes, most notably in the case of coordination, but also in many cases of constructions consisting of a function word and a content word, which in LFG are analysed as headed by the function word, but in UD are assumed to be headed by the content word. Returning to coordination, in UD it is assumed to be headed by the first conjunct – rather than by the conjunction – so the `obj` relation from *płacić* 'pay' goes to *podatki* 'taxes', rather than to *i* 'and'.[2] Since the basic UD structure is a tree, there may be only one incoming edge to *pracodawca*, rather than the two present in the initial dependency structure; the other one is only present in the enhanced dependency graph (and marked in red here, as are all dependency relations absent in the basic dependency tree). Similarly for the shared ADJUNCT – renamed to `det` in compliance with UD guidelines – *wszelkie* 'all, any'. The third enhanced dependency absent in the basic tree is the `obj` dependency from *płacić* 'pay' to *ubezpieczenia* 'insurances' – according to UD guidelines, if a coordinate structure as a whole is a dependent of some head, all conjuncts in this structure should be enhanced dependents of this head. These are only some of many possible differences between the initial dependency structure and the final UD representation (and there are also a few more possible differences between the basic UD tree and the enhanced UD representation) – they are discussed in more detail in Section 7.2.

In the remainder of this chapter, particular steps of the conversion procedure are often illustrated with dependency structures intermediate between the initial LFG structures and the final UD structures. In such cases, the target UD structures are provided in Appendix C.

7.1 From LFG to initial dependencies

At this first stage, two dependency structures are created. The first, illustrated in Figure 7.3 above, represents the complete dependency information present in LFG representations. This structure does not have to be a tree (it is not in this figure), and it is the input to the conversion into the final enhanced UD representation. The procedure of arriving at such initial dependency structures is described in Sections 7.1.1–7.1.3.

The second structure, shown in Figure 7.5 below, is derived from this complete representation, but it is reduced to a tree in a way described in Section 7.1.4. This tree representation is the input to the conversion into the final basic UD representation.

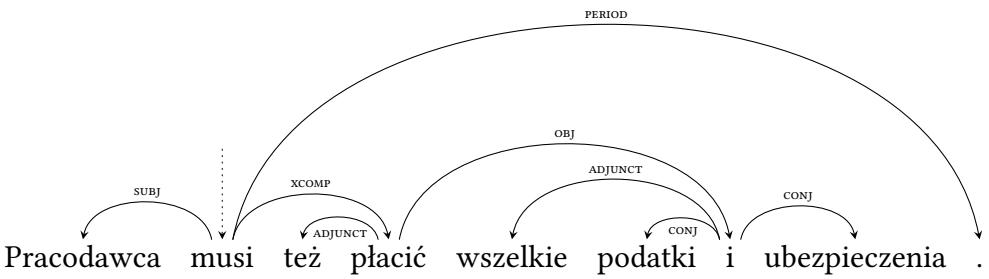


Figure 7.5: Initial dependency representation of (7.1) – basic tree

[2]Other structural differences between the two treatments of coordination are discussed below.

7.1.1 Finding true heads

Consider the following sentence and its LFG representations in Figures 7.6 and 7.7:[3]

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

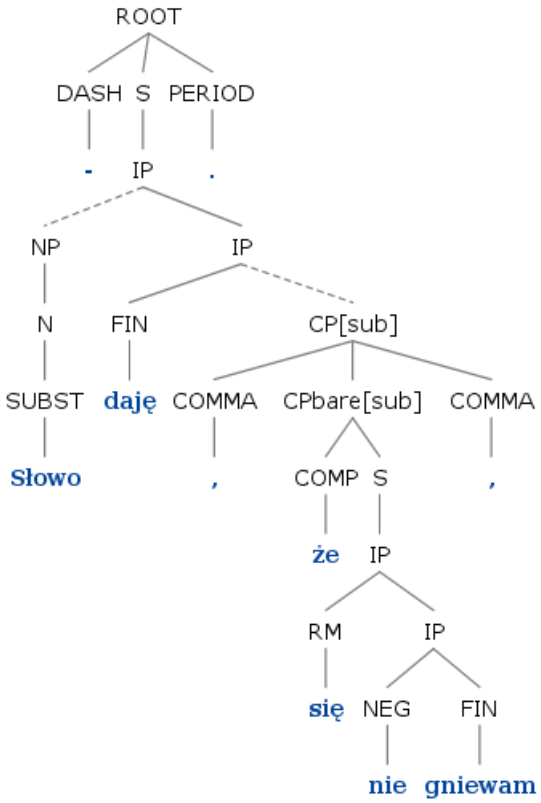


Figure 7.6: C-structure of (7.2)

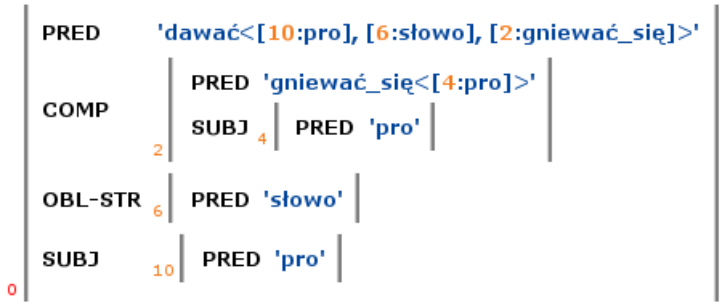


Figure 7.7: Schematic f-structure of (7.2)

There are ten tokens (leaves) in the c-structure and only five feature structures (marked with indices 0, 2, 4, 6, 10) in the f-structure, two of which (4 and 10) do not correspond directly to any

[3] RM stands for a 'reflexive marker', here an inherent part of the verb GNIEWAĆ SIĘ 'be angry', and NEG – for the marker of negation.

of the tokens in the sentence, but rather represent the *pro*-dropped subjects of the two verbs. Hence, the preterminals of the ten tokens will be mapped to only three feature structures:

- the initial DASH, the FIN of the main verb and the final PERIOD are mapped to the main feature structure (with index 0),
- the SUBST preterminal of *słowo* 'word' is mapped to the value of OBL-STR (the structure with index 6),
- the other six preterminals are all mapped to the value of COMP (the structure with index 2).

The f-structure makes two dependency relations available (apart from the SUBJ dependencies which do not end in an actual token): OBL-STR and COMP. Which tokens are related by these dependencies?

Let us assume for the time being that, in answering this question, we disregard punctuation (but, as we will see below, this is not a safe assumption). If so, it is clear that OBL-STR is the label of a dependency edge from the main verb, *daję* 'give' (preterminal FIN), to *słowo* 'word' (preterminal SUBST). Also the COMP dependency starts with *daję* 'give', but it is not immediately clear where it ends: *że* (preterminal COMP), *się* (RM), *nie* (NEG) or *gniewam* (FIN)? That is, which of the co-heads – the four tokens (six – if the two surrounding commas are included) mapping into the same feature structure – should be chosen as the real head?

The basic algorithm is simple, but is complicated by the fact – to be discussed in more detail below – that in some cases a comma may be the head of a coordinate structure:

- if there is a verbal token among the co-heads, select it as the true head; a verbal token is defined here as having one of the following preterminals: FIN, PRAET, INF, IMPS, IMPT, PRED, WINIEN, BEDZIE, PCON, PANT; note that this clause immediately selects the two FIN verbs, *daję* and *gniewam*, as the true heads in the respective sets of co-heads;
- otherwise, if there is a nominal or adjectival token, it is the true head; this concerns tokens with the following preterminals: SUBST, DEPR, GER, PPRON12, PPRON3, SIEBIE, NUM, ADJ, PACT, PPAS;
- otherwise, if there is an explicit – lexical – conjunction (preterminal CONJ), it is the head; note that in the case of discontinuous conjunctions (as in the English *either... or...*), only the final part of the conjunction is marked as CONJ, and the initial – as PRECONJ.

If none of the above three conditions is met, the most common situation is that all co-heads are punctuation marks. Usually punctuation marks are co-heads with real words, i.e., they fall under one of the three cases above, but it is also possible that there is no lexical conjunction and a comma acts as the conjunction, as in the following example:

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

Here, the comma acting as the conjunction is a co-head with the final period, so we need a rule selecting the comma, not the period, as the true head of the whole sentence. The preliminary version of this additional clause is:

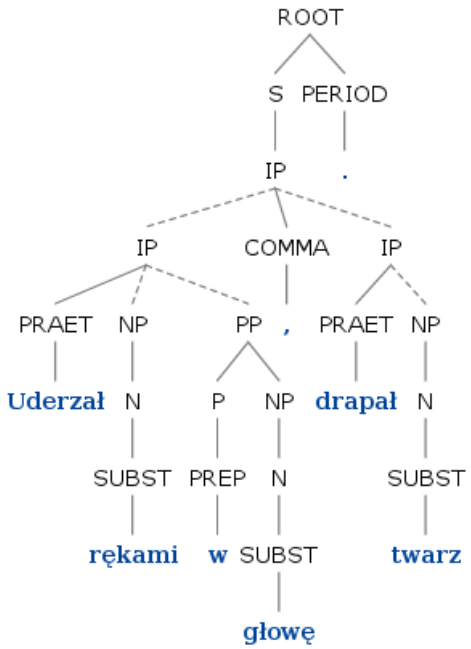


Figure 7.8: C-structure of (7.3)

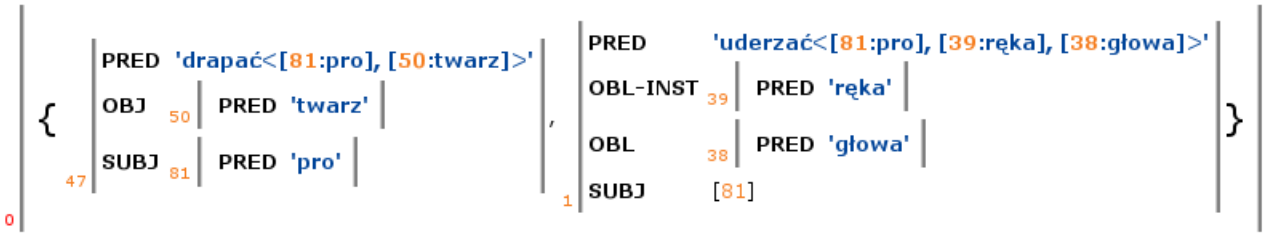


Figure 7.9: Schematic f-structure of (7.3)

- otherwise, select the final (linearly rightmost) comma as the true head.

Note that this formulation properly takes care of cases of more than two conjuncts: as all adjacent conjuncts will normally be separated by commas, there will be more than one comma in the co-head set, but only the final one should be selected as the true conjunction. Moreover, this clause disregards any other punctuation marks, as there may be many of them in the co-head set: an initial dash, a number of sentence-ending punctuation marks (e.g., *?!* or the ellipsis, …, written as three periods), etc.

There are two complications, though, both illustrated with the following example:

(7.4) Wydawało się, że wojna jednak go przerosła, przeraziła.
seemed.3SG.N RM COMP war.NOM.SG.F after_all him.ACC overwhelmed.3SG.F scared.3SG.F
'It seemed that after all the war overwhelmed and scared him.'

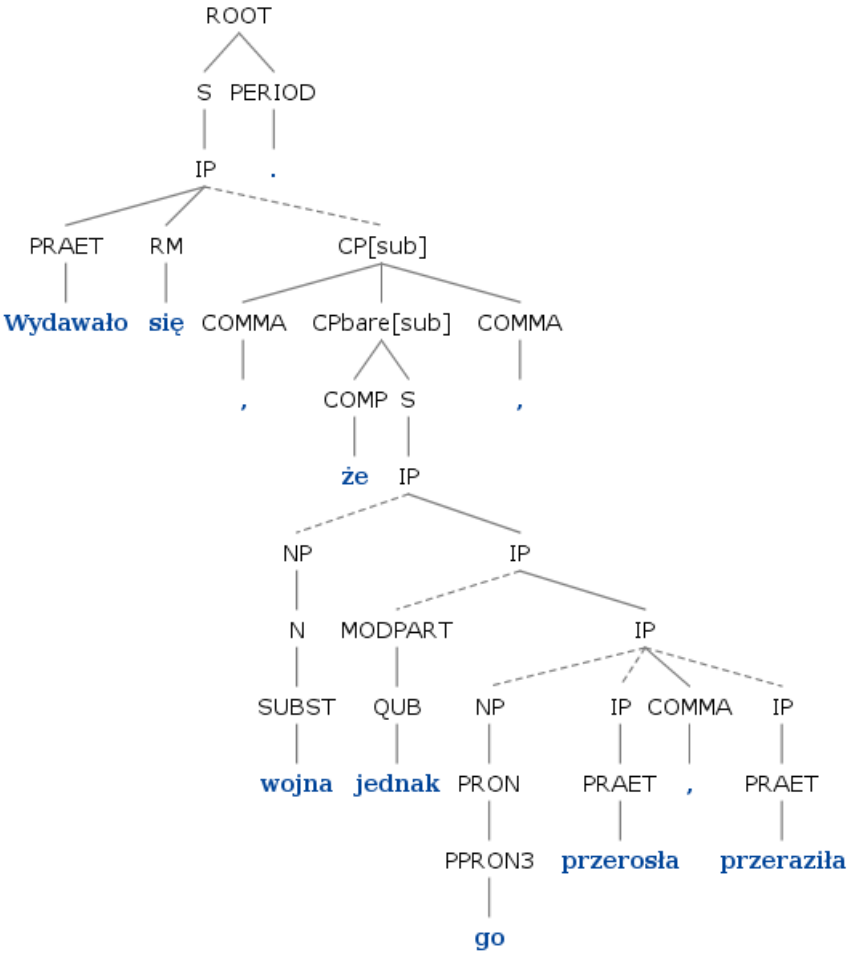


Figure 7.10: C-structure of (7.4)

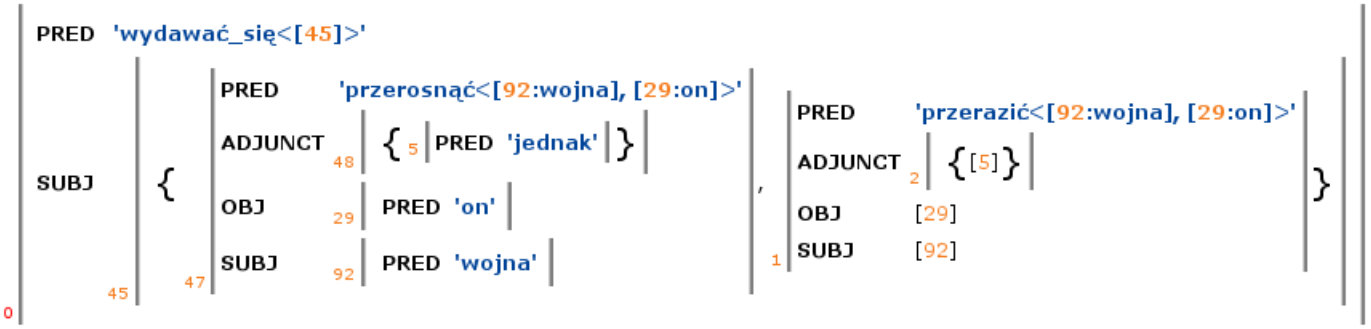


Figure 7.11: Schematic f-structure of (7.4)

First of all, a coordinate structure headed by a comma may be introduced by a grammatical rule which requires it to be surrounded by commas. This is the case with the subordinate clause in Figure 7.10, dominated by the c-structure node `CPbare[sub]`. The two commas around it, both dominated by `CP[sub]`, are co-heads with the comma which marks coordination. In this case selecting the final comma as the true head is not appropriate; rather, the penultimate comma plays this role. Fortunately, such cases are relatively rare and the following rule of thumb, replacing the above preliminary version, seems to work in most relevant cases in $\text{UD}^{\text{PL}}_{\text{LFG}}$:[4]

- otherwise select the final comma as the true head in case there are at most two commas in the co-head set, or the penultimate comma in case there are more than two commas in the co-head set.

The second complication concerns subordinate clauses. In the LFG structure bank, some complementisers introducing such clauses are heads, projecting a separate feature structure, and others are co-heads of the main verbs within the subordinate clause, marking their presence via an additional attribute. This difference corresponds to whether the complementiser introduces a semantic relation, e.g., the complementiser ŻEBY 'in order to' may express causality, or whether it is idiosyncratically selected by the higher verb. Some complementisers, including ŻEBY, may have either function, as illustrated by the following ambiguous sentence (slightly simplified with respect to the real treebank sentence):

(7.5) Piszę, żeby uratować ludzkość.
write.1SG COMP save.INF humanity.ACC
'I write in order to save humanity.'
'I am writing that humanity should be saved.'

In both cases the c-structure is the same, cf. Figure 7.12. But the two f-structures differ: that in Figure 7.13, corresponding to the semantic use of ŻEBY, where it expresses causality, has one more level of feature structure than that in Figure 7.14, featuring the idiosyncratic use of ŻEBY (and the grammatical function of the subordinate clause differs). In the latter case, since the asemantic complementiser projects into the same feature structure as the verbal head of the subordinate clause (i.e., they are co-heads), the verb will be selected as the true head, according to the procedure described above. However, we do not yet have a clause for the former case, where the semantic complementiser is the true head, so there should be an additional rule, positioned before the final clause (the one concerned mostly with punctuation), saying that when a complementiser is one of the co-heads (other co-heads being at this stage only punctuation marks), it is the true head.

Unfortunately, things are even more complicated by the fact that such an asemantic complementiser may introduce a conjunction-less coordinate clause, as in the above example (7.4). So, going back to Figure 7.10, there are four elements in the co-head set corresponding to the subordinate coordinate structure: the two surrounding commas, the internal comma acting as the conjunction, and the asemantic complementiser ŻE. But, this time, it is not the complementiser

[4]There are some cases of asyndetic coordination with multiple conjuncts, where this rule wrongly selects the penultimate comma as the conjunction, but this is rectified in further steps of conversion, where the first conjunct – rather than the conjunction – becomes the head of the coordinate structure.

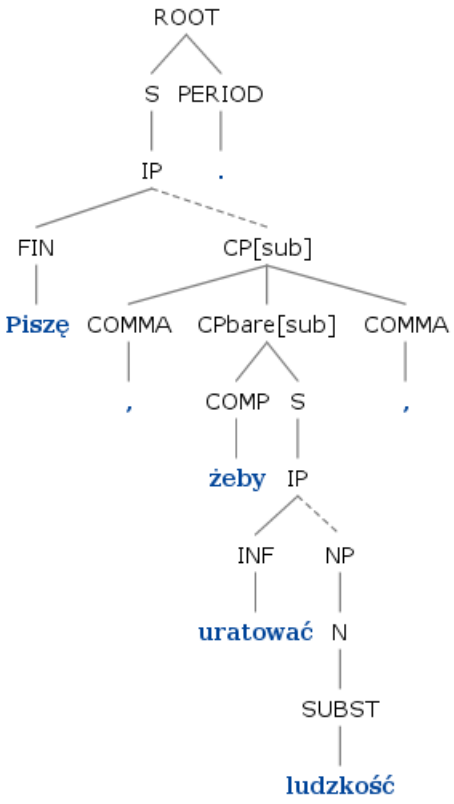


Figure 7.12: C-structure of (7.5)

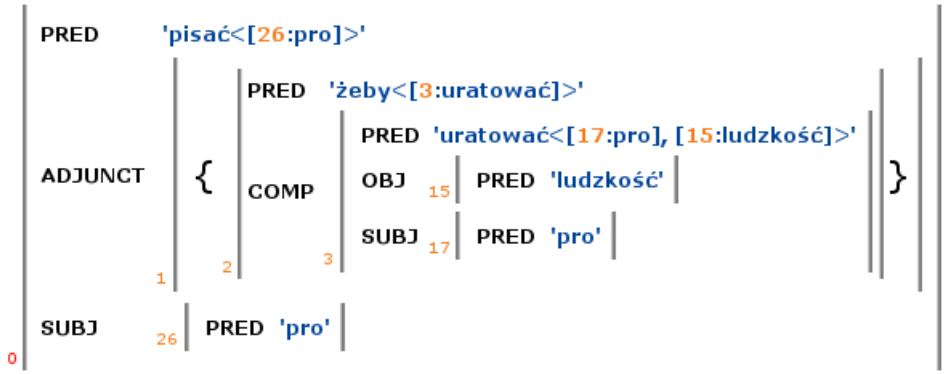


Figure 7.13: Schematic f-structure of (7.5) with the semantic complementiser ŻEBY

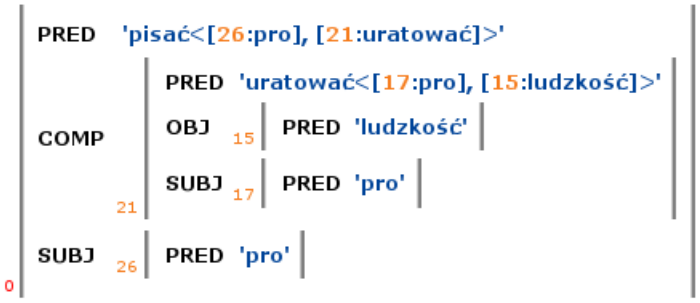


Figure 7.14: Schematic f-structure of (7.5) with the asemantic complementiser ŻEBY

that is the true head of this structure – the comma expressing conjunction still plays this role. Hence, whenever a co-head set contains only punctuation marks and a complementiser, two situations must be distinguished: either it is a semantic complementiser, in which case it is the true head, or it is an asemantic complementiser, in which case it is ignored and one of the commas is selected as the true head, on the assumption that it heads a coordinate structure.

To summarise, the following ordered rules are responsible for selecting the true head from a set of co-heads:

- if there is a verbal token among the co-heads, select it as the true head; a verbal token is defined here as having one of the following preterminals: `FIN`, `PRAET`, `INF`, `IMPS`, `IMPT`, `PRED`, `WINIEN`, `BEDZIE`, `PCON`, `PANT`;
- otherwise, if there is a nominal or adjectival token, it is the true head; this concerns tokens with the following preterminals: `SUBST`, `DEPR`, `GER`, `PPRON12`, `PPRON3`, `SIEBIE`, `NUM`, `ADJ`, `PACT`, `PPAS`;
- otherwise, if there is an explicit conjunction (preterminal `CONJ`), it is the head;
- otherwise, if there is a complementiser (preterminal `COMP`) of the semantic kind (this is determined on the basis of the corresponding f-structure), select it as the true head;
- otherwise select the final comma as the true head in case there are at most two commas in the co-head set, or the penultimate comma in case there are more than two commas in the co-head set.

7.1.2 Dependencies between true heads

The backbone of the initial dependency representation consists of true heads connected with dependency relations read directly off the f-structure. For example, in the case of the sentence (7.2), repeated below for convenience, whose f-structure is given in Figure 7.7 on page 117, there are four grammatical functions specified in this f-structure: OBL-STR, COMP and – twice – SUBJ.

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

The values of the two SUBJ attributes are phonetically empty pronouns, which do not correspond to any tokens in the sentence, so they will not surface in the initial dependency representation.[5] The attribute OBL-STR relates the whole f-structure to the substructure (with index 6) corresponding to *słowo* 'word'. There are three tokens corresponding to the whole f-structure: the two punctuation marks at the extremes of the sentence and the verb *daję* 'give'. Since it is the verb that is the true head among these three co-heads, the OBL-STR dependency relates this verb and the noun *słowo*. Similarly, among the six tokens corresponding to the value of COMP (the two commas, *że*, *się*, *nie* and *gniewam*), the procedure outlined in the previous subsection selects the verb *gniewam* 'be angry' as the true head, so the COMP dependency will connect

[5] While UD v.2 allows for empty tokens, at the moment this possibility is constrained to elided predicates, and not applicable to *pro*-dropped dependents.

the matrix verb *daję* ‘give’ and this embedded verb *gniewam* ‘be angry’. Hence, the backbone dependency structure for (7.2) is that shown in Figure 7.15.

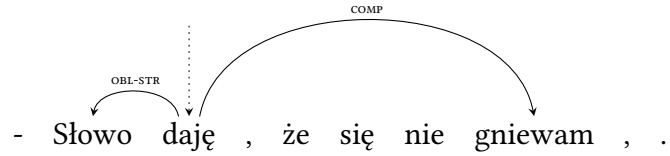


Figure 7.15: Initial dependency representation of (7.2) – the backbone

Similarly, in the case of (7.3), also repeated below, the dependencies which more or less directly correspond to grammatical functions in the LFG representation are shown in Figure 7.16. One novelty here concerns coordinate structures and consists in the translation of set membership in the f-structure representation of coordination into the CONJ dependency.

(7.3) Uderzał rękami w głowę, drapał twarz.
hit.3SG.M hands.INS in head.ACC scratched.3SG.M face.ACC
‘He pounded his head with his fists, scratched his face.’

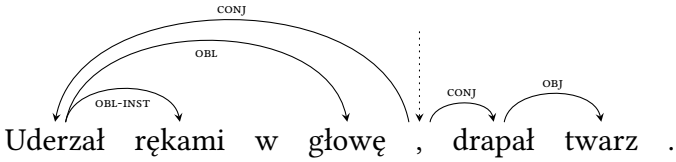


Figure 7.16: Initial dependency representation of (7.3) – the backbone

Note that this initial backbone does not have to be a tree. It is not in the case of example (7.4), where the two coordinated verb forms, *przerosła* ‘overwhelmed’ and *przeraziła* ‘scared’, share a number of dependents: the SUBJect *wojna* ‘war’, the OBJect *go* ‘him’, and the ADJUNCT *jednak* ‘after all’. As this dependent-sharing is explicitly represented in the f-structure (see Figure 7.11 on page 120), all these dependencies will be reflected in the backbone of the initial dependency representation, as illustrated in Figure 7.17.

(7.4) Wydawało się, że wojna jednak go przerosła,
seemed.3SG.N RM COMP war.NOM.SG.F after_all him.ACC overwhelmed.3SG.F
przeraziła.
scared.3SG.F
‘It seemed that after all the war overwhelmed and scared him.’

Similarly, multiple incoming dependencies will also occur in cases of control and raising.

For completeness, Figures 7.18–7.19 present the initial backbones of the two meanings of (7.5).

(7.5) Piszę, żeby uratować ludzkość.
write.1SG COMP save.INF humanity.ACC
‘I write in order to save humanity.’
‘I am writing that humanity should be saved.’

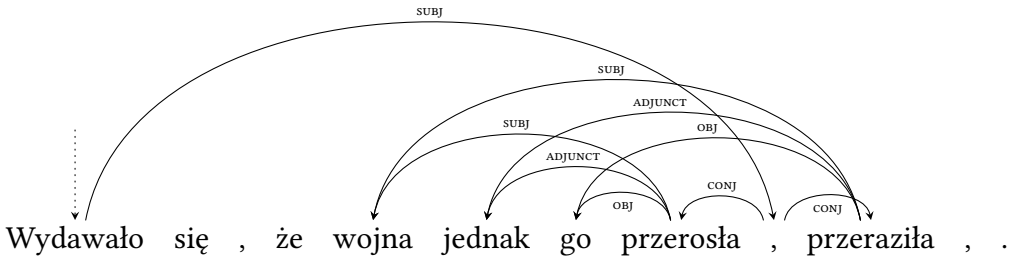


Figure 7.17: Initial dependency representation of (7.4) – the backbone

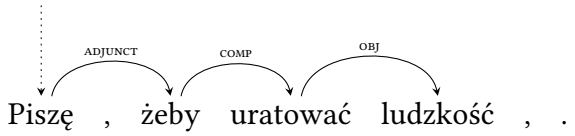


Figure 7.18: Initial dependency representation of (7.5) with the semantic complementiser ŻEBY – the backbone

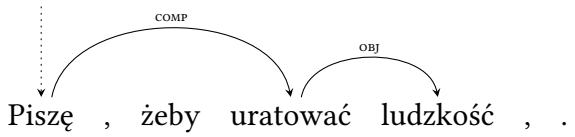


Figure 7.19: Initial dependency representation of (7.5) with the asemantic complementiser ŻEBY – the backbone

7.1.3 Adding dependencies to other co-heads

In order to complete initial dependency representations, co-heads which were not selected as true heads must be made dependents of the respective true heads. What should be the labels of dependencies added this way? As new labels will have to be assigned in the final UD representations anyway, at this stage the basic algorithm is very simple: take as the dependency label the preterminal name of the dependent. Thus, in the case of the first example of this section, (7.2), whose constituency tree is given in Figure 7.6 on page 117, and the backbone dependencies – in Figure 7.15 on page 124, the complete initial dependency representation will have the form shown in Figure 7.20. There, the dependency from *gniewam* 'be angry' to its co-head *się* (the inherent reflexive marker) is called RM simply because the preterminal symbol of *się* is `RM`, etc. The only minor exception to this rule concerns complementisers, whose preterminal is `COMP`: as there is a grammatical function of the same name in LFG f-structures (for subordinate clauses which are arguments but not subjects or objects), the dependency from the true head of the subordinate clause to the asemantic complementiser is renamed to COMP-FORM, as also illustrated in Figure 7.20.

For completeness, Figures 7.21–7.24 present complete initial dependency structures for all the other examples used in this section. Such initial dependency structures are subsequently transformed into final enhanced UD representations.

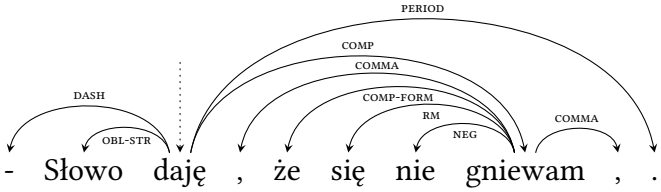


Figure 7.20: Initial dependency representation of (7.2)

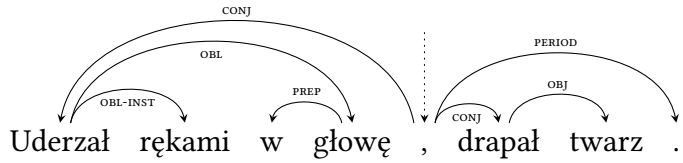


Figure 7.21: Initial dependency representation of (7.3)

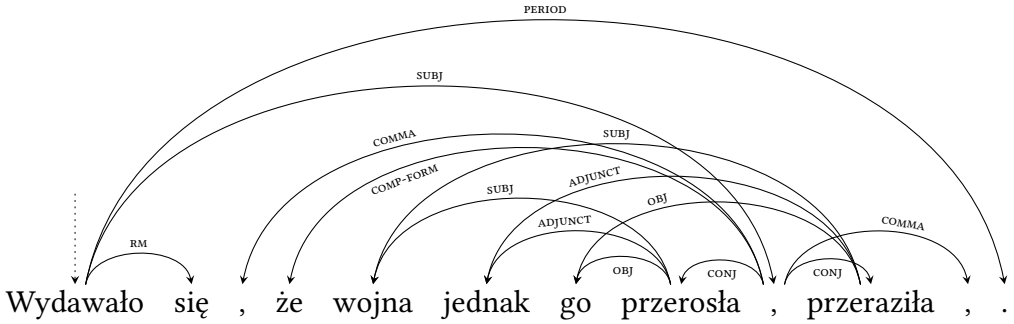


Figure 7.22: Initial dependency representation of (7.4)

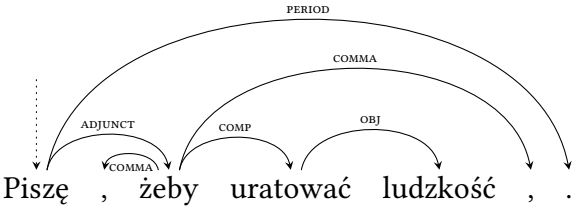


Figure 7.23: Initial dependency representation of (7.5) with the semantic complementiser ŻEBY

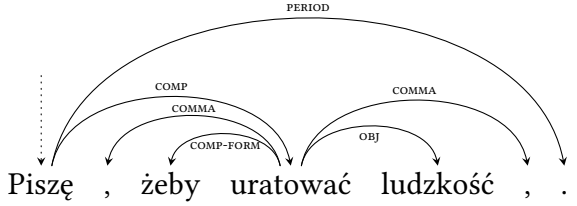


Figure 7.24: Initial dependency representation of (7.5) with the asemantic complementiser ŻEBY

7.1.4 Converting to initial basic dependency tree

As can be seen in Figure 7.22, the initial dependency structure does not have to be a tree. One reason, illustrated in this figure, is the possibility of dependent-sharing in coordinate structures. Here, the two asyndetically coordinated verbs, *przerosła* 'overwhelmed' and *przeraziła* 'scared', share the subject *wojna* 'war', the object *go* 'him' and an adjunct *jednak* 'after all' – each of these three dependents has two incoming edges in this graph. Such dependency graphs are further converted into enhanced UD structures (see the next section), which also do not have to be trees, but a simplified initial dependency tree is also created at this stage, which is subsequently converted into the basic UD tree.

In order to derive the initial dependency tree from the complete initial dependency structure, the edges from all conjuncts to a given shared dependent are merged into one edge from the conjunction to this dependent. The effect of applying this strategy to the structure in Figure 7.22 is shown in Figure 7.25.

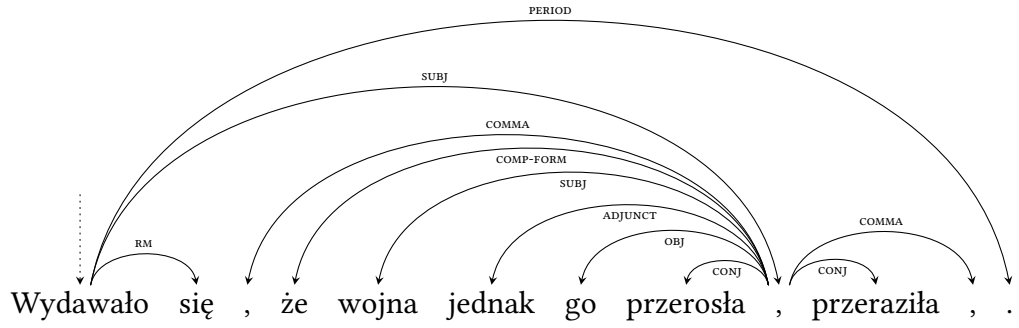


Figure 7.25: Initial dependency representation of (7.4) – basic tree

One complication is that a dependent may be shared across more than one level of coordination. Consider the following sentence:

(7.6) Dyrektor zapoznał Grodzickiego z katechetą, potem pożegnał się i
director.NOM introduced Grodzicki.ACC to catechist then said_goodbye RM and
wyszedł.
left
'The director introduced Grodzicki to the catechist, and then said goodbye and left.'

There are two coordinations in this sentence: asyndetic at the main level, with the two conjuncts: *Dyrektor... z katechetą* 'the director introduced Grodzicki to the catechist' and *potem... wyszedł* 'then (he) said goodbye and left', and syndetic within the second conjunct, with the adjunct *potem* 'then' shared between the two embedded conjuncts: *pożegnał się* 'said goodbye' and *wyszedł* 'left'. The initial dependency representation of this sentence is shown in Figure 7.26. The technical difficulty is that the subject, *dyrektor*, is a shared dependent of conjuncts from two different levels of coordination: it is the subject of *zapoznał* 'introduced' at the top level, but also of the two verbs coordinated *within* the second top-level conjunct: *pożegnał się* 'said goodbye' and *wyszedł* 'left'. Hence, the appropriate dependency tree should be as

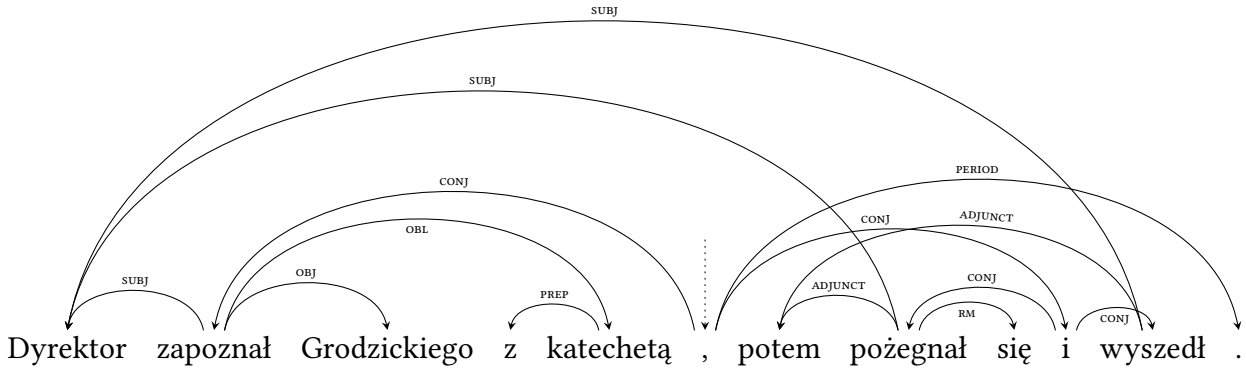


Figure 7.26: Initial dependency representation of (7.6)

in Figure 7.27, on the understanding that the sharing of a dependent marked by an edge from the conjunction (here, comma) to this shared dependent 'percolates down' in case some of the conjuncts are coordinate structures themselves.

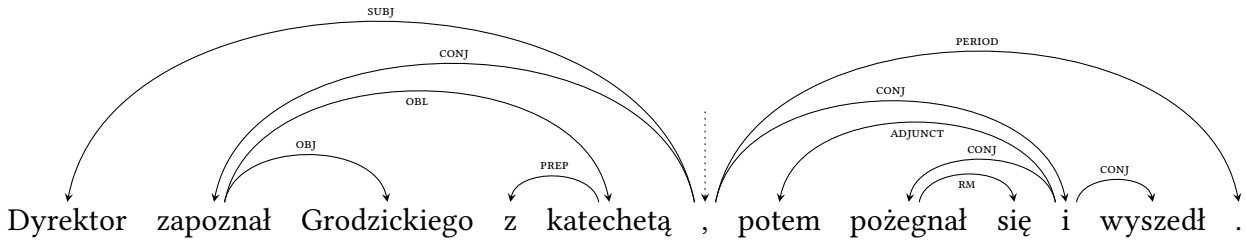


Figure 7.27: Initial dependency representation of (7.6) – basic tree

Apart from coordination, another obvious case of multiple edges to the same token is control, understood here widely as also including raising and predicative constructions. Both are present in the following sentence, whose initial complete and basic dependency structures are presented in Figures 7.28–7.29.

(7.7) Blondyn zaczął być zły.
blond.NOM.SG.M began.3SG.M be.INF angry.NOM.SG.M
'The blond guy started to be angry.'

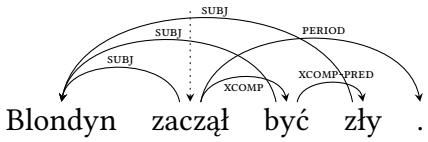


Figure 7.28: Initial dependency representation of (7.7)

In the input LFG representations, infinitival complements in control and raising constructions bear the XCOMP relation to the head, and predicative complements are marked as XCOMP-PRED;

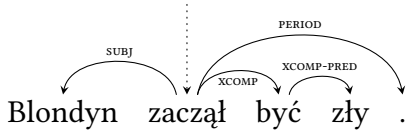


Figure 7.29: Initial dependency representation of (7.7) – basic tree

hence the labels in the two figures. In this case the operation leading to the basic tree is simple: remove SUBJ edges coming from tokens which have one of the following incoming relations: XCOMP, XCOMP-PRED (both illustrated here) or XADJUNCT (controlled adjuncts, often adverbial participles). This rule is extended to adjectival participles, as in (7.8), where the participle *zbierającej* 'collecting' is a modifier of the head noun, *osoby* 'person', but also has this head noun as its subject, as shown in Figure 7.30. This last dependency is removed in the initial basic tree; cf. Figure 7.31.

(7.8) Sprawdź dokumenty osoby zbierającej datki.
check.IMP.2SG documents.ACC person.GEN collecting.GEN contributions.ACC
'Verify the documents of the person collecting contributions.'

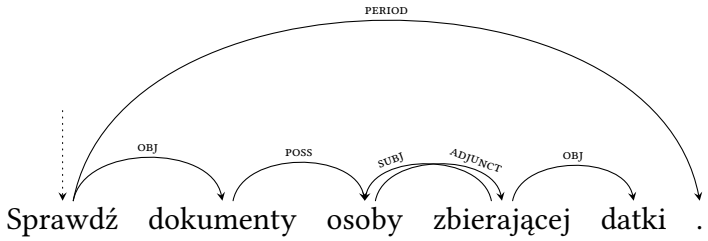


Figure 7.30: Initial dependency representation of (7.8)

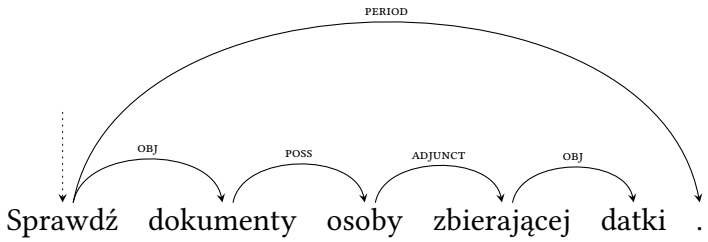


Figure 7.31: Initial dependency representation of (7.8) – basic tree

Again, coordination complicates the above procedure a little, as the target of the XCOMP, XCOMP-PRED or XADJUNCT dependency may be the conjunction in a coordination of controlled dependents, rather than a controlled dependent itself. This is illustrated in Figure 7.32, involving the coordination of two predicative adjectives: *szary* 'grey' and *niemrawy* 'sluggish'. Both adjectives have *dzień* 'day' as its subject, but – in order to simplify this graph to a tree – both SUBJect dependencies need to be removed since *dzień* is also a dependent of the finite verb *wstał* 'arose'. The complication is that the XADJUNCT dependency licensing this removal

targets the conjunction *i* ‘and’ rather than the adjectives directly. The pruning procedure recognises such situations and recursively descends into coordination, resulting here in the basic tree in Figure 7.33.

(7.9) Dzień wstał szary i niemrawy.
day.NOM.SG.M arose.3SG.M grey.NOM.SG.M and sluggish.NOM.SG.M
‘The day started grey and dim.’

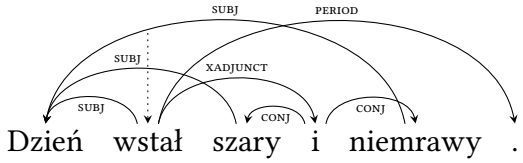


Figure 7.32: Initial dependency representation of (7.9)

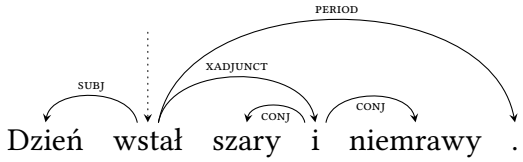


Figure 7.33: Initial dependency representation of (7.9) – basic tree

The third, after coordination and broadly understood control, situation that gives rise to multiple incoming dependencies, concerns free relatives, as in the following example:

(7.10) Ktokolwiek zostawił plecak, nie zamieszkiwał tutaj.
whoever.NOM.SG.M left.3SG.M rucksack.ACC NEG lived.3SG.M here
‘Whoever left the rucksack didn’t live here.’

According to the input LFG structures, *ktokolwiek* ‘whoever’ is the subject of the main verb, *zamieszkiwał* ‘lived’. The representation of the relative clause *ktokolwiek zostawił plecak* ‘whoever left the rucksack’ is an adjunct of *ktokolwiek*. Moreover, *ktokolwiek* is the subject of *zostawił* ‘left’ within this relative clause. Hence, there are two SUBJect dependencies targeting the pronoun, as shown in Figure 7.34. In order to turn this dependency graph into a tree, the SUBJect dependency from *zostawił* ‘left’ to *ktokolwiek* ‘whoever’ must be removed, as shown in Figure 7.35.

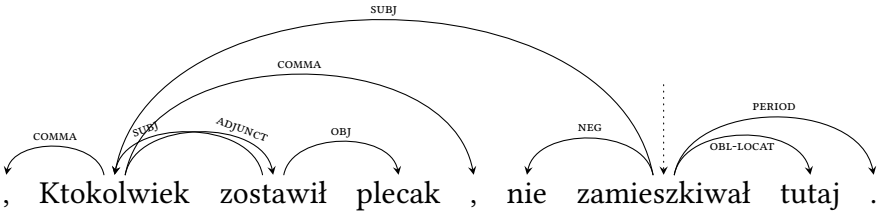


Figure 7.34: Initial dependency representation of (7.10)

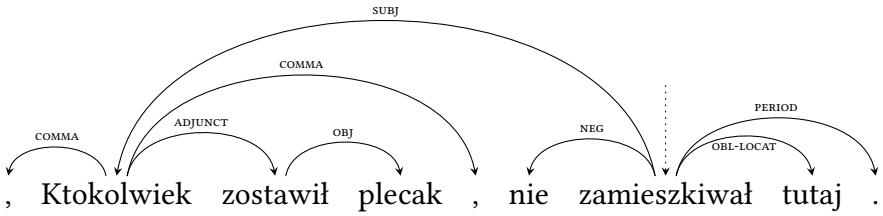


Figure 7.35: Initial dependency representation of (7.10) – basic tree

7.2 From initial dependencies to UD v.2

7.2.1 Tokenisation

Initial dependency representations described in the previous section are very close to LFG representations: dependency relations are based directly on functional structures and – in case of dependencies between co-heads – on constituent structures, tokenisation follows that assumed in constituent trees. Hence, the very first step in converting initial dependency structures into final UD structures consists in converting tokenisation, as described in Chapter 5: mobile inflections are stripped off of the preceding '+', multi-token words are split into separate tokens, spurious commas absent in the original input sentence are removed.

In the case of the above sentence (7.10) involving a free relative, the effect of this step is shown in Figure 7.36. Note that, as above, the tree above the sentence is the – very partial, so far – result of converting the initial basic tree in Figure 7.35 to the basic UD tree, and the tree below the text is the – again, very partial at this point – result of converting the initial complete dependency structure in Figure 7.34 into enhanced UD.[6] In this case, the only effect of this first step of the second conversion stage is the removal of the initial comma.

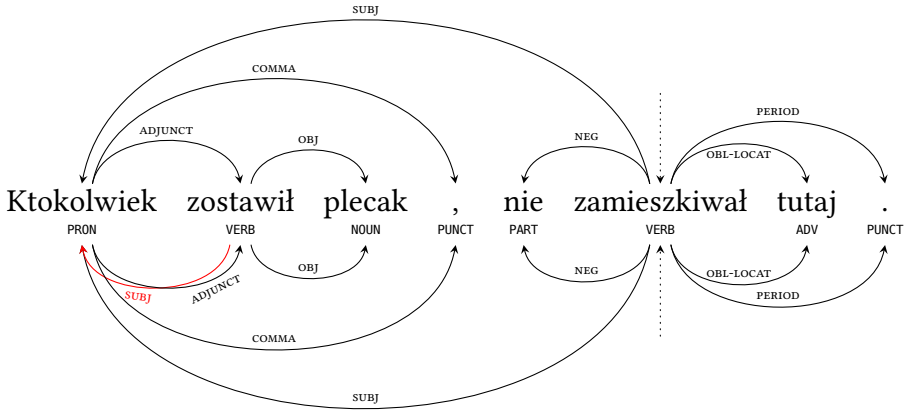


Figure 7.36: Towards UD representation of (7.10) – after tokenisation

[6] As above, whenever the two structures are identical, only one is displayed (above the tokens), and when they are different, the differences are shown in red. Moreover, all representations illustrating the second conversion stage include coarse parts of speech of all tokens.

A slightly more interesting example is (7.11), involving a multi-token word *na pewno* ‘for sure, certainly’. Its initial dependency representations are given in Figures 7.37–7.38, and the result of the first step of conversion into UD – in Figure 7.39. Note that here an actual UD dependency relation was introduced, namely, `fixed`.

(7.11) - Reforma na pewno nie zostanie zaniechana.
reform.NOM.SG.F for sure NEG become.FUT.3SG abandoned.NOM.SG.F
‘The reform will certainly not be abandoned.’

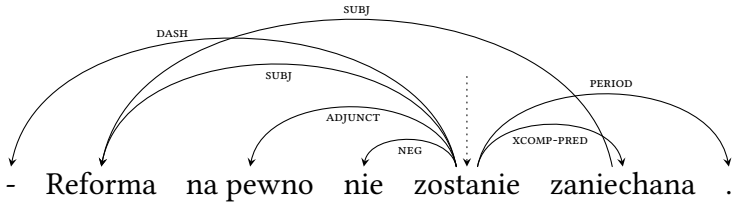


Figure 7.37: Initial dependency representation of (7.11)

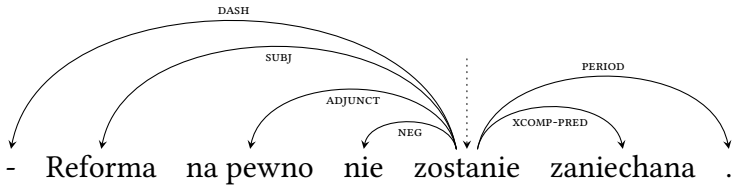


Figure 7.38: Initial dependency representation of (7.11) – basic tree

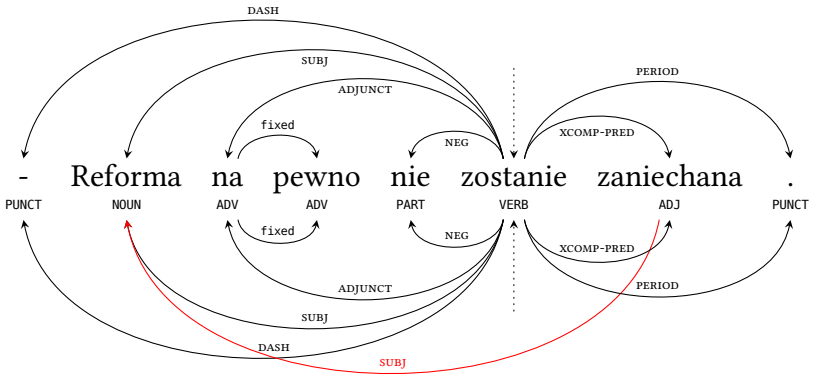


Figure 7.39: Towards UD representation of (7.11) – after tokenisation

7.2.2 Initial conversion of coordination

Recall – from the introduction to this chapter – sentence (7.1), repeated below.

(7.1) Pracodawca musi też płacić wszelkie podatki i ubezpieczenia.
employer.NOM must also pay.INF all.ACC taxes.ACC and insurance.ACC
'The employer must also pay any taxes and insurance.'

Its complete initial dependency representation is given in Figure 7.3 on page 115 (and the final UD representation – in Figure 7.4). After the re-tokenisation step, which in this case does not change the tokenisation at all, the two dependency structures look as in Figure 7.40 (with the differences between them, again, shown in red). Note that neither of the two representations

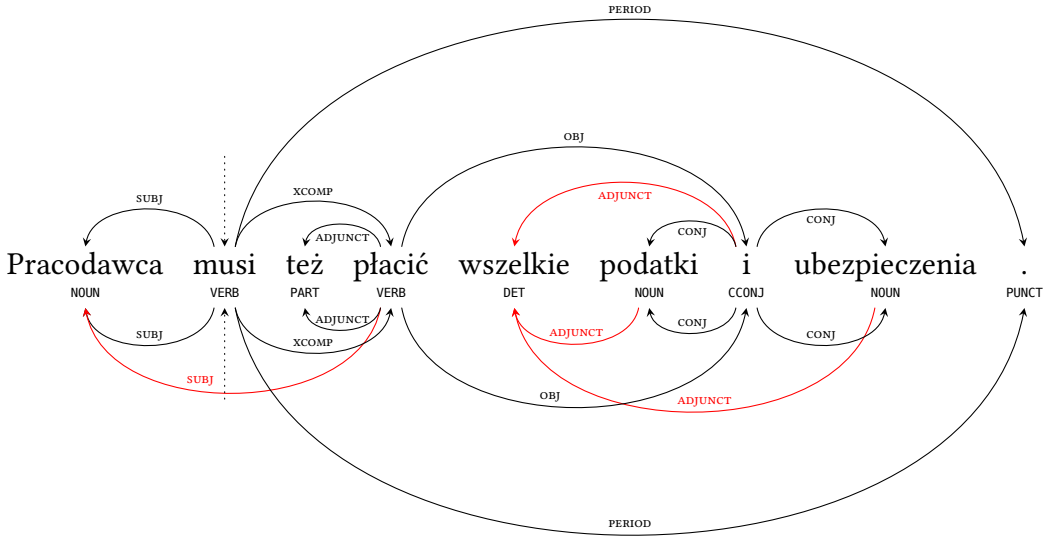


Figure 7.40: Towards UD representation of (7.1) – after tokenisation

of coordination satisfy UD guidelines. In both structures it is the conjunction, rather than the first conjunct, that is the head of the coordinate structure. In the basic tree representation at the top, both conjuncts are dependents of the conjunction, the OBJ dependency aimed at the whole coordinate structure targets the conjunction, and similarly the shared dependent of the two conjuncts is represented as a dependent of the conjunction. Matters are only slightly better in the fuller dependency representation in the lower part: the shared dependent does depend on both conjuncts rather than on the conjunction, but it is the conjunction that is the head of the coordinate structure in all other respects. The first non-trivial conversion step from initial dependency structures to UD reorganises dependencies to, from and within coordinate structures in a way compliant with UD guidelines. The effect of this step in case at hand is presented in Figure 7.41. Note that after this step it is the first conjunct that is the head of the coordinate structure. In the basic tree, it receives incoming dependencies from outside (here, OBJ), it is the source of outgoing dependencies to the outside (here, ADJUNCT), and it is the head of all other conjuncts (here, just one), with the final conjunct heading the conjunction.

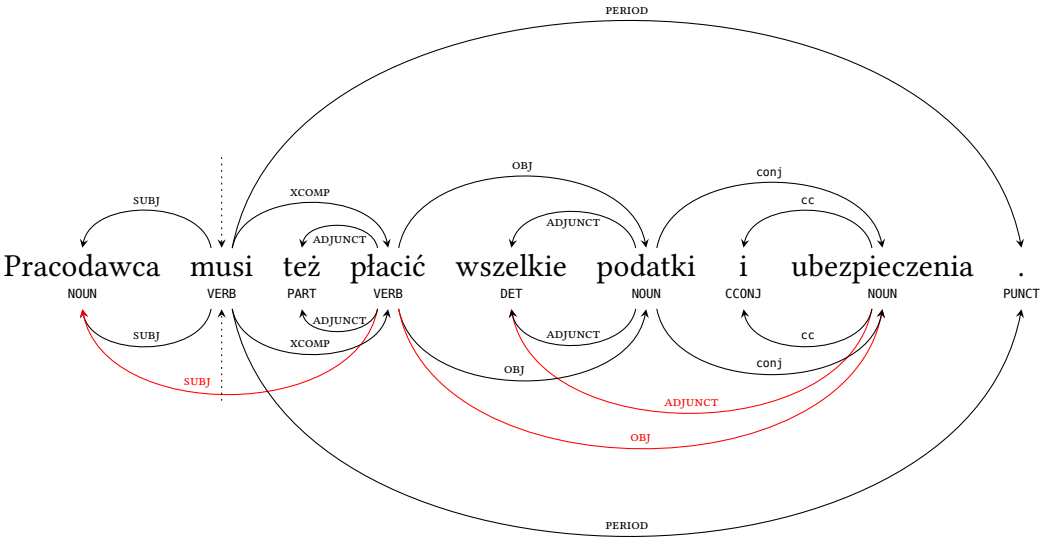


Figure 7.41: Towards UD representation of (7.1) – after initial conversion of coordination

Additionally, in enhanced dependencies, there are additional edges from outside to the non-initial conjunct, and from the non-initial conjunct to the shared dependent.

One complication in this procedure concerns multi-token conjunctions, as in the following example:

(7.12) Sprawca ten okazał się nie tylko złodziejem, ale
perpetrator.NOM.SG.M this.NOM.SG.M turned_out.3SG.M RM NEG only thief.INS but
i sadystą.
and sadist.INS
'This perpetrator turned out to be not only a thief, but also a sadist.'

In this example, the conjunction consists of two parts, each being a two-word token: *nie tylko* 'not only' and *ale i* 'but also'. Hence, the representation of this sentence after the tokenisation step is as in Figure 7.42. The complication is that while normally all dependents of the conjunction must be either carried over to the first conjunct (in the case of the basic tree) or distributed over all conjuncts (in the case of enhanced dependencies), `fixed` dependents are exempt from this rule; otherwise, the `fixed` dependent *i* of *ale* would become a `fixed` dependent of the two conjuncts. The correct representation of this sentence after the coordination step is given in Figure 7.43.

7.2.3 Punctuation

As can be seen in Figure 7.43, as a result of the previous step, commas within coordinate structure have the UD incoming dependency of type `punct`. The next – trivial – step is to convert all other punctuation dependencies, including COMMA, PERIOD, DASH, etc., into the UD dependency `punct`.[7] The representation of (7.12) after this step is given in Figure 7.44, and the representation of (7.11) – in Figure 7.45 (to be compared with Figure 7.39 on page 132).

[7] The actual algorithm converts the dependency relation into `punct` whenever the dependent is a punctuation mark, i.e., has the UPOS value `PUNCT`.

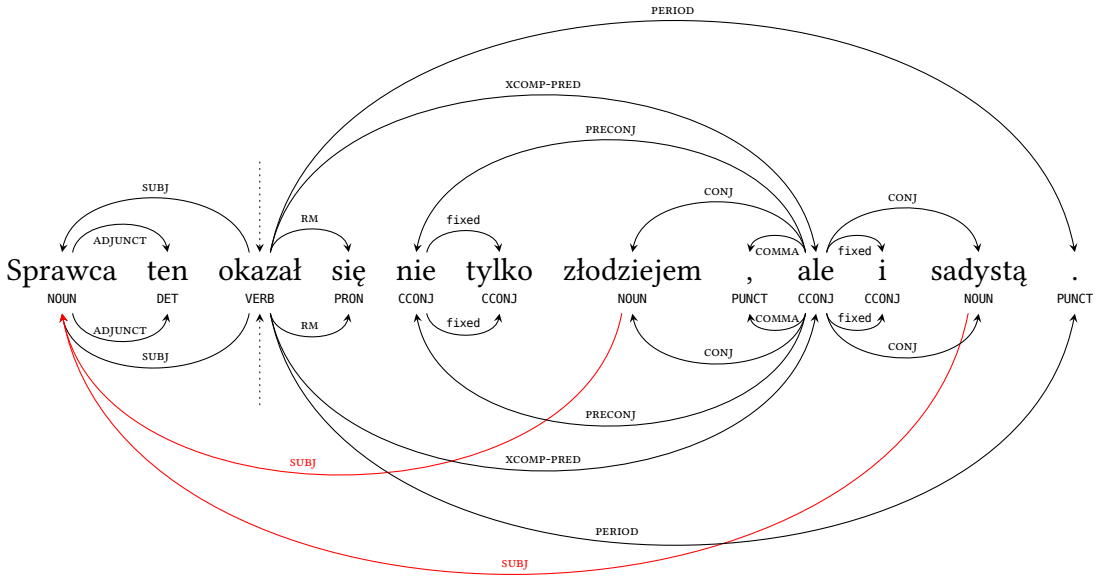


Figure 7.42: Towards UD representation of (7.12) – after tokenisation

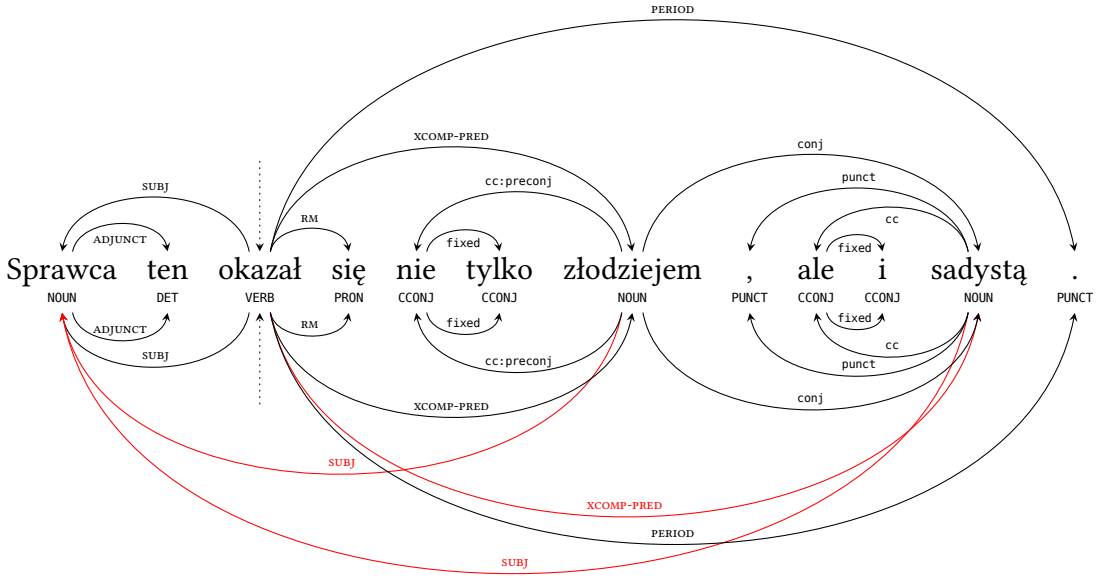


Figure 7.43: Towards UD representation of (7.12) – after initial conversion of coordination

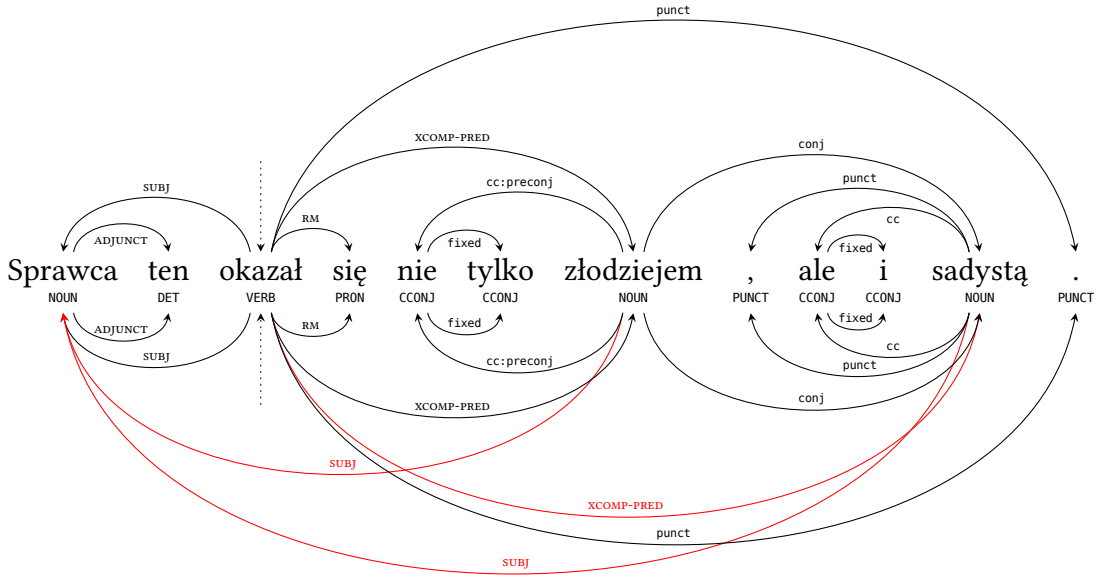


Figure 7.44: Towards UD representation of (7.12) – after conversion of punctuation

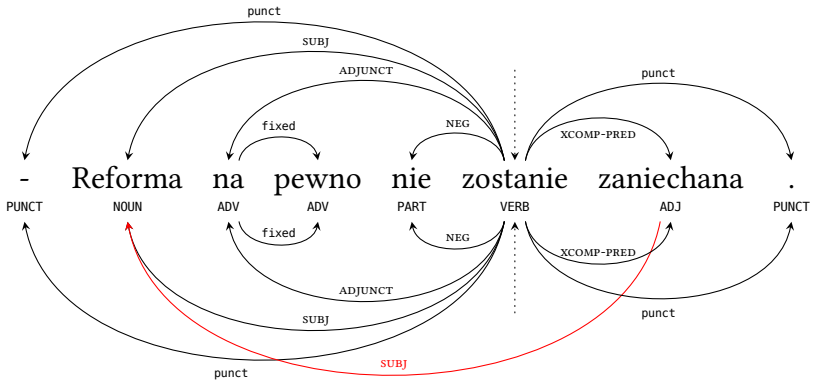


Figure 7.45: Towards UD representation of (7.11) – after conversion of punctuation

7.2.4 Reversing dependencies

The next step is much more important and it consists in reversing dependencies between functional and content words. In LFG, as in many contemporary linguistic theories, functional elements such as adpositions and complementisers are taken to be heads, and the content words in the phrases they combine with (usually, nouns and verbs, respectively) – are their dependents.[8] Similarly, numerals (hence, also adnumeral determiners) are taken to be true heads of numeral phrases. In UD, these dependencies need to be reversed, due to the principle of the primacy of content words.[9]

[8]In fact, in the LFG structure bank adpositions and complementisers are always heads, but in two different ways, depending on whether they introduce a semantic relation, as is for example the case with locative prepositions or those complementisers which introduce causal relations, or whether they are idiosyncratic markers of the nominal or clausal constituents they combine with. In the former – semantic — case, they are sole heads, so the dependency needs to be reversed. In the latter – asemantic – case, they are co-heads, together with the content words, and the content words are chosen as the true heads at an earlier stage (see Section 7.1.1), so nothing needs to be done here.

[9]`http://universaldependencies.org/u/overview/syntax.html`

Let us illustrate the effect of this step with the example (7.13):

(7.13) Odbywają się one w 100 fabrykach i PGR-ach.
happen.3PL RM they.NOM.PL in 100.LOC factories.LOC and PGRs.LOC
'They take place in 100 factories and PGRs (state-owned collective farms).'

Before this step, the basic tree – shown in the upper part of Figure 7.46 – makes it clear that the numeral *100* combines with the whole coordinate structure *fabrykach i PGR-ach* 'factories and PGRs' (otherwise the conj dependency to *PGR-ach* would originate in *100*, and not in *fabrykach*). After this step, this information is lost at the level of the basic tree. The upper part of Figure 7.47 is ambiguous between two syntactic structures: one where the numeral pertains to the whole coordination, and another where it combines with *fabrykach* 'factories' only; on this latter interpretation, there are 100 factories and an unspecified number of PGRs. However, the right interpretation may be read off the enhanced representation, where the nummod dependency to *100* originates not only from *fabrykach* 'factories', but also from *PGR-ach* 'PGRs' (and similarly for the case dependency to the preposition).

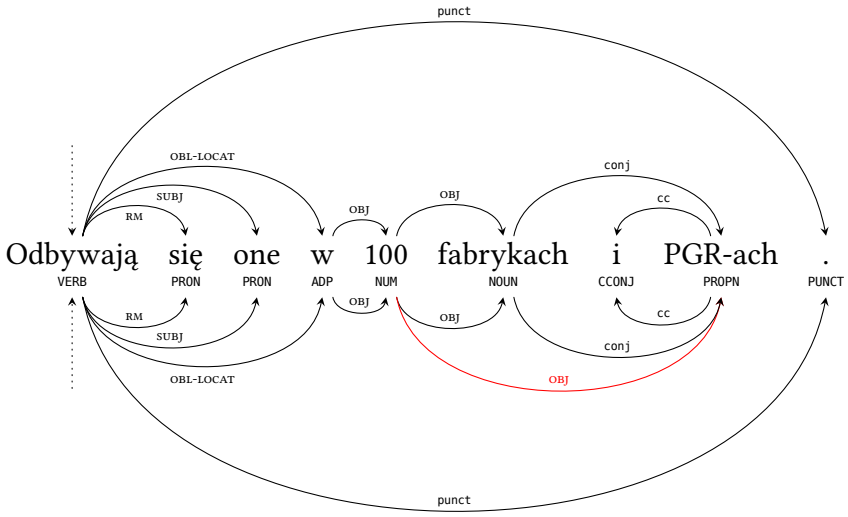


Figure 7.46: Towards UD representation of (7.13) – after conversion of punctuation

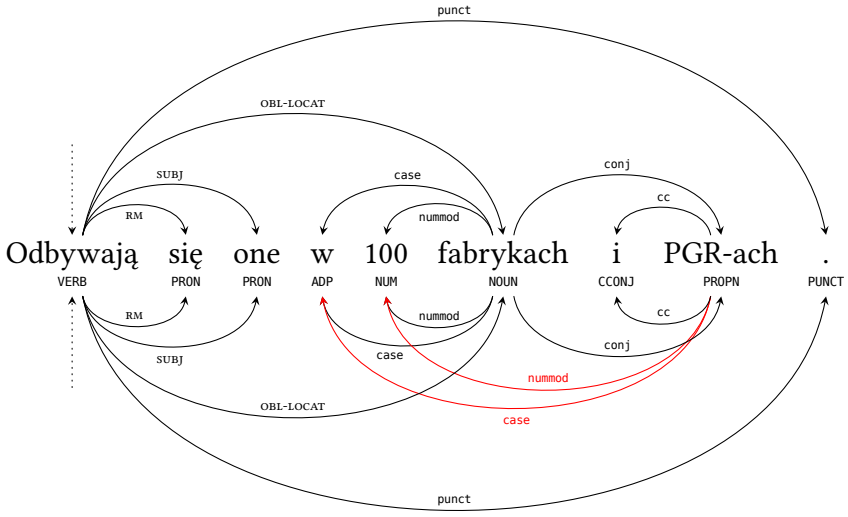


Figure 7.47: Towards UD representation of (7.13) – after reversing dependencies

This should be contrasted with example (7.14), where the numeral combines only with the first conjunct, as shown in Figure 7.48.

(7.14) Nad wszystkim czuwać będzie trzech lekarzy i personel pielęgniarski.
over everything watch.INF will three doctors and personnel nursing
'Three doctors and the nurses will watch over everything.'

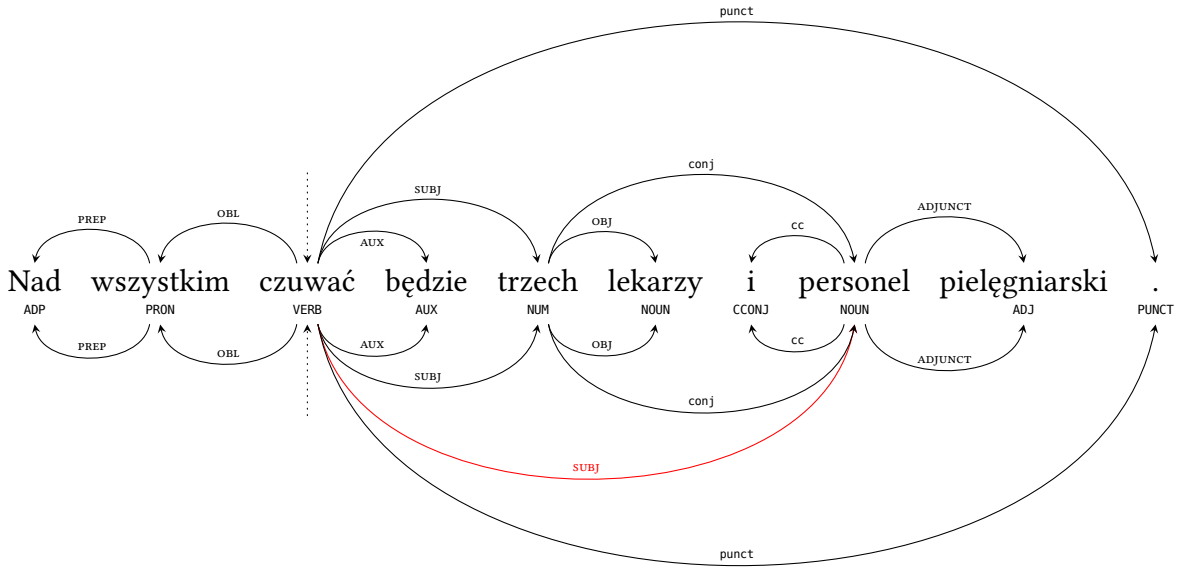


Figure 7.48: Towards UD representation of (7.14) – after conversion of punctuation

In this case, after reversing dependencies, the basic tree is also ambiguous between the two syntactic analyses, and again the right interpretation follows from enhanced dependencies: as shown in Figure 7.49, there is only one `nummod` dependency to the numeral, namely, from the first conjunct.

This step is relatively complex, as care must be taken to rearrange various incoming and outgoing dependencies. For example, in Figure 7.46, the OBL-LOCAT dependency from the main verb to the preposition *w* 'in' must be modified to target the new head of the prepositional phrase, i.e., the first conjunct, as shown in Figure 7.47. Also, in the enhanced representation another instance of the OBL-LOCAT dependency should be introduced, to the second conjunct, but this happens at a later stage of processing. Moreover, it would not be sufficient to reverse the dependency between the preposition and the numeral, as – according to UD guidelines – both should be dependents of the noun, and they should not be directly connected.

The above description concerns four kinds of dependencies:

- the COMP dependency originating from a complementiser, i.e., a token whose UPOS is `SCONJ`; in this case the reversed dependency has the UD label `mark`,
- the OBJ dependency originating from a numeral (UPOS `NUM`); the reversed dependency is `nummod`,

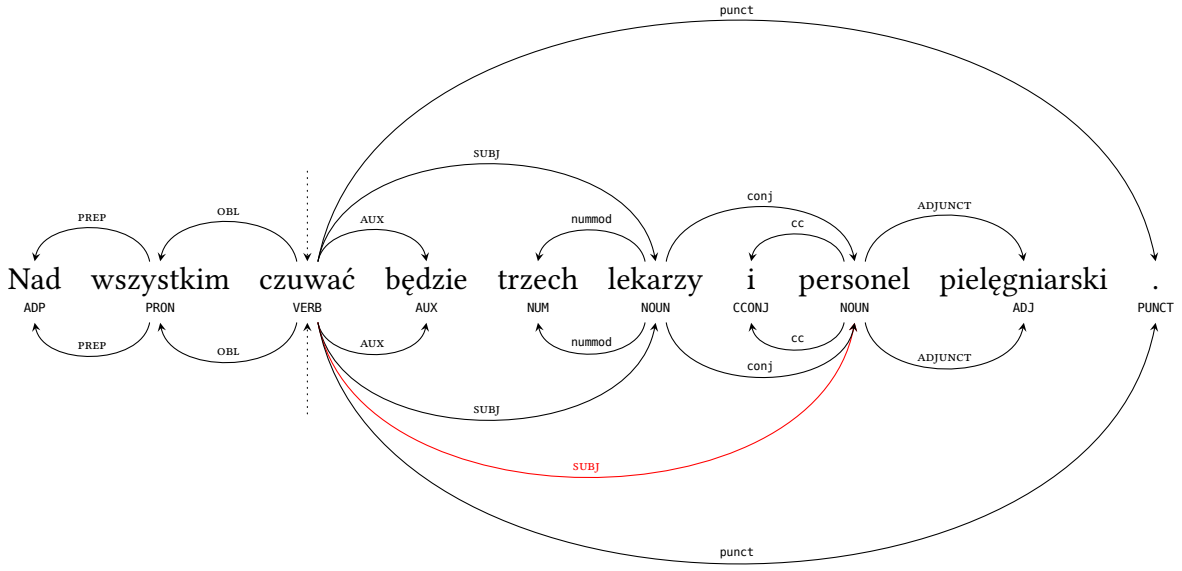


Figure 7.49: Towards UD representation of (7.14) – after reversing dependencies

- the OBJ dependency originating from a determiner (UPOS `DET`); the reversed dependency is `det`,
- the OBJ dependency originating from an adposition (UPOS `ADP`); the reversed dependency is `case`.

Apart from these, also the copula and some auxiliaries must be reanalysed from heads to dependents. Consider the following example and its dependency representation in Figure 7.50 (before the dependency reversal step):

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

After the initial conversion of coordination (see Section 7.2.2), the XCOMP-PRED dependency from *jest* 'is' to the coordinate structure is distributed to the two conjuncts: *wysoko zapięta pod szyję* 'buttoned up high to the neck' and *wysmukła jak kwiat* 'lean like a flower'. The first of these conjuncts is headed by an adjectival passive participle (*zapięta* 'buttoned up'), and the second – by an ordinary adjective (*wysmukła* 'lean'). Hence, according to the UD guidelines, *jest* 'is' acts as a passive auxiliary with respect to the first conjunct, and as a copula with respect to the second conjunct. This is shown in the enhanced dependency structure in the lower part of Figure 7.51; the tree in the upper part does not contain the information about this dual role of the function word *jest* 'is'.

Let us note in passing that not all XCOMP-PRED dependencies are reversed and translated to `aux:pass` or `cop`, but only those that indicate dependents of appropriate auxiliary or copular verbs (see below for a more precise description). In the LFG structure bank, the XCOMP-PRED grammatical function is also used to indicate predicative arguments in other constructions, in-

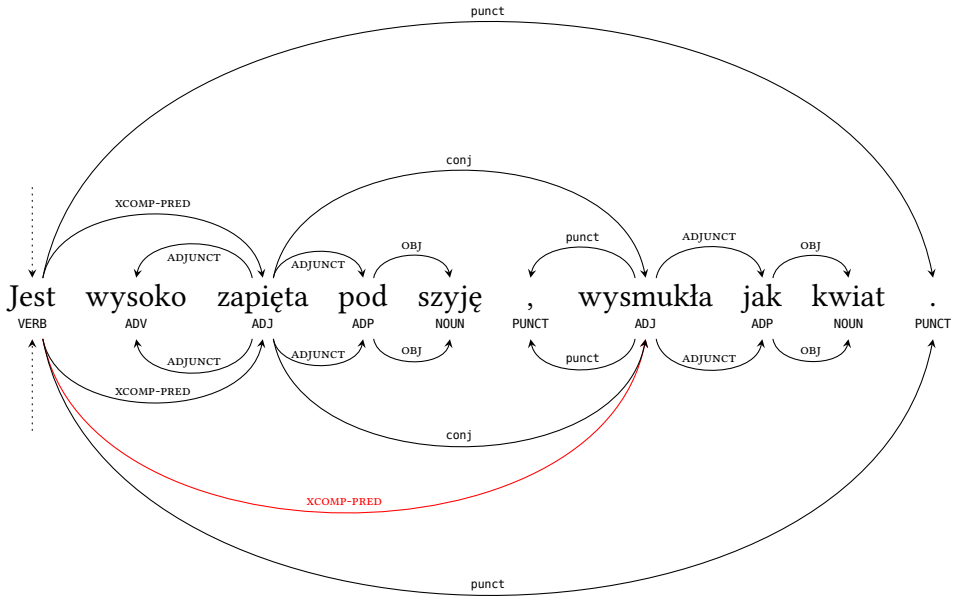


Figure 7.50: Towards UD representation of (7.15) – after conversion of punctuation

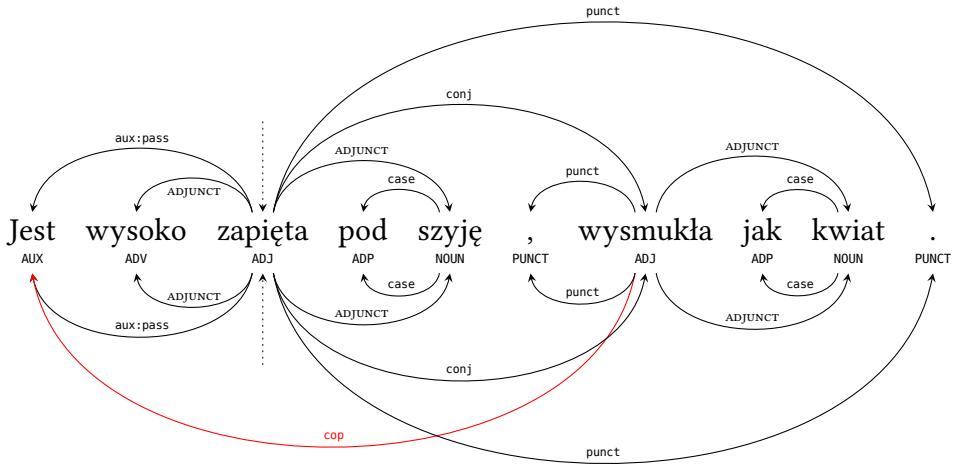


Figure 7.51: Towards UD representation of (7.15) – after reversing dependencies

volving content verbs such as OKAZAĆ SIĘ ‘turn out’, CZUĆ SIĘ ‘feel’ (as in ‘feel good’), STAĆ SIĘ ‘become’, ZOSTAĆ ‘remain’ (as in ‘remain alone’), CZYNIĆ ‘make’ (as in ‘make somebody popular’), etc. In such cases, the XCOMP-PRED relation is simply translated to `xcomp`, as illustrated in Figure 7.52, to be compared to Figure 7.44 on page 136.

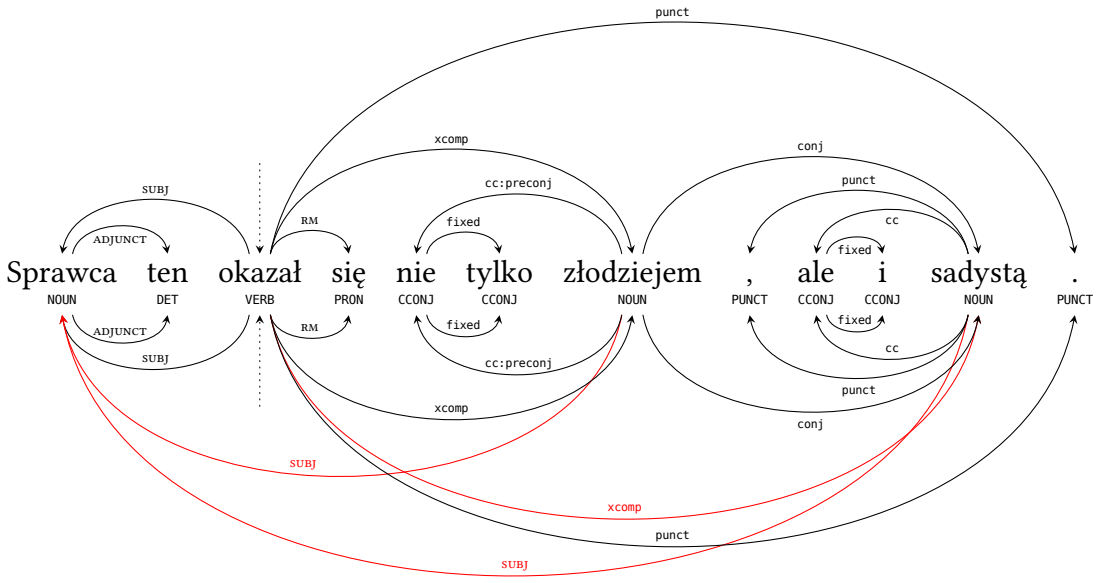


Figure 7.52: Towards UD representation of (7.12) – after conversion of XCOMP-PRED

Conversely, not only the XCOMP-PRED relation is reversed and translated as `cop`, but also some cases of the OBL-LOCAT relation, when it indicates dependents of copular verbs. In such cases the `cop` relation is subtyped with the `locat` qualifier, indicating locative copular constructions.[10] This is illustrated with example (7.16) whose initial dependency representation is given in Figure 7.53, and the representation after reversing dependencies – in Figure 7.54. While this is a simple 7-token sentence, its representations before the conversion of coordination and reversing dependencies and after these steps are dramatically different: even disregarding the labels, only one of seven dependencies survived these steps (the one from *słoikach* ‘jars’ to *tych* ‘these’).

(7.16) a co jest w tych słoikach?
and what.NOM is in these.LOC jars.LOC
‘And what’s in these jars?’

In (partial) summary, this part of the dependency reversing step concerns the following kinds of dependencies:

- the XCOMP-PRED dependency from a possible passive auxiliary (a form of BYĆ ‘be’, BYWAĆ ‘be (habitual)’, ZOSTAĆ ‘become’, ZOSTAWAĆ ‘become (habitual)’) to a passive participle (a token whose XPOS starts with `ppas`); in this case the reversed dependency has the UD

[10] Compare `http://universaldependencies.org/u/overview/simple-syntax.html#nonverbal-clauses`?.

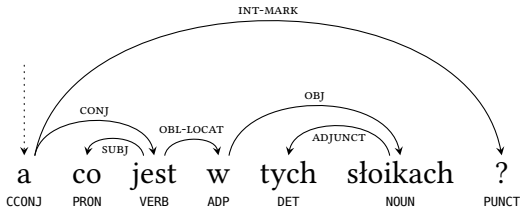


Figure 7.53: Towards UD representation of (7.16) – after tokenisation

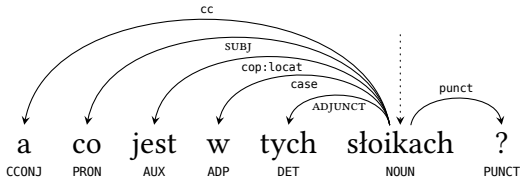


Figure 7.54: Towards UD representation of (7.16) – after reversing dependencies

label `aux:pass` (and the copula gets the UPOS `AUX`, even if it was assigned the tag `VERB` at an earlier stage; see `AUX` and `VERB` in Section 6.2),
- the XCOMP-PRED dependency originating from a possible copula (a form of BYĆ 'be', BYWAĆ 'be (habitual)' or TO – see, e.g., Bondaruk 2013 for arguments for the copula status of TO); the reversed dependency is `cop` (other occurrences of XCOMP-PRED are not reversed but they are translated into `xcomp`),
- the OBL-LOCAT dependency originating from a possible copula (as above); the reversed dependency is `cop:locat`.

While the two parts of the procedure of reversing dependencies are described in this subsection jointly, this second part, concerned with passive auxiliaries and copulas, interacts with the conversion of some of the grammatical functions, so it is actually performed at a slightly later stage, after the conversion of objects and adjuncts (but before subjects and obliques).

7.2.5 Converting grammatical functions

Subjects

The basic conversion of SUBJects into UD dependencies is relatively simple:

- if the target of the SUBJ relation is a token whose UPOS is `VERB`, then change the label to `csubj`,
- otherwise, i.e., if the target has a broadly nominal part of speech (`NOUN`, `PROPN`, `DET`, `NUM`, `ADJ`), change the label to `nsubj`.

The effect of this rule is illustrated with example (7.17), which involves a clausal subject at the matrix level and a nominal subject within the subordinate clause; see Figures 7.55 and 7.56 for representations before and after this step.

(7.17) Zdawało się, że dopiero teraz Maria Rosa ją zauważyła.
seemed.3SG.N RM COMP only now Maria.NOM Rosa.NOM her.ACC noticed
'It seemed that Maria Rosa has noticed her only now.'

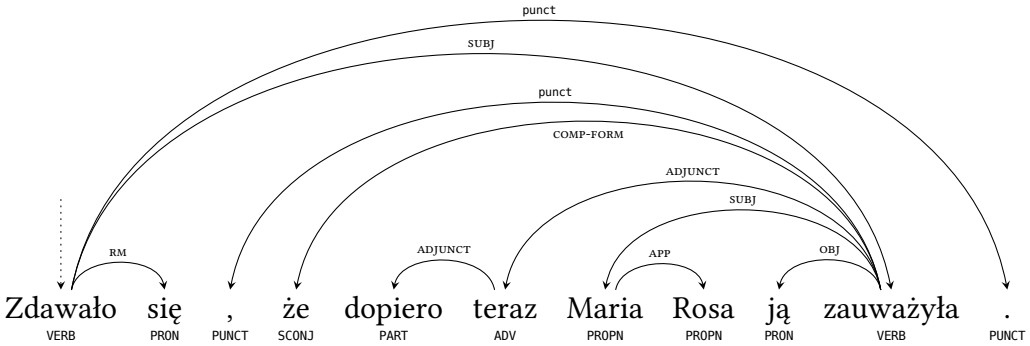


Figure 7.55: Towards UD representation of (7.17) – before converting subjects

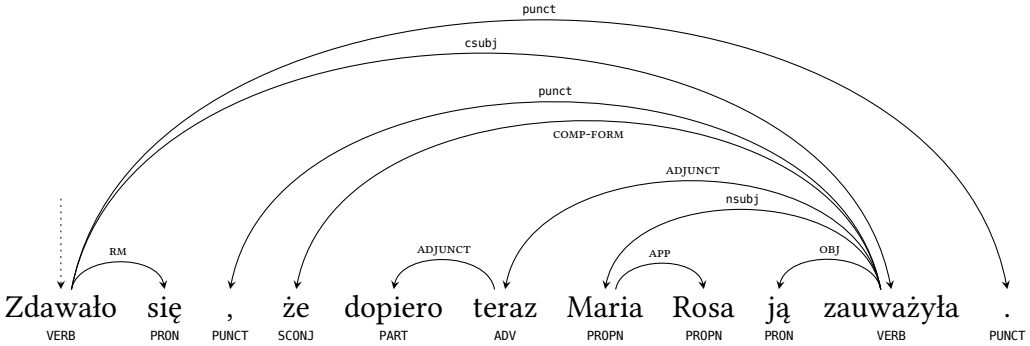


Figure 7.56: Towards UD representation of (7.17) – after converting subjects

There are three complications to this simple rule:

1. if the origin of the relation is a passive participle (its XPOS starts with `ppas`), then the subtype `pass` is added to the UD relation; in practice, there are `nsubj:pass` subjects in $\text{UD}^{\text{PL}}_{\text{LFG}}$, but no `csubj:pass` subjects,
2. SUBJ may be translated to `csubj` despite a nominal target, namely, when this nominal target actually represents a clausal construction, i.e., is a predicative element in a broadly copular construction (has an outgoing dependency `cop`, `cop:locat` or `aux:pass`),
3. if the origin of the relation is a gerund, the relation is `nmod` (instead of `nsubj`) or `acl` (instead of `csubj`).

The first two points are illustrated with example (7.18), where the matrix subject is a passive clause, headed by the adjectival passive participle. Such adjectival participles receive the UPOS `ADJ` in $\text{UD}^{\text{PL}}_{\text{LFG}}$ (see Figure 7.57). While in most cases `ADJ` subjects would be treated as having the incoming `nsubj` dependency, here the dependency label is `csubj`, as shown in Figure 7.58. Moreover, the subject within this passive clause is marked as `nsubj:pass`.

(7.18) Wydaje mi się, że sytuacja została opanowana.
seems.3SG me.DAT RM COMP situation.NOM.SG.F became.3SG.F controlled.NOM.SG.F
'It seems to me that the situation is under control now.'

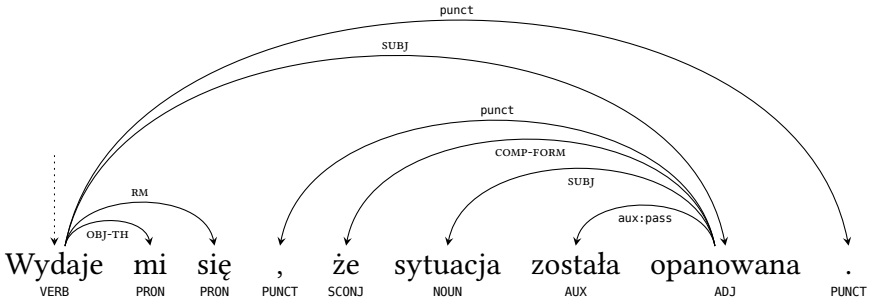


Figure 7.57: Towards UD representation of (7.18) – before converting subjects

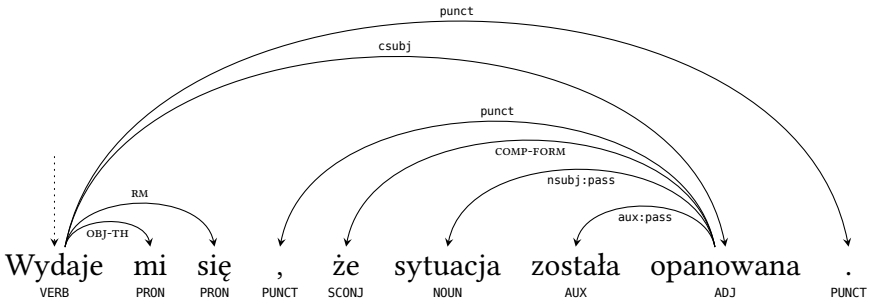


Figure 7.58: Towards UD representation of (7.18) – after converting subjects

The third point may be illustrated with example (7.19). There are two gerunds in this sentence: *uznanie* 'acknowledging' and *istnienia* 'existing, existence'. The first, *uznanie*, takes an argument corresponding to the object of the active verb, namely, *jej istnienia* 'her existence', and the second, *istnienia*, takes an argument corresponding to the subject, *jej* 'her'. In the LFG structure bank such arguments receive the grammatical functions OBJ and SUBJ – see the relevant dependencies in Figure 7.59.[11] On the other hand, since gerunds are treated as `NOUN`s in $\text{UD}^{\text{PL}}_{\text{LFG}}$, their dependents cannot – according to UD guidelines – be marked as `subj`, `obj`, etc., but rather as `nmod` (in the prototypical case, i.e., when these dependents are nominal) or as `acl` (in the case of clausal subject or object). See the result of converting subjects and objects in Figure 7.60.

(7.19) Samo uznanie jej istnienia wymaga
alone.NOM.SG.N acknowledging.NOM.SG.N her.GEN existing.GEN.SG.N requires.3SG
niemal religijnej wiary.
almost religious.GEN faith.GEN
'The acknowledgement of her existence alone requires almost religious faith.'

[11] This representation is not optimal, as *niemal* 'almost' should be analysed as a dependent of *religijnej* 'religious', rather than the current *wiary* 'faith'.

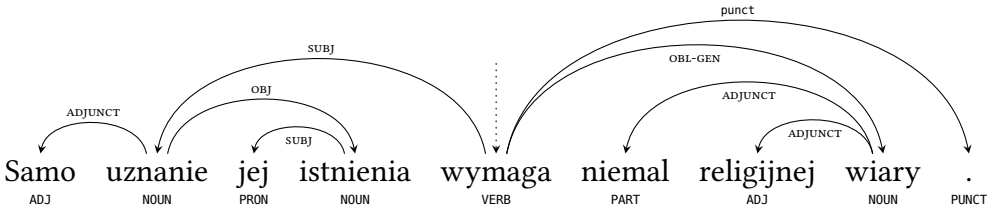


Figure 7.59: Towards UD representation of (7.19) – before converting subjects and objects

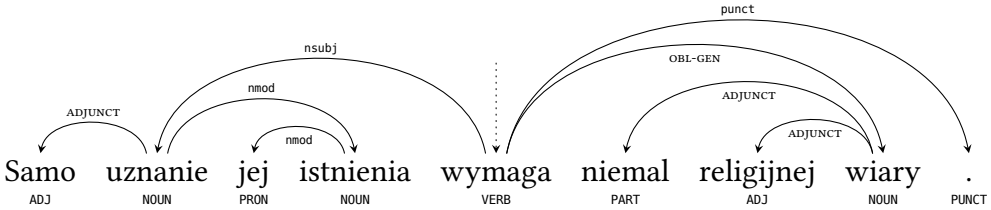


Figure 7.60: Towards UD representation of (7.19) – after converting subjects and objects

Note that there are many other cases where the head is nominal and the SUBJ dependent is still marked as `nsubj` rather than `nmod`. This is the case in copular constructions, where – according to UD guidelines – the nominal predicate is the head, but also in other constructions with predicative nominals, as in example (7.20). There are two dependencies in Figure 7.61 that will eventually be translated into `xcomp`: the XCOMP from the finite verb to its infinitival argument, and the XCOMP-PRED – already translated into `xcomp` (see Section 7.2.4 above) – from the infinitival verb to the nominal predicate *aktorem* 'actor'. All three predicates – two verbal and one nominal – share the same subject, *Solter*, in the enhanced part of the representation, also after the conversion of the subject relation, as shown in Figure 7.62.

(7.20) Soter pragnął zostać aktorem.
Soter.NOM.SG.M strived.3SG.M become.INF actor.INS
'Soter strived to become an actor.'

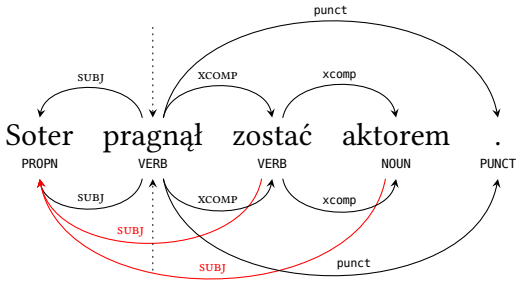


Figure 7.61: Towards UD representation of (7.20) – before converting subjects

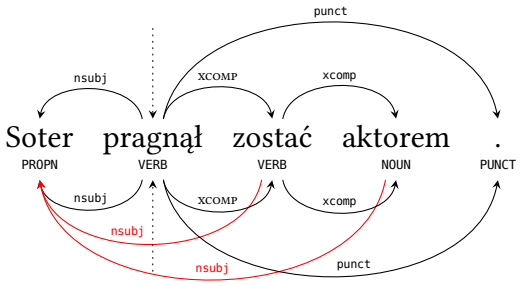


Figure 7.62: Towards UD representation of (7.20) – after converting subjects

Objects

In the LFG structure bank, direct objects have the grammatical function OBJ, and indirect objects – OBJ-TH. In Polish, direct objects are determined on the basis of the passivisation test: whichever argument becomes the subject under passivisation, it is understood as the object in the active construction. This means that not only – and not all – accusative arguments are objects, but also some instrumental and genitive arguments, as well as some clausal arguments.

Unfortunately, while UD allows subjects to be clauses, it assumes that objects must be nominal – the `obj` relation is reserved for nominal dependents only. All clausal arguments are marked as `ccomp`. In $\text{UD}^{\text{PL}}_{\text{LFG}}$, in order to distinguish clausal objects from other clausal arguments (marked as `ccomp`), the former are marked with the label `ccomp:obj`. This is illustrated with example (7.21), where the two objects – clausal in the matrix clause and nominal in the embedded clause (see Figure 7.63) – are translated into `ccomp:obj` and `obj`, respectively (see Figure 7.64).

(7.21) - Sojusz zapowiadał, że poprze reformę
alliance.NOM.SG.M announced.3SG.M COMP support.FUT.3SG reform.ACC
samorządową.
council.ADJ.ACC
'The alliance announced that it will support the council reform.'

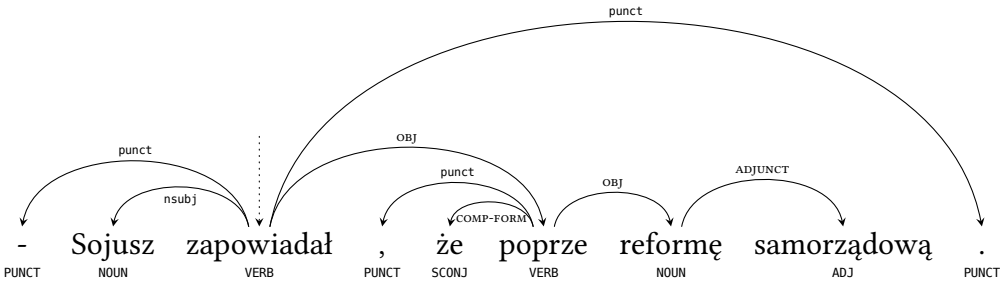


Figure 7.63: Towards UD representation of (7.21) – before converting objects

In contrast to direct objects, indirect objects are determined on the basis of grammatical case and they are defined simply as nominal arguments in the dative case. This definition of the OBJ-TH grammatical function in the LFG structure bank is consistent with the approach to

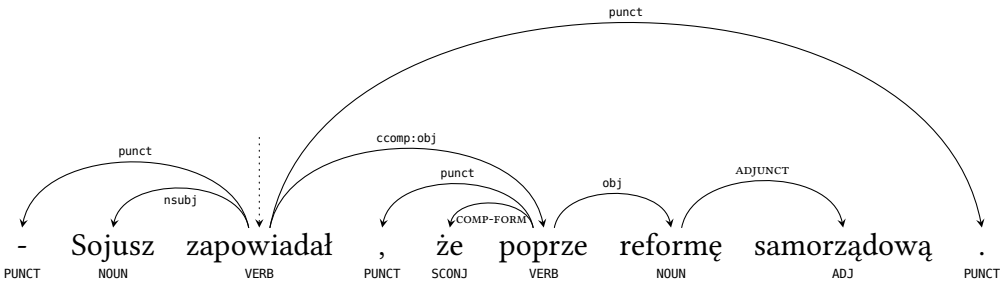


Figure 7.64: Towards UD representation of (7.21) – after converting objects

indirect objects in UD v.2, so OBJ-TH labels are simply translated into `iobj`. As there are no clausal indirect objects, there is no need for `ccomp:iobj` here.

Again, as shown in connection with subjects, when objects – whether direct or indirect – are dependents of gerunds, they are marked as `nmod` (if they are nominal) or as `acl` (if the direct object is clausal). On the other hand, since adjectival participles are treated as reduced relative clauses (see Adjuncts below), their dependents are treated as if they were dependents of verbs, even though such participles have the UPOS `ADJ`. Hence, objects of such adjectival participles are marked as `obj` or `iobj`, and not as `nmod` (or `acl`).

Obliques

In the original LFG structure bank, there is a large number of oblique grammatical functions, including those distinguished semantically: OBL-LOCAT (for locative arguments), OBL-ADL (adlative), OBL-TEMP (temporal), OBL-MOD (manner), etc., but also indicating the grammatical case of the argument, e.g., OBL-INST (instrumental) or OBL-STR (structural, i.e., accusative or genitive, depending on the presence of negation and other factors).

The semantic subtypes of oblique arguments are lost in the conversion. The main reason is not that they could not be represented in UD – they could with the help of language-specific extensions – but that similar information is not available in the case of adjuncts, which in UD are not distinguished from obliques. Hence, preserving semantic subtypes of oblique arguments would result in inconsistency: only some of broadly understood oblique dependents (arguments or adjuncts) would have such information, e.g., only some temporal dependents would be marked as such. Thus, in the basic dependency tree, all LFG grammatical relations starting with OBL are translated into the UD relations `obl`, `advmod` or `nmod`, depending on parts of speech of the head and the dependent (see below for details). The only exception to this rule is OBL-AG, expressing the demoted agent in passive constructions, translated to `obl:agent`.

On the other hand, in the case of the enhanced representation, the resulting `obl` and `nmod` relations may be subtyped with the adposition, in case the dependent is an adpositional phrase. This, and various ways of mapping OBL relations into UD, is illustrated with example (7.22), involving three oblique arguments: two arguments of the finite form of the verb BYĆ 'be', and an argument of the gerund form *meldowaniem* 'checking in' – see Figure 7.65. These oblique arguments are translated into `advmod`, `obl` (enhanced to `obl:z`) and `nmod` (enhanced to `nmod:w`) – see Figure 7.66.

(7.22) - Ale jak będzie z meldowaniem w hotelu?
but how be.FUT with checking_in in hotel
'But what shall we do about checking in at the hotel?'

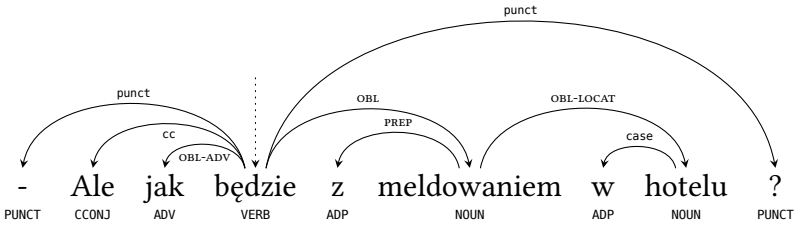


Figure 7.65: Towards UD representation of (7.22) – before converting obliques

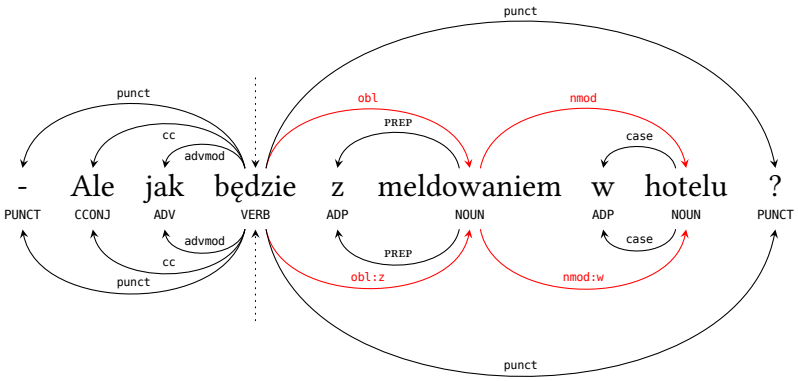


Figure 7.66: Towards UD representation of (7.22) – after converting obliques

The decision on whether an OBL relation is translated into `obl`, `advmod` or `nmod` is made as follows:

- translate to `obl` if the head's UPOS is `VERB`, `ADJ` or `ADV`, and the dependent's UPOS is broadly nominal, i.e., one of: `NOUN`, `PRON`, `PROPN`, `DET`, `ADJ`, `NUM`,
- translate to `advmod` if the head's UPOS is `VERB`, `ADJ` or `ADV`, and the dependent's UPOS is `ADV`,
- translate to `nmod` if the head's UPOS is `NOUN`, and the dependent's UPOS is broadly nominal (as defined above).

There are two reasons for lumping `ADJ` and `ADV` together with `VERB` above. First, adjectival participles are marked with the UPOS `ADJ`, but – just in the case of object dependents of such participles discussed above – oblique arguments of adjectival participles are treated as if they were dependents of verbs, i.e., as `obl` (or `advmod`) rather than `nmod`. Second, the other situation where adjectives and adverbs may have oblique arguments is when they are heads of comparative constructions (in which case the original relation is OBL-COMPAR). In such cases, the dependent should be marked as an `obl`, according to the UD guidelines.[12] This is illustrated with example (7.23) and the 'before' and 'after' Figures 7.67–7.68.

[12] `http://universaldependencies.org/u/overview/specific-syntax.html#comparatives`

(7.23) To było silniejsze od ciebie?
this.NOM.SG.N was.3SG.N stronger.NOM.SG.N than you.GEN
'Was this stronger than you?'

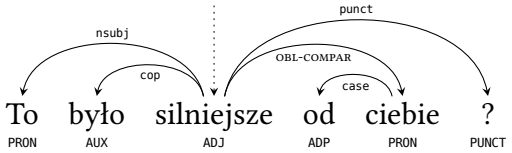


Figure 7.67: Towards UD representation of (7.23) – before converting obliques

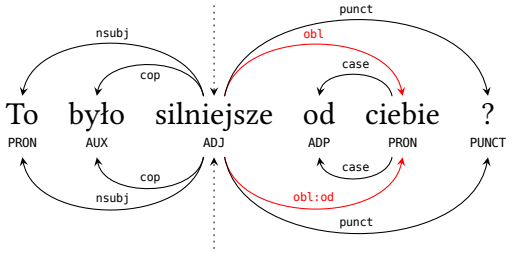


Figure 7.68: Towards UD representation of (7.23) – after converting obliques

Clausal arguments

Those clausal arguments which do not have any more specific grammatical functions (i.e., which are not subjects or objects) bear the COMP relation in the LFG structure bank. This grammatical function is almost invariably translated into the UD relation `ccomp`. One exception is when the source of the dependency relation is a noun or a pronoun, especially, in so-called correlative constructions, where a pronoun introduces the subordinate clause; in such a case the appropriate relation is `acl`. Both the prototypical situation and this exception are present in example (7.24), as shown in Figures 7.69–7.70. Another exception is when the source of the dependency is an adjective; in this case the usual UD relation is `advcl` (as suggested by Joakim Nivre, p.c.), perhaps subtyped by the complementiser in the enhanced representation.

(7.24) Przypuszczam, że chodzi raczej o to, iż wybrał się samowolnie!
presume.1SG COMP goes rather about this COMP set_off.3SG.M RM lawlessly
'I presume that it's rather about the fact that he set off without permission!'

Open (controlled) clausal arguments

In control and raising constructions the open (controlled) clausal argument bears the XCOMP grammatical function in the LFG structure bank. As the UD `xcomp` is taken directly from LFG, the XCOMP relation is trivially translated into `xcomp`, without any additional conditions attached. Note, however, that – as discussed in Section 7.2.4 above (see page 141) – also some instances of the XCOMP-PRED LFG relation are translated into UD as `xcomp`.

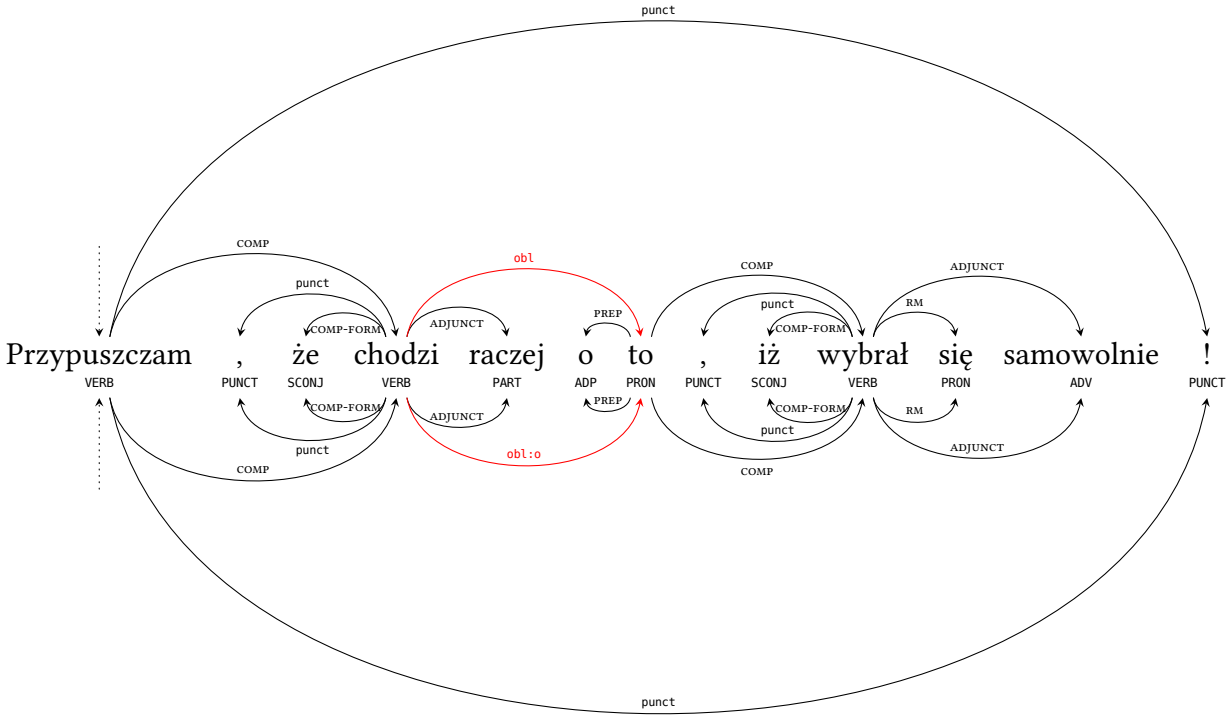


Figure 7.69: Towards UD representation of (7.24) – before converting clausal arguments

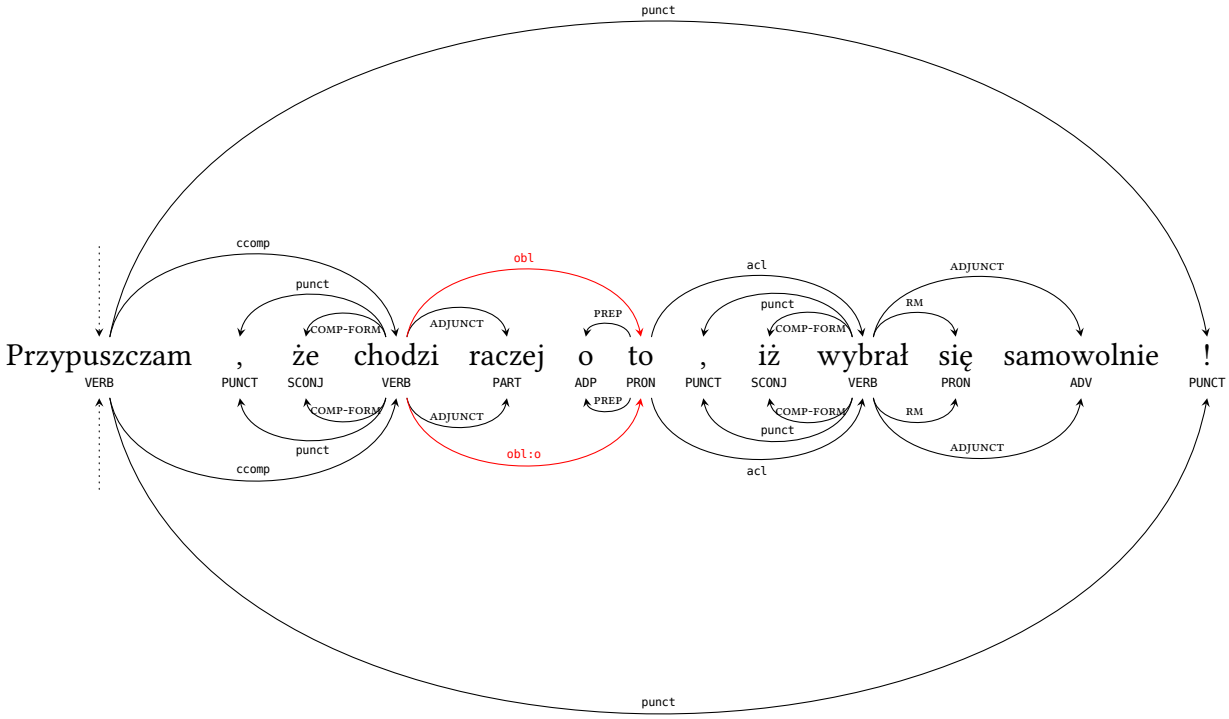


Figure 7.70: Towards UD representation of (7.24) – after converting clausal arguments

Adjuncts

Similarly to the translation of dependency labels of oblique arguments, conversion of adjuncts depends on whether the head is broadly verbal (`VERB`, `ADV` or an adjectival participle) or broadly nominal (`NOUN`, `PROPN`, `PRON`, `DET`, `NUM`, non-participial `ADJ`).

In the former case, when the head is broadly verbal:

- if the dependent is a `VERB`, translate ADJUNCT to `advcl`, or – only in the enhanced representation – a subtype of this relation containing information about the complementiser,
- if the dependent is an adverb (`ADV`) or a particle (`PART`), translate it to `advmod`,
- if the dependent is broadly nominal:
  - in case it is vocative (bears the vocative case or is a proper noun in the nominative case),[13] translate ADJUNCT to `vocative`,
  - otherwise translate it to `obl`.

These four possible translations of ad-verbal ADJUNCTs are illustrated with examples (7.25) and (7.26), and their respective 'before' and 'after' dependency structures in Figures 7.71–7.72 and 7.73–7.74.

(7.25) Tu, bracie, obcujesz z przyrodą.
here brother.VOC.SG.M commune.2SG.M with nature.INS
'Here, my brother, you commune with nature.'

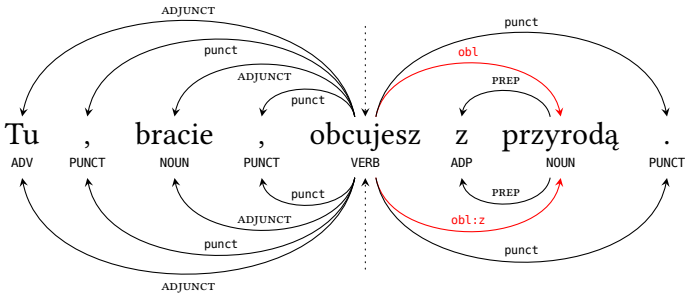


Figure 7.71: Towards UD representation of (7.25) – before converting adjuncts

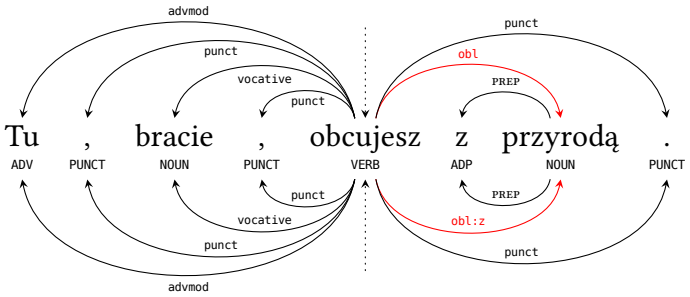


Figure 7.72: Towards UD representation of (7.25) – after converting adjuncts

[13] This rule of thumb concerning the treatment of proper name adjuncts in the nominative as functionally vocative, is relatively robust: about 90% of dependents it classifies as vocative are indeed functionally vocative. Removing the proper name condition would significantly lower the precision: only about 15% of such nominative non-proper nominal adjuncts are functionally vocative.

(7.26) Radujmy się z nimi, bo żyją!
rejoice.IMP.1PL RM with them because live.3PL
'Let's rejoice with them, because they are alive!'

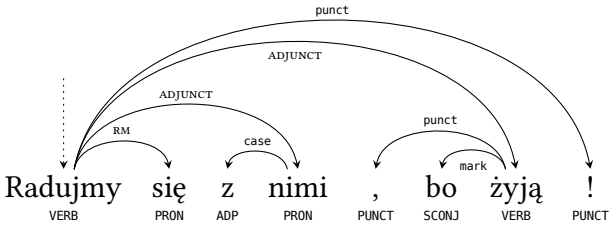


Figure 7.73: Towards UD representation of (7.26) – before converting adjuncts

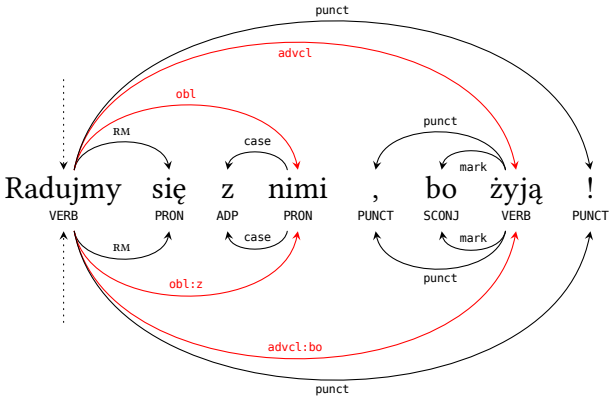


Figure 7.74: Towards UD representation of (7.26) – after converting adjuncts

In the case of broadly nominal heads of the ADJUNCT relation, this label may be translated to:

- `det`, if the dependent's UPOS is `DET`,
- `amod`, if the dependent is marked as `ADJ` and it is not an adjectival participle,
- `acl`, if
  - the dependent is a `VERB`,
  - or the dependent is an adjectival participle; such adjectival participles, despite the fact that their UPOS is `ADJ`, are treated here on a par with reduced relative clauses (as suggested to us by Joakim Nivre, p.c.); on the other hand, they are not marked as `acl:relcl`, as this dependency label is reserved for true relative clauses,
- `acl:relcl`, if the dependent is a relative clause,
- `advmod`, if the dependent is an adverb (`ADV`) or a particle (`PART`),
- `nmod`, if the dependent is nominal (`NOUN`, `PROPN`, `PRON`, `NUM`), perhaps with an appropriate subtype indicating an adposition.

An exception is made for those `DET` and `ADJ` dependents which themselves have a `case` dependency, i.e., which are arguments of prepositions: such dependents are very likely to be elective or otherwise represent nominal constructions, so the ADJUNCT label is translated to `nmod` with an appropriate subtype indicating the preposition. An example illustrating conversion of four

ad-nominal ADJUNCT dependencies to two nmods, an amod and a det is (7.27), with the 'before' and 'after' Figures in 7.75–7.76.

(7.27) Inny produkt z tej serii to torba na zakupy.
another.NOM.SG.M product.NOM.SG.M from this.GEN series.GEN is bag.NOM.SG.F for shopping.ACC
'Another product in this series is a shopping bag.'

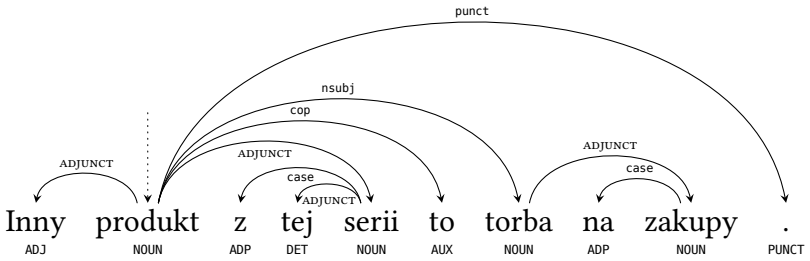


Figure 7.75: Towards UD representation of (7.27) – before converting adjuncts

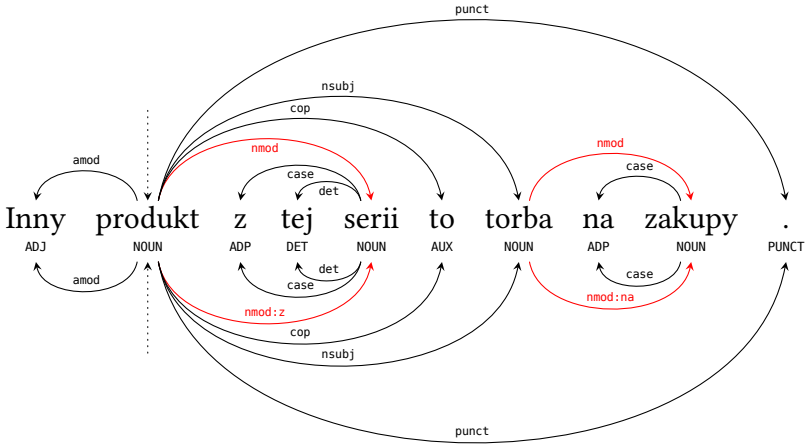


Figure 7.76: Towards UD representation of (7.27) – after converting adjuncts

Apart from ADJUNCT, there is another grammatical function for a particular kind of adjuncts, namely POSS, for possessive modifiers, trivially translated to the standard UD relation nmod:poss.

Open (controlled) adjuncts

A separate grammatical function, XADJUNCT, is used in the LFG structure bank to represent open adjuncts, i.e., adjuncts whose subject is obligatorily controlled by another element in the clause. There are two situations when this grammatical function is used: to mark adverbial participles and to mark secondary predicates. Both are illustrated with example (7.28), which

involves an adverbial participial modifier, *nie ważąc się…* ‘not daring…’, and a secondary predicate, *niezdecydowani* ‘undecided’. Both are originally marked as XADJUNCT (see Figure 7.77) and both are translated to `advcl` (see Figure 7.78).

(7.28) Stali dłuższą chwilę niezdecydowani, nie ważąc się na ryzykowny
stood.3PL.M longer.ACC while.ACC undecided.NOM.PL.M NEG daring RM on risky
krok.
step
‘They stood for a longer while, undecided, not daring to take the risky step.’

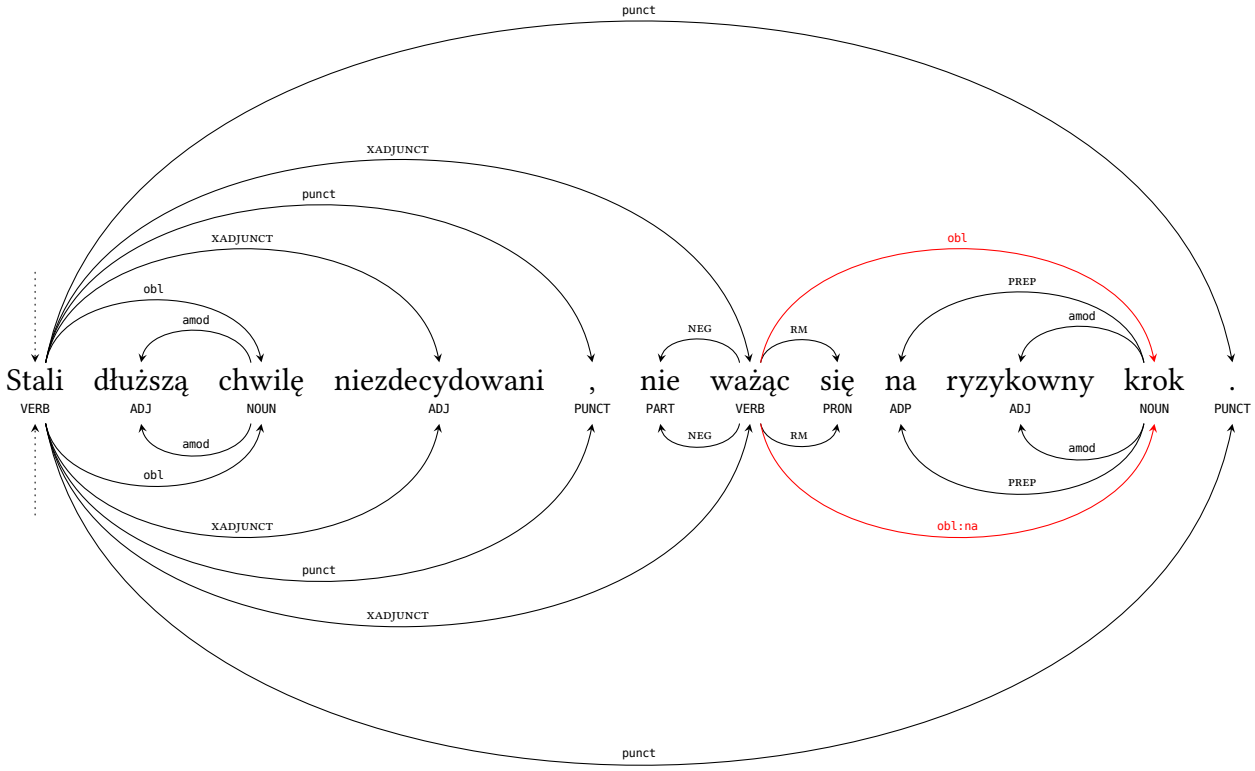


Figure 7.77: Towards UD representation of (7.28) – before converting open adjuncts

Not all secondary predicates are translated as `advcl`, only those which attach to the main verb. This is the case above: while the secondary predicate *niezdecydowani* ‘undecided’ refers to the subject of the verb *stali* ‘stood’, this subject is not overtly realised (it is *pro*-dropped), so – in compliance with UD guidelines – the secondary predicate attaches to the verb and bears the `advcl` relation. If the subject were overtly realised, the predicate would be a dependent of this subject, with the dependency label `acl`. This is illustrated with example (7.29) and Figures 7.79–7.80. Here, the secondary predicate *pierwszy* ‘first’ is in direct relation with the subject, *król* ‘king’. Note that, as a result of this conversion step, the number of enhanced dependencies is reduced: according to the UD guidelines, the fact that *król* ‘king’ is understood as the subject of the secondary predicate does not have to be represented directly, as it is inferable from the `acl` relation between these two words.

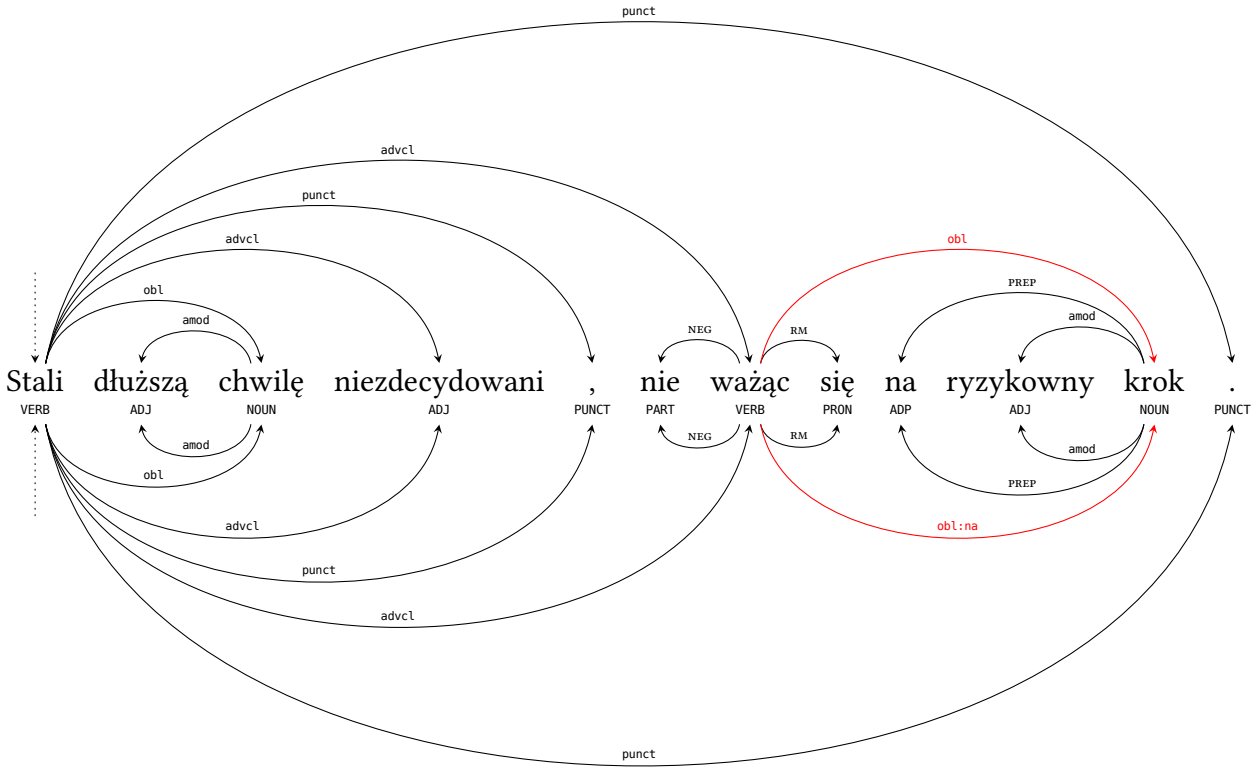


Figure 7.78: Towards UD representation of (7.28) – after converting open adjuncts

(7.29) Król zaatakował pierwszy.
king.NOM.SG.M attacked.3SG.M first.NOM.SG.M
'The king attacked (as) first.'

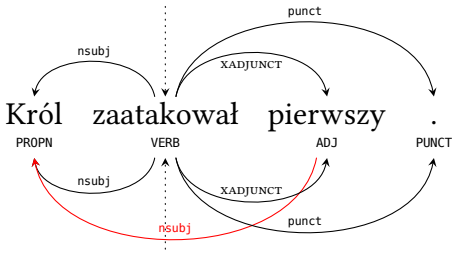


Figure 7.79: Towards UD representation of (7.29) – before converting open adjuncts

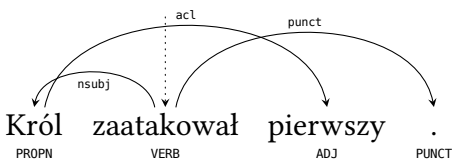


Figure 7.80: Towards UD representation of (7.29) – after converting open adjuncts

7.2.6 Other dependency relations

Of the remaining labels from the initial dependency representation, almost all are translated into UD relation via the deterministic mapping shown in Table 7.1. Two more, RM and APP, are a little more complicated and will be discussed below.

Table 7.1: Deterministic translation of some of the initial dependency labels into their UD equivalents

initial label	UD label
ROOT	root
AUX	aux
AGLT	aux:aglt
MM	aux:mood
NEG	advmod
CNEG	advmod
COMP-FORM	mark
RSM	mark
QUB[INT]	mark
PREP	case

ROOT

There is exactly one token marked as ROOT in the initial dependency representation of each sentence, and exactly one marked as root in the final UD representation. Note that these may be two different tokens: as described in Section 7.2.4, incoming dependencies – so also the ROOT dependency – may be moved to other tokens in the process of reversing dependency relations. This is illustrated above with the example (7.15) and the dependency representations in Figures 7.50–7.51 (pages 139–140), and similarly with the example (7.16) and Figures 7.53–7.54 (pages 141–142).

AUX, AGLT and MM

Apart from standard auxiliaries in periphrastic future tense, in passive constructions, etc., there are two additional kinds of functional elements treated as auxiliaries. The first one is the mobile inflection (see Section 5.1) expressing number and person; it is marked as aux:aglt, i.e., with the language-specific subtype aglt. The other type consists of two particles expressing mood: BY, expressing the conditional, and NIECH (and its variant NIECHAJ), expressing the imperative mood. Such particles are treated as aux:mood auxiliaries, where mood is a language-specific subtype of the aux relation. These two language-specific subtypes of aux are illustrated with example (7.30) and Figures 7.81–7.82. As these figures also show, not all tokens with the UPOS AUX are aux dependents: the initial token, while marked as AUX, bears the cop dependency.

(7.30) Byłbym bardziej autentyczny.
be.COND.1SG.M more authentic.NOM.SG.M
'I would be more authentic.'

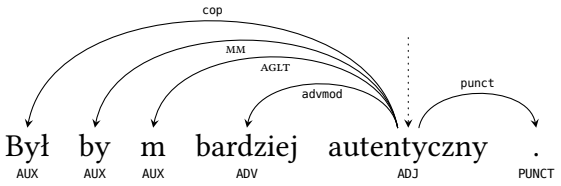


Figure 7.81: Towards UD representation of (7.30) – before converting other dependency relations

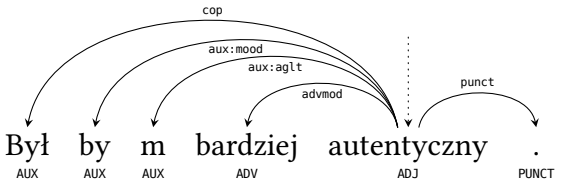


Figure 7.82: Towards UD representation of (7.30) – after converting other dependency relations

NEG and CNEG

The negation particle, NIE, whose initial dependency is either NEG (eventuality negation) or CNEG (constituent negation), is simply translated into `advmod`, with no information lost about the fact that this `advmod` expresses negation – the fact that this is the negation particle can be decoded from its lemma and its UPOS `PART` – but with information lost about the eventuality (verbal, sentential) vs. constituent status of this negation (Przepiórkowski and Patejuk 2015).

COMP-FORM, RSM and QUB[INT]

The UD relation `mark` is used to indicate a complementiser. Semantic complementisers, initially solely heading the subordinate clause, are made into `mark` dependents in the dependency reversing step – see Section 7.2.4. Asemantic complementisers are dependents marked with the COMP-FORM relation, so this relation should now be translated into `mark`. Additionally, two specific complementiser-like elements are singled out with the initial relations RSM and QUB[INT]. The former is the functional element CO introducing a certain kind of relative clauses, namely, relative clauses which may contain resumptive pronouns. The latter is the question particle CZY. All these functional elements are already dependents of the head of the clause, so at this stage it is sufficient to change the names of their labels to `mark`.

PREP

Again, semantic prepositions, initially solely heading the prepositional phrase, are dealt with in the dependency reversing step. What is left is asemantic prepositions, which – as a result of the process of finding the true head among co-heads (see Section 7.1.1) – are turned into dependents of the nominal heads (see Section 7.1.3). Such asemantic prepositions bear the PREP relation to their heads, so this relation must now be translated to `case`.

RM

The various functions of the so-called reflexive marker SIĘ (see, e.g., Patejuk and Przepiórkowski 2015a and references therein), initially marked with the RM dependency, are clustered into three dependency labels in $\text{UD}^{\text{PL}}_{\text{LFG}}$:

- `expl:pv`, a subtype of `expl` already used in a number of UD treebanks to indicate the inherent use of the reflexive marker – in such cases SIĘ is a part of the verbal lemma,[14]
- `expl:impers`, a subtype of `expl` used earlier in Italian and Romanian treebanks to indicate the use of the reflexive marker in impersonal constructions,[15]
- `obj`, for those uses of SIĘ which correspond to direct objects.

Sentences (7.31)–(7.32) illustrate the three kinds of SIĘ. In fact, example (7.31) also illustrates the phenomenon of the haplology of the reflexive marker, where one occurrence of *się* simultaneously plays two roles (Kupść 1999; Patejuk and Przepiórkowski 2015a). This is the case with the sequence *modliło się* 'one would pray', which is an impersonal form of the inherently reflexive verb MODLIĆ SIĘ 'pray'. Since it is not possible to have two labels on a single relation between two tokens, the dependency between *modliło* and *się* is marked as `expl:pv` in Figure 7.84 (to be compared with Figure 7.83, showing the input to this conversion step). On the other hand, in *uczestniczyło się* 'one would participate', *się* is unequivocally impersonal, so it is marked as `expl:impers`.

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

[14] `http://universaldependencies.org/cs/dep/expl-pv.html`
[15] `http://universaldependencies.org/it/dep/expl-impers.html`

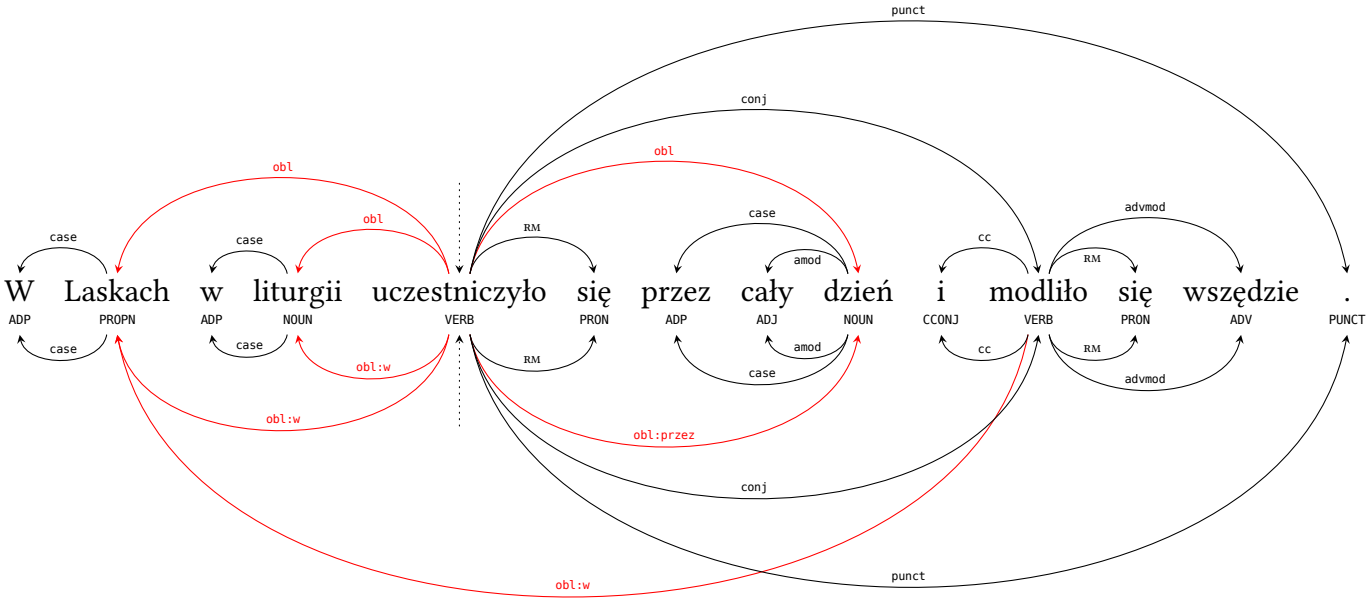


Figure 7.83: Towards UD representation of (7.31) – before converting other dependency relations

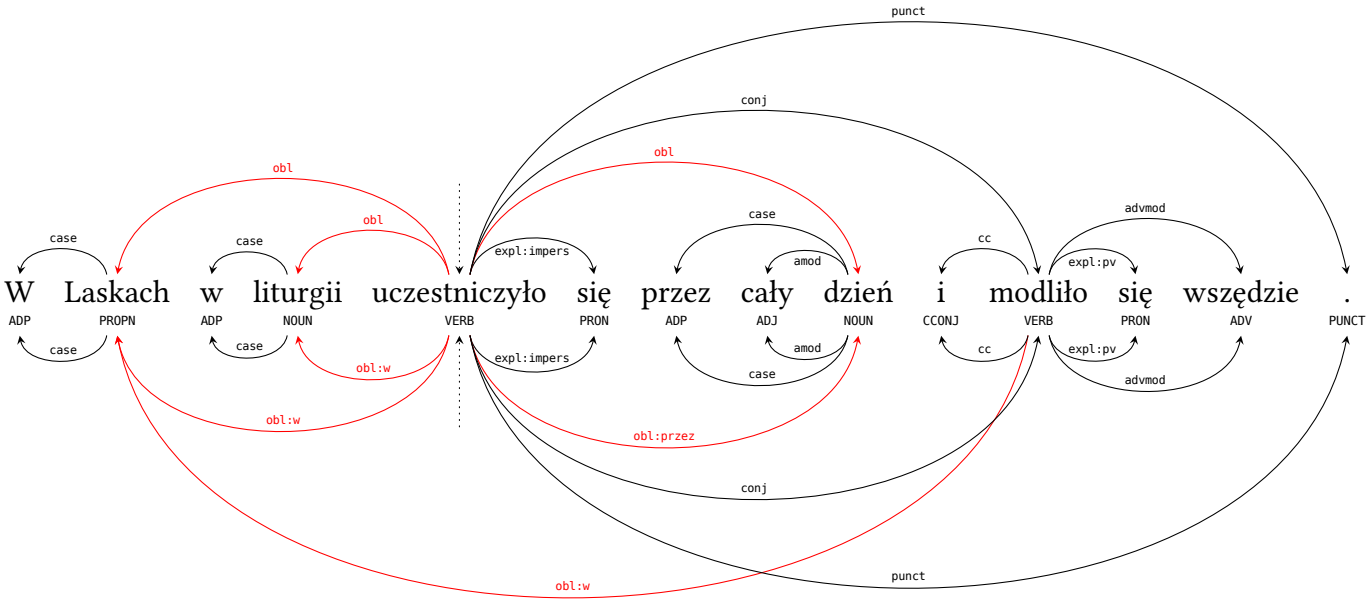


Figure 7.84: Towards UD representation of (7.31) – after converting other dependency relations

The third possibility is illustrated with example (7.32), where the verb UKRYĆ ‘hide’ takes the reflexive *się* instead of a direct object; see Figures 7.85–7.86.

(7.32) A myśl ukryła się w tłumie.
and thought.NOM.SG.F hid.3SG.F RM in crowd
‘And the thought concealed itself in the crowd.’

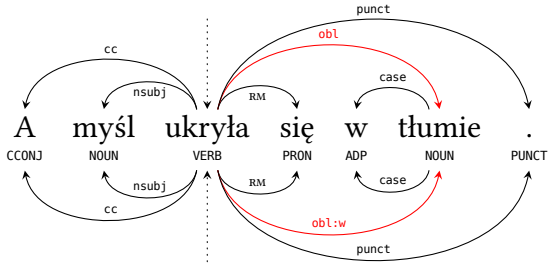


Figure 7.85: Towards UD representation of (7.32) – before converting other dependency relations

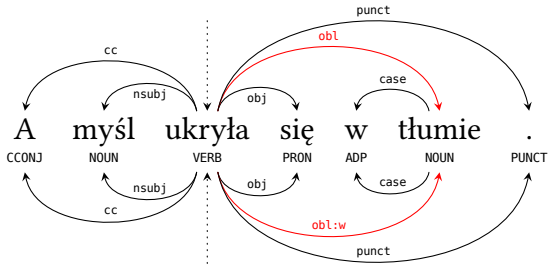


Figure 7.86: Towards UD representation of (7.32) – after converting other dependency relations

APP

In the LFG structure bank, the ‘grammatical function’ APP combines the roles of two UD relations: `appos` and `flat`. The former, `appos`, is used in cases of ordinary apposition, where the apposed constituents may in principle be reversed. The latter, `flat`, is used in the case of names and other apposition-like constructions, where the two parts should occur in the fixed order. In this conversion step some attempt is made to recover such `flat` relations from those marked as APP. To this end the following rule of thumb is used:

- translate APP to `flat` if the dependent is a proper noun, or the head is a form of PAN ‘Mr.’ or PANI ‘Ms.’, or the ⟨head, dependent⟩ pair belongs to a small dictionary of known `flat` pairs (e.g., *inżynier metalurg* ‘metallurgist’),
- otherwise translate it to `appos`.

The precision of this rule of thumb is very high: well over 95% of appositions classified as `flat` have a rigid word order. The recall is much lower, as about 50% of the remaining appositions, classified as `appos`, have a relatively fixed word order. Both translations of APP are illustrated with example (7.33) and the ‘before’ and ‘after’ representations given in Figures 7.87–7.88.

Here, *Lech Kaczyński*, consisting of the first name and the surname, is a relatively fixed apposition of the `flat` kind, and the larger *Lech Kaczyński, prezydent RP* 'Lech Kaczyński, the president of the RP', is a typical apposition with mutable word order.

(7.33) Zdecydował o tym Lech Kaczyński, prezydent RP.
decided.3SG.M about this Lech.NOM.SG.M Kaczyński.NOM.SG.M president.NOM.SG.M RP.GEN
'It was decided by Lech Kaczyński, the president of the Republic of Poland.'

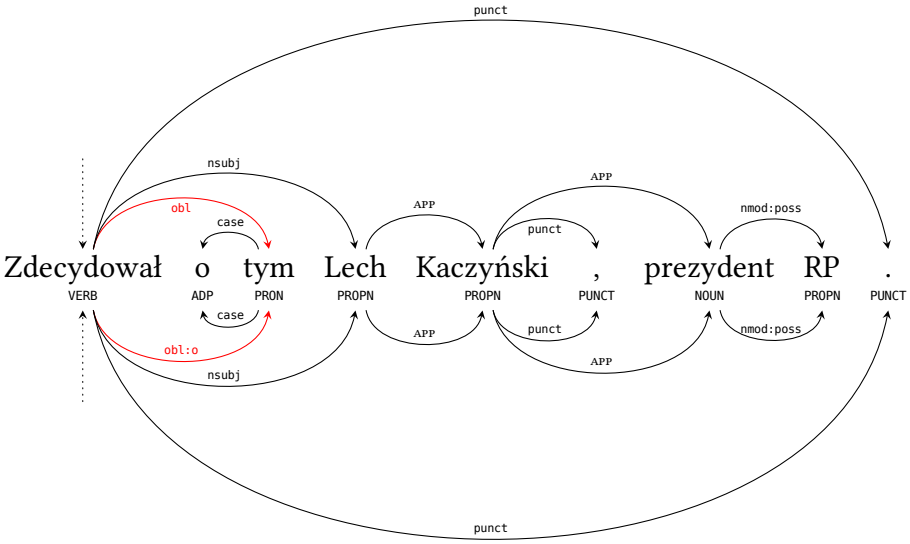


Figure 7.87: Towards UD representation of (7.33) – before converting other dependency relations

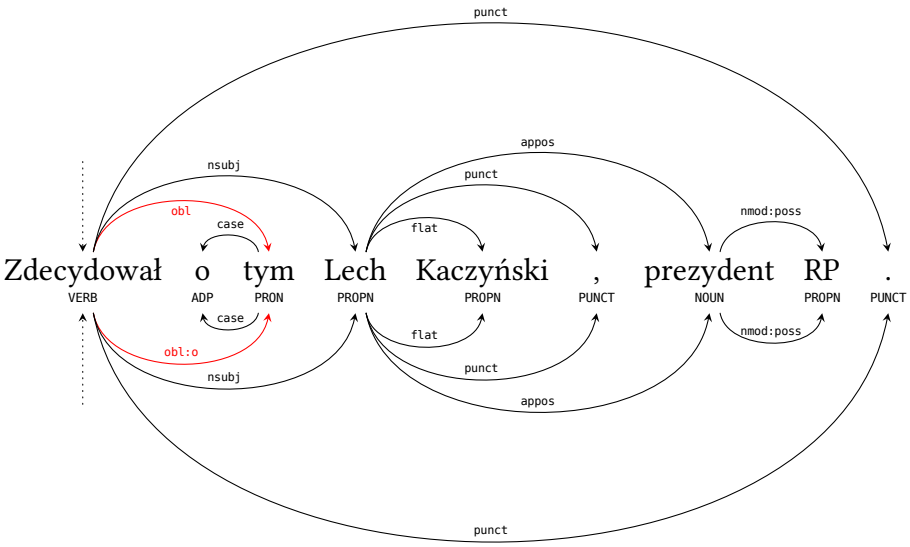


Figure 7.88: Towards UD representation of (7.33) – after converting other dependency relations

Note that this step may also require rearranging dependency structures. This is because, in the LFG structure bank, appositions form chains, while in UD, the first element of an apposition, whether `appos` or `flat`, governs all other elements. The exact rearranging rules are rather subtle. For example, in Figures 7.87–7.88 the origin of the APP dependency between *Kaczyński* and *prezydent RP*, renamed to `appos`, is moved to *Lech*, which is the `flat` governor of *Kaczyński*. But the dual operation, moving the `flat` dependent up the `appos` dependency, would usually give wrong results, as exemplified by (7.34). There, as shown in Figures 7.89–7.90, the `appos` relation holds between two `flat` constituents, *Henryk Sadurski* and *inżynier metalurg* 'metallurgy engineer', so moving the second `flat` dependency up would result in a `flat` dependency between *Henryk* (and *Sadurski*) and *metalurg* 'metallurgist', with the exclusion of *inżynier* 'engineer', related to *Henryk* via `appos`.

(7.34) Henryk Sadurski, inżynier metalurg, nie ma pracy od kilku lat.
Henryk.NOM.SG.M Sadurski.NOM.SG.M engineer.NOM.SG.M metallurgist.NOM NEG has.3SG work.GEN from several years
'Henryk Sadurski, a metallurgy engineer, has been unemployed for a few years.'

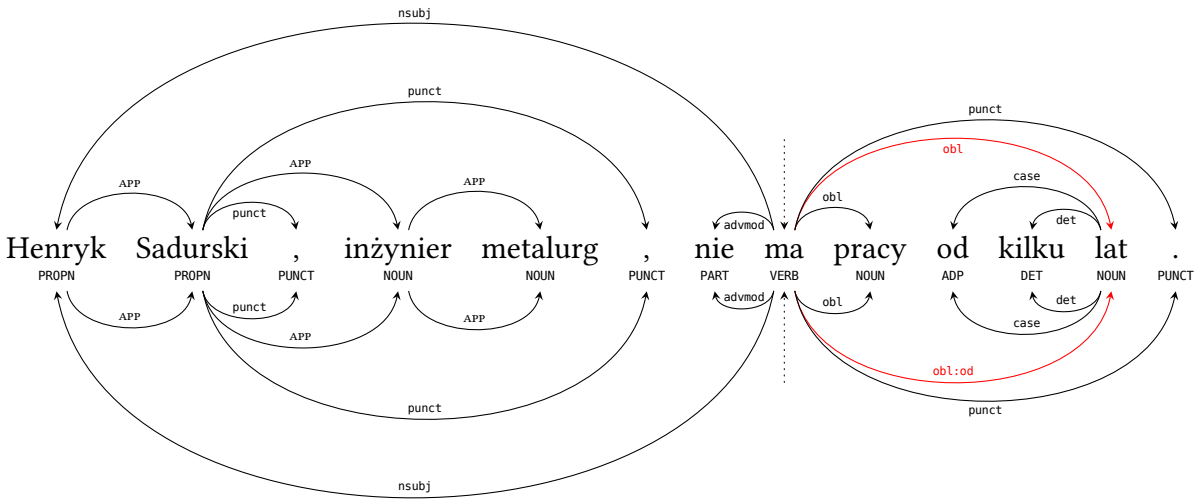


Figure 7.89: Towards UD representation of (7.34) – before converting other dependency relations

7.2.7 Propagating coordination

The final conversion step consists in propagating coordination in the enhanced representation: if the whole coordinate structure is a dependent, the dependency targets the head of this coordinate structure – the first conjunct – in the basic tree, but it should also target the other conjuncts in the enhanced representation. For example, in the case of (7.35), one more enhanced dependency should be added to those in Figure 7.91, as shown in Figure 7.92. (For completeness, the f-structure of this sentence is given in Figure 7.93 and the initial dependency structure – in Figure 7.94.)

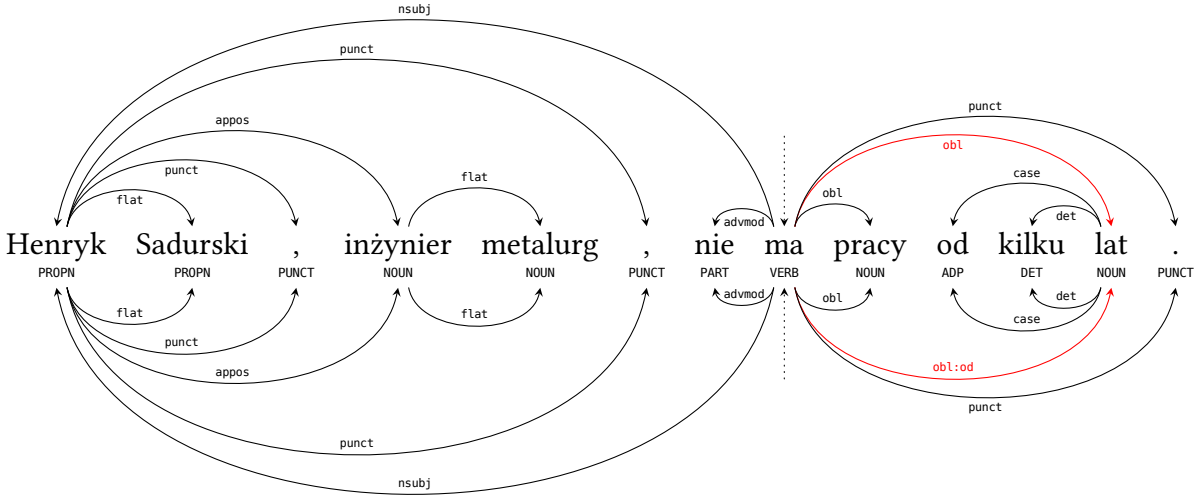


Figure 7.90: Towards UD representation of (7.34) – after converting other dependency relations

(7.35) Nie mieszkam też w Wenecji czy Paryżu.
NEG live.1SG also in Venice or Paris
'I also don't live in Venice or Paris.'

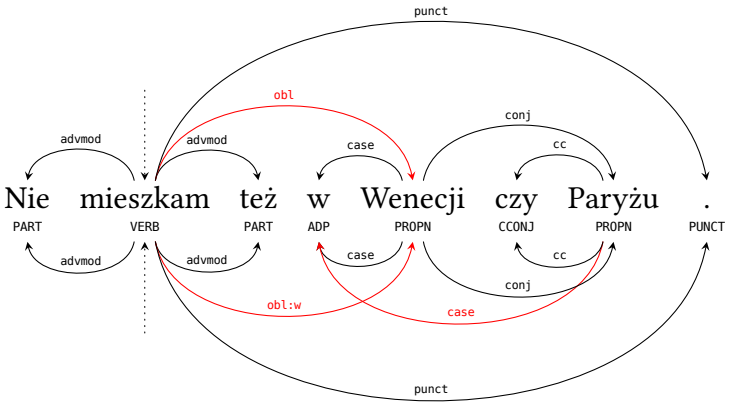


Figure 7.91: Towards UD representation of (7.35) – before propagating coordination

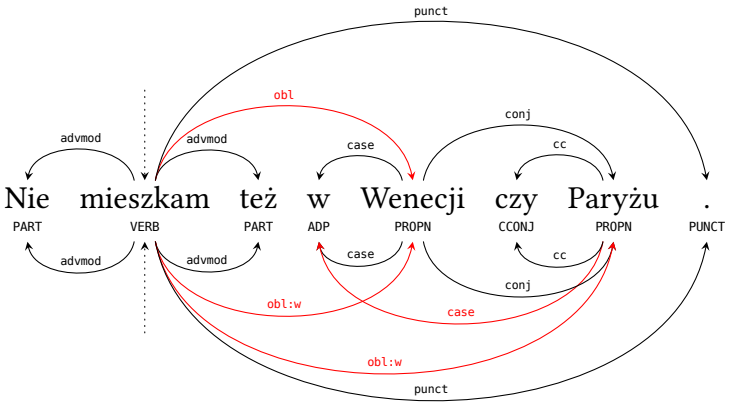


Figure 7.92: Towards UD representation of (7.35) – after propagating coordination

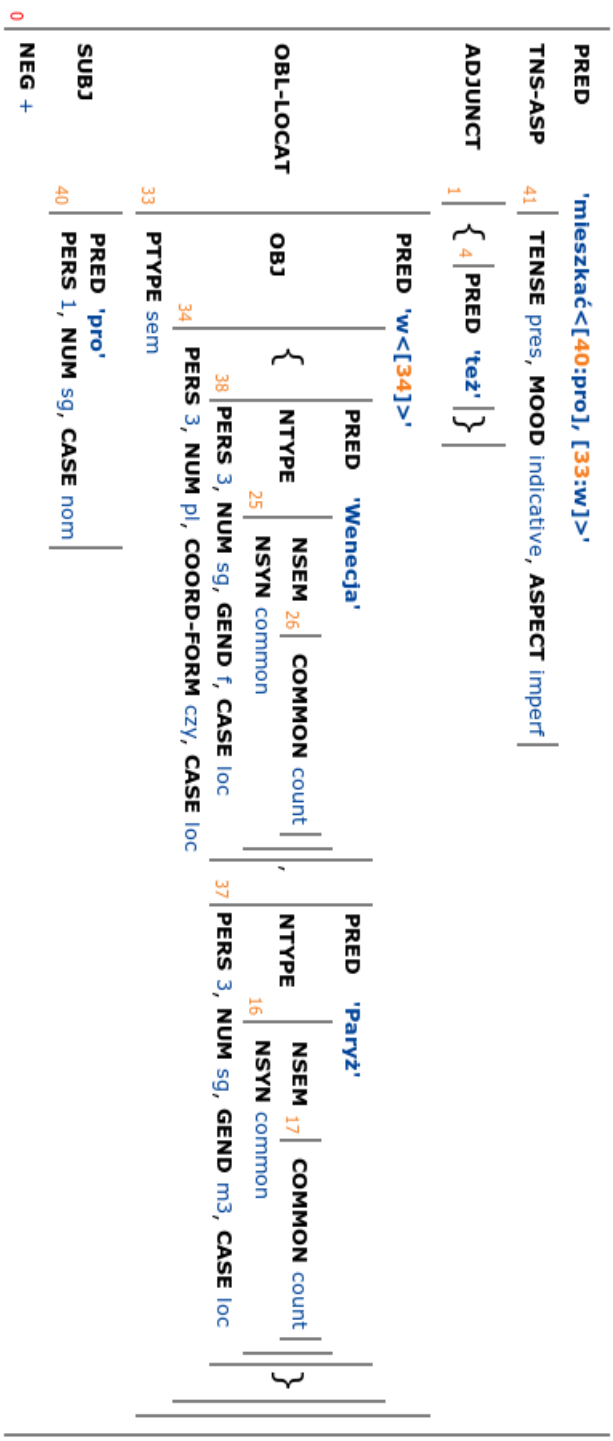


Figure 7.93: F-structure of (7.35)

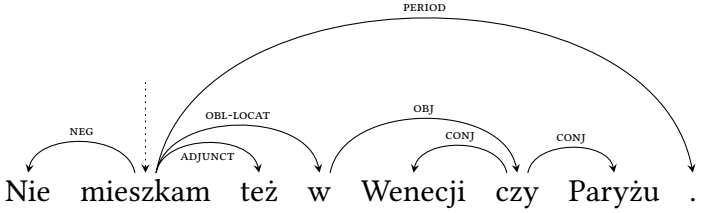


Figure 7.94: Initial dependency representation of (7.35)

As dependent-sharing is handled at earlier stages, nothing more needs to be done to propagate coordination in enhanced representations. This step ends the conversion procedure; the resulting treebank is described in the following chapter.

Part III

Enhanced UD Treebank of Polish

Chapter 8

Enhanced UD Treebank of Polish

The aim of this chapter is summarise the main features of the $\text{UD}^{\text{PL}}_{\text{LFG}}$ treebank of Polish. Many of these were mentioned in, or can be inferred from, the preceding chapters, but here they are presented in a way that does not require any knowledge of the input LFG structure bank or the conversion procedure.[1]

8.1 Tokenisation

Tokenisation follows the principles of the previous corpora of Polish (Przepiórkowski 2004b; Przepiórkowski et al. 2012), i.e., tokens never contain any spaces (or other whitespace characters), but some "orthographic words" (i.e., words "from space to space") may be split into smaller tokens. This happens in two broad kinds of situations.

First, there is a closed class of "orthographic words" such as *weń* 'in him(/it/her)' or *doń* 'to him(/it/her)', which consist of a pronominal form (*we*, *do*) and a short postprepositional form of the personal pronoun (*ń*), prescriptively interpreted as masculine,[2] but occurring in texts also with neuter and feminine references. Such "orthographic words" are split into two, and tagged separately as an adposition and a pronoun.

Second, conditional particles *by* and "mobile inflections" expressing person and number, e.g., *śmy* '1PL', are separated from the words they attach to. For example, *przyszlibyśmy* 'we would have come', is split into *przyszli*, *by* and *śmy*, with appropriate morphosyntactic information: *przyszli* is treated as a finite verb with appropriate features (including aspect, gender and number), *by* – as a conditional auxiliary with mood as the only feature, *śmy* – as an auxiliary with number and person (but not gender) among its features.

"Words with spaces" are represented as separate tokens connected with the `fixed` dependency relation and all tagged with the same morphosyntactic information, namely, information that pertains to the whole "word with spaces". For example, the complex preposition *w czasie*

[1] Occasional references are made below to the CoNLL-U representation of UD structures; see Section 4.3 and `http://universaldependencies.org/format.html`.

[2] See, e.g., `http://sjp.pwn.pl/poradnia/haslo/;6283` (in Polish).

'during', is split into *w* 'in' and *czasie* 'time' and both tokens are tagged as an adposition combining with the genitive case, even though, in separation, *w* is a preposition combining with the locative or the accusative, and *czasie* is a noun in the locative case.

Punctuation marks are usually separate tokens, unless they are integral parts of a word. This can occur in two situations. First, some stems contain punctuation, so all forms of such words contain this punctuation. Typical examples are certain proper nouns, e.g., *Rolls-Royce*, *O'Donell* and *Yahoo!*, but also some common (often morphologically complex) words may contain hyphens, e.g., *e-mail*, *stop-klatka* 'freeze-frame' or *22-latek* 'a/the 22-year old' . Second, inflectional affixes may be added with the help of one of two punctuation marks: the apostrophy (in the case of certain foreign words), as in *ragtime'y* 'ragtimes', or the hyphen (in the case of certain acronyms), as in *SMS-a* 'SMS.GEN/ACC'.

In the CoNLL-U representation used in UD, there is also a feature in the MISC column which is relevant for tokenisation, namely `SpaceAfter=No`. As the name suggests, it appears on tokens directly followed by other tokens, with no intervening whitespace. (Other tokens do not have the `SpaceAfter` feature at all.)

8.2 Morphosyntax

Three columns in the CoNLL-U representation contain morphosyntactic information:

- UPOS: a coarse-grained part of speech,
- XPOS: a fine-grained legacy tag (see Appendix A),
- FEATS: a `|`-separated list of morphosyntactic features in the `Feature=Value` format.

The XPOS value, as well as the LEMMA, are mostly taken directly from the manually annotated input data (see Section 8.4 below). One exception to this rule concerns multi-token words such as *w czasie* 'during' mentioned in the previous section. Each token within such a multi-token word is assigned the morphosyntactic tag of the whole word, in this case, `prep:gen`, saying that the whole word is a preposition combining with a genitive complement.[3] On the other hand, the 'lemma' of each token in such a complex word is the token itself, i.e., *w* and *sprawie* in this case. This may seem a little inconsistent, but this representation seems more reasonable than the more consistent alternatives; unfortunately, the current UD guidelines do not make clear recommendations about the morphosyntactic treatment of such `fixed` expressions.

The values of UPOS and FEATS are explained in the ensuing subsections, organised by coarse parts of speech, with morphosyntactic features introduced where they first become relevant. Note that of the 17 universal parts of speech defined in UD, `SYM` and `X` are not used in $\text{UD}^{\text{PL}}_{\text{LFG}}$.

[3] As mentioned in Section 6.1, the legacy tagset is well documented elsewhere, including: `http://nkjp.pl/poliqarp/help/en.html`, but it is also summarised in Appendix A.

8.2.1 Verbs (VERB and AUX)

There are two UPOS tags for verbs: VERB and AUX, with some (de)verbal forms tagged as ADJ (adjectival participles) or NOUN (gerunds). The following tokens are assigned the AUX UPOS:

- those forms of BYĆ 'be' and BYWAĆ 'be (habitual)' which are used as past or future tense auxiliaries,
- those forms of BYĆ, BYWAĆ, ZOSTAĆ 'become' and ZOSTAWAĆ 'become (habitual)' which are used as passive auxiliaries,
- the "mobile inflections" *(e)m* '1SG', *(e)ś* '2SG', *śmy* '1PL' and *ście* '2PL' – they are lemmatised to BYĆ,
- those forms of BYĆ and BYWAĆ which are used as copulas,
- the form *to* used as a copula (hence, there are altogether three copular lemmata in Polish),
- the conditional particle *by*,
- the imperative particle *niech* (and its variant *niechaj*).

All verbal forms, including those of auxiliaries (but apart from "mobile inflections" and the two mood particles), gerunds and all participles, have the VerbForm feature with the following values:

- Vnoun – in the case of gerunds, i.e., (de)verbal forms tagged as NOUN,
- Part – in the case of adjectival participles (both active and passive), i.e., (de)verbal forms tagged as ADJ,
- Conv – in the case of adverbial participles (they are tagged as VERB or – in principle – AUX (in the case of adverbial participial forms of copulas), although no adverbial participial auxiliaries actually occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$),
- Inf – in the case of infinitival forms (they are tagged as VERB or AUX),
- Fin – all other verbal forms, including not only the prototypical finite forms, but also:
  - morphologically imperative forms,
  - morphologically impersonal forms (ending in *-no/-to*) – such forms bear the Person=0 feature,
  - forms of the two morphosyntactically unique verb-like lexemes WINIEN and POWINIEN 'ought to',
  - and words which analytically inflect for tense and may act as the main predicate in an utterance, sometimes called 'quasi-verbs', e.g., TRZEBA 'one must', BRAK 'there is no' or TO used as a copula – such words bear the VerbType=Quasi feature.

Apart from 'quasi-verbs', all (de)verbal forms have the Aspect feature, with the following values:

- Imp – imperfective aspect,
- Perf – perfective aspect.

As is common in Slavic linguistics, aspect is treated here as a lexical (not inflectional) feature of verbs.

All finite verb forms, as well as the two mood particles, also bear the Mood feature, which may have one of the following values:

- `Cnd` – conditional mood, only marked on the conditional particle *by*,
- `Imp` – imperative mood, only marked on morphologically imperative forms of verbs and the imperative particle *niech* (and its variant *niechaj*),
- `Ind` – indicative mood, marked on all other finite verbal forms.

Verbs in the indicative mood have the `Tense` feature, with one of the following values:

- `Past` – preterite forms (sometimes called 'l-participles'), as well as impersonal *-no/-to* forms,
- `Pres` – present forms of imperfective verbs, as well as 'quasi-verbs',
- `Fut` – future forms of perfective verbs, as well as future forms of the word BYĆ 'be'.

Note that `Tense` is a morphosyntactic feature of particular tokens, not a semantic feature of the whole utterance. In particular, in sequences such as *będę spał* 'I'll be sleeping', lit. 'will.1SG sleep.SG.M', the tense of the whole utterance is unequivocally future, as reflected by `Tense=Fut` on the future auxiliary *będę*, but the preterite form *spał* used in this construction is still marked as `Tense=Past`.

Apart from verbs in the indicative mood, also adverbial participles bear the `Tense` feature, although its interpretation is different. Its value is `Past` in the case of anterior adverbial participles, e.g., *zrobiwszy* 'having done', and `Pres` in the case of the contemporary adverbial participles, e.g., *robiąc* 'doing'. No other forms bear the `Tense` feature.

Most verbal forms also bear the `Voice` feature, which may have one of two values: `Pass` (only passive adjectival participles) and `Act` (all other verbal forms exhibiting the category of voice).

Many verbal forms also have the `Person` feature, with the following possible values:

- `0` – in the case of morphologically impersonal verbal forms (ending in *-no/-to*), e.g., *kupiono* 'one bought', *nabyto* 'one acquired',
- `1` – in the case of finite and imperative first person forms, as well as "mobile inflections" *(e)m* '1SG' and *śmy* '1PL',
- `2` – in the case of finite and imperative second person forms, as well as "mobile inflections" *(e)ś* '2SG' and *ście* '2PL',
- `3` – in the case of finite third person forms.

The `Person` feature is also present on some forms of pronouns and determiners – see Section 8.2.3 below.

Many verbal forms also carry the `Number` feature, whose values are `Sing` and `Plur`: not only singular and plural finite forms, but also imperative forms, "mobile inflections" and the WINIEN-class forms. Additionally, preterite forms and WINIEN-class also have the `Gender` feature and, hence, possibly also the language-specific `SubGender` feature; their possible values are given in Section 8.2.4 below.

It should also be mentioned that deverbal forms tagged as `NOUN` (gerunds) and `ADJ` (adjectival participles), but not truly verbal forms, also have the `Polarity` feature, whose values are:

- `Neg` – negative polarity,
- `Pos` – positive (affirmative) polarity.

The reason for the presence of Polarity on deverbal – but not fully verbal – tokens is that, according to Polish orthographic rules, the negative marker *nie* is written together with such deverbal forms but separately from truly verbal forms. This orthographic rule is an idiosyncrasy of Polish (in Czech, verbal negation is always attached to the following form), one that does not have confirmation in linguistic facts (arguments for the morphological status of verbal negation in Polish are given in Kupść and Przepiórkowski 2002), so it would also make sense to analyse negated verbs as single tokens. As such "tokens with spaces" are not allowed in UD, only deverbal – gerundial and adjectival participial – forms are marked for the presence or absence of the negation prefix. (In the case of negated truly verbal forms, it is the separate negative marker token, *nie*, that bears the Polarity=Neg feature.)

Apart from the above universally defined features, $\text{UD}^{\text{PL}}_{\text{LFG}}$ also makes use of two language-specific features relevant for the representation of some verbal forms. First, Agglutination distinguishes these rare situations where the preterite has different forms depending on whether the "mobile inflection" auxiliary directly attaches to it or not, e.g., *on mógł* 'he could' (Agglutination=Nagl) vs. *mógł* in *ja mogłem* 'I could' (Agglutination=Agl). There are 26 different verbs for which this distinction is relevant in $\text{UD}^{\text{PL}}_{\text{LFG}}$ (and the Agglutination feature is used 163 times altogether).

Second, the multi-purpose Variant feature is used to distinguish basic from vocalised forms of "mobile inflections", i.e., *m* (Short) from *em* (Long) '1SG' and *ś* (Short) from *eś* (Long) '2SG'. (Variant=Short is also redundantly present on *śmy* '1PL' and *ście* '2PL'.)

8.2.2 Adverbs (ADV)

Adverbs often inflect for degree, so many – but not all – tokens marked as ADV have the Degree feature with the following values:

- Pos – the positive degree,
- Cmp – the comparative degree,
- Sup – the superlative degree.

This feature is also present on typical adjectives (see Section 8.2.5).

Some adverbs, e.g., TUTAJ 'here' and KIEDYŚ 'once (temporal)' would traditionally be classified as pronouns, so they bear the PronType feature – see Section 8.2.3 below for possible values as well as lists of adverbial lemmata of particular pronominal types in $\text{UD}^{\text{PL}}_{\text{LFG}}$. Additionally, two such pronominal adverbs have emphatic variants: GDZIEŻ 'where' (vs. the neutral GDZIE) and JAKŻE 'how' (vs. the neutral JAK). Hence, they are marked as Emphatic=Yes – see, again, Section 8.2.3 for details.

8.2.3 Pronouns (PRON and DET)

Pronouns and determiners (PRON and DET) are two broadly pronominal closed classes of words (21 and 47 different lemmata, respectively). All forms of these words have the PronType feature, which may have the following values:

- `Prs` – personal pronouns, i.e.:
  - `PRON` tokens with lemmata: JA 'I', TY 'you.SG', ON 'he' (all genders and numbers), MY 'we', WY 'you.PL', but also the so-called reflexive pronouns SIEBIE and SIĘ (see the discussion below),
  - and `DET` tokens with lemmata: MÓJ 'my', TWÓJ 'your.SG', NASZ 'our', WASZ 'your.PL', but also the reflexive possessive SWÓJ 'one's',
- `Dem` – demonstrative pronouns, i.e.:
  - `PRON` tokens with lemmata: TO 'this' and TAMTO 'that',
  - `DET` tokens with lemmata: ÓW 'this/that', TEN 'this', TAMTEN 'that', TAKI, TAKIŻ 'such', TYLE 'so many',
  - as well as 13 adverbial demonstrative pronouns: DLATEGO 'for this reason, therefore', DOTĄD 'until now/then', ODTĄD 'from now/then', STAMTĄD 'from there', STĄD 'from here', TAK 'so', TAM 'there', TAMTĘDY 'through there', TU, TUTAJ 'here', WTEDY, WÓWCZAS, WTENCZAS 'then',
- `Ind` – indefinite pronouns, i.e.:
  - `PRON` tokens with lemmata: COŚ 'something', KTOŚ 'somebody', COKOLWIEK 'whatever', KTOKOLWIEK 'whoever',
  - `DET` tokens with lemmata: CZYJŚ 'somebody's', DUŻO 'much, many', JAKIKOLWIEK 'whatever like', JAKIŚ 'some', KILKA 'several', KILKANAŚCIE 'dozen or so', KILKADZIESIĄT 'several tens', KILKASET 'several hundred', KTÓRYŚ 'one of which', MAŁO 'little, few', MNIEJ 'fewer, less', MNÓSTWO 'great quantity', NAJWIĘCEJ 'most', NIECO 'some', NIEJAKI 'certain', NIEJEDEN 'not one', NIEKTÓRY 'some', NIEMAŁO 'not little, not few', NIEWIELE 'not many', PARĘ 'a few', PEWIEN 'certain', SPORO 'considerably many, much', TROCHĘ 'some', WIELE 'many', WIĘCEJ 'more',
  - as well as 7 adverbial indefinite pronouns: GDZIENIEGDZIE 'in some places', GDZIEŚ 'somewhere', JAKOŚ 'in some way', KIEDYKOLWIEK 'whenever', KIEDYŚ 'sometime', NIEKIEDY 'sometimes', SKĄDŚ 'from somewhere',
- `Neg` – negative pronouns, i.e.:
  - `PRON` tokens with lemmata: NIKT 'nobody', NIC 'nothing',
  - `DET` tokens with the lemma ŻADEN 'none',
  - as well as two adverbial pronouns: NIGDY 'never', NIGDZIE 'nowhere',
- `Tot` – collective pronouns, i.e.:
  - `PRON` tokens with lemmata: WSZYSCY 'all (human)', WSZYSTKO 'all (non-human)',
  - `DET` tokens with lemmata: KAŻDY 'each', WSZELKI, WSZYSTEK 'each, all',
  - as well as two adverbial pronouns: WSZĘDZIE 'everywhere', ZAWSZE 'always',
- `Int` – interrogative pronouns, i.e.:
  - `PRON` tokens with lemmata: CÓŻ 'what (emphatic)', KTÓŻ 'who (emphatic)', as well as appropriate occurrences of CO 'what' and KTO 'who',
  - `DET` tokens with lemmata: CZYJ 'whose', ILE 'how many', ILEŻ 'how many (non-human)', ILUŻ 'how-many (human)', JAKIŻ 'what kind', and also appropriate occurrences of JAKI 'what kind' and KTÓRY 'which',
  - as well as ten adverbial pronouns: CZEMU, DLACZEGO 'why', DOKĄD 'where to', GDZIEŻ 'where', JAK, JAKŻE 'how', ODKĄD 'since when', SKĄD 'where from', and appropriate occurrences of GDZIE 'where' and KIEDY 'when',
- `Rel` – relative pronouns, i.e.:
  - appropriate occurrences of `PRON` tokens with lemmata: CO 'what', KTO 'who',

- appropriate occurrences of DET tokens with lemmata: JAKI 'what kind', KTÓRY 'which',
- appropriate occurrences of the adverbial pronouns GDZIE 'where' and KIEDY 'when',
- as well as tokens of the form *co* used to introduce a certain kind (so-called 'resumptive') of relative clauses; the part of speech of such tokens is SCONJ (i.e., subordinate conjunction).

Note that Polish part of speech classifications do not normally envisage the existence of determiners; Polish words corresponding to, say, English determiners are usually classified as adjectives or numerals. However, given the strong emphasis in UD on cross-lingual consistency, Slavic UD treebanks usually make use of the DET part of speech. Similarly, the list of determiners presented above has been constructed with the intention to maximise cross-lingual consistency, at the cost of going against the received wisdom in Polish (morpho)syntax.

A much more controversial aspect of the above annotation principles is the classification of all occurrences of the so-called reflexive pronouns, SIEBIE and SIĘ, as personal pronouns. This is done in the interest of consistency with other Slavic UD treebanks (as recommended by Dan Zeman, p.c.), but this is linguistically wrong in some instances of SIEBIE and in almost all instances of SIĘ. In the case of the lexeme SIEBIE, which overtly only inflects for case (but has no nominative form), there are verbs inherently combining with a form of this lexeme, i.e., without any pronominal or reflexive role played by SIEBIE. One example would be the verb PODPIĆ SOBIE 'drink too much'. There are also constructions involving SIEBIE, such as *rzeka płynie sobie doliną* 'the river flows along the valley', lit. 'river.NOM flows SIEBIE.DAT valley.INS', where the role of SIEBIE is difficult to grasp (Danielewiczowa 2015), but it certainly does not act as a reflexive personal pronoun in this case. The situation is even more clear in the case of SIĘ: out of 3256 occurrences in $\text{UD}^{\text{PL}}_{\text{LFG}}$, 3045 are cases of inherent SIĘ, i.e., part of a verbal lemma, with no pronominal or anaphoric meaning, 146 form an impersonal construction, so they are not pronominal either, and only the remaining 65 could perhaps be classified as pronominal, although even here it could be argued that SIĘ is not really a pronoun but a morpheme reducing the argument structure of the verb.[4] Finally, also not all occurrences of the possessive DET SWÓJ should be classified as PronType=Prs, as sometimes it occurs in multi-word constructions, without its original meaning, as in *chłopcy zrobili swoje* 'the boys did what they should', lit. 'boys.NOM did self's.ACC'.

Nevertheless, since all occurrences of all forms of SIEBIE, SIĘ and SWÓJ are marked – as either PRON (SIEBIE, SIĘ) or DET (SWÓJ) – with the PronType=Prs feature, they are also all marked with the Reflex=Yes feature (no other tokens bear the Reflex feature). In the case of SIĘ these are the only two features it bears. In the case of SIEBIE, there is also the Case feature (see Section 8.2.4 below).

Apart from the single-value Reflex=Yes feature, another broadly pronominal single-value feature is Poss=Yes, marking possessive determiners: not only SWÓJ, but also MÓJ 'my', CZYJ 'whose', etc. (However, genitive forms of the third person pronoun are not marked as possessive.) Forms of four such words also bear the Number[psor] feature, indicating the number of the possessor:

- Sing in the case of MÓJ 'my' and TWÓJ 'your.SG',
- Plur in the case of NASZ 'our' and WASZ 'your.PL'.

[4] See the references in fn. 4 on p. 101.

Yet another such single-value feature is the language-specific feature `Emphatic=Yes`, which marks broadly pronominal forms with the emphatic particle *ż(e)* (treated as an integral part of the word): this concerns `PRON` tokens with lemmata: CÓŻ 'what' (vs. the neutral CO) and KTÓŻ 'who' (vs. KTO), `DET` tokens with lemmata: ILEŻ 'how many (non-human)', ILUŻ 'how many (human)' (both contrasted with ILE, which inflects for gender), JAKIŻ 'what kind' (vs. JAKI), TAKIŻ 'such' (vs. TAKI), the adverbs GDZIEŻ 'where' (vs. GDZIE) and JAKŻE 'how' (vs. JAK), as well as the question `PART`icle CZYŻ (vs. CZY).

There are two multi-purpose features which are used, *inter alia*, to make certain distinctions within the class of truly personal (non-reflexive) `pronouns`. First, the language-specific `Variant` feature distinguishes long (accentable) from short (not accentable) forms of such pronouns, e.g., *jego* and *niego*, with `Variant=Long`, from *go* and *ń*, with `Variant=Short` (all four forms may be interpreted as singular, 3rd person, masculine, accusative). Second, the universal `PrepCase` feature marks some forms, such as *go* and *jego*, as not being able to act as dependents of prepositions (`PrepCase=Npr`), and other – such as *ń* and *niego* – as acting solely as dependents of prepositions (`PrepCase=Pre`).

Another feature important for personal pronouns (but also occurring on some verbs, see Section 8.2.1 above) is `Person`; here, the possible values are:

- `1` – in the case of pronouns JA 'I' and MY 'we' and determiners MÓJ 'my' and NASZ 'our',
- `2` – in the case of pronouns TY 'you.SG' and WY 'you.PL' and determiners TWÓJ 'your.SG' and WASZ 'your.PL',
- `3` – in the case of the multiple forms of the pronoun ON 'he' (inflecting for gender, among other grammatical categories).

Some of the `DET` tokens are morphosyntactically numerals – mostly indefinite (`PronType=Ind`), e.g., DUŻO 'much, many', KILKA 'several', etc. (17 different lemmata altogether), but also interrogative (`PronType=Int`; ILE, ILEŻ and ILUŻ 'how many') and demonstrative (`PronType=Dem`; TYLE 'so many'). Such determiners have the numeral feature `NumType=Card` (see Section 8.2.6 below).

Finally, many `PRON` and `DET` pronouns inflect for case, number and/or gender, so they will have the nominal features `Case`, `Number` and/or `Gender` (hence, in some cases also the language-specific feature `SubGender`); see Section 8.2.4 below.

8.2.4 Nouns (`NOUN` and `PROPN`)

Common nouns (`NOUN`) and proper nouns (`PROPN`) inflect for case and – usually – number, and have lexically specified gender. In Polish, there are seven values of the `Case` feature:

- `Nom` – nominative, the usual case of nominal subjects, but also of some nominals within prepositional phrases, etc.,
- `Acc` – accusative, a frequent case of direct objects, but note that not all direct objects are in the accusative, and not all accusative nominal phrases are direct objects; they may also be temporal dependents, elements of prepositional phrases, etc.;
- `Gen` – genitive,

- `Dat` – dative,
- `Ins` – instrumental,
- `Loc` – locative, occurs only within prepositional phrases,
- `Voc` – vocative.

Apart from common and proper nouns, `Case` is also a feature of all numerals (see Section 8.2.6), almost all adjectival forms (see Section 8.2.5), and many broadly pronominal forms. In some UD treebanks `Case` is also a feature of adpositions, even though, in this case, it is not a morphological feature, but a purely syntactic (valency) feature. For this reason, in $\text{UD}^{\text{PL}}_{\text{LFG}}$, `Case` understood as a feature of adpositions is present in the MISC field of the CoNLL-U format, not in the FEATS field, which represents morphological (or morphosyntactic) features.

Another feature shared by all nominal forms is `Number`, with the expected values:

- `Sing` – singular,
- `Plur` – plural.

Apart from common and proper nouns, `Number` is also a feature of all numerals and determiners, almost all adjectives, a great majority of verbal tokens (`VERB` and `AUX`), and many pronominal tokens.

All proper and common nouns also share the lexical `Gender` feature. Apart from nouns, `Gender` is also borne by all numerals and determiners, almost all adjectival tokens, most verbal tokens (including a great majority of auxiliaries) and most pronominal tokens. After Mańczak 1956, five genders are standardly assumed in Polish linguistics (and in Polish tagsets): three masculine, one feminine and one neuter. The three masculine genders are often called 'human masculine', 'animate masculine' and 'inanimate masculine', but the correlation with the semantic animacy feature is far from perfect. In particular, there are many 'animate masculine' semantically inanimate nouns (including all masculine names of dances, and many more), as well as 'animate masculine' nouns which are semantically human and feminine (some derogatory forms for women, e.g., *babsztyl*), or which are human and, well, no longer animate (*trup* 'corpse'), or which are 'superhuman' (e.g., *diabeł* 'devil' and *anioł* 'angel', but not *bóg* 'god', which is 'human masculine'). For the sake of cross-linguistic consistency, three values are assumed for the 'Gender' feature, i.e.:

- `Masc` – one of the three masculine genders,
- `Fem` – feminine,
- `Neut` – neuter,

but there must be another feature which distinguishes the three masculine genders.

In $\text{UD}^{\text{PL}}_{\text{LFG}}$, a language-specific feature, `SubGender`, is used to this end. It has the following possible values:

- `Masc1` – 'human masculine', usually marked in Polish tagsets as `m1`,
- `Masc2` – '(non-human) animate masculine', usually marked in Polish tagsets as `m2`,
- `Masc3` – 'inanimate masculine', usually marked in Polish tagsets as `m3`.

The `SubGender` feature occurs if and only if `Gender=Masc` is present among the features.

A feature which only occurs on some NOUN tokens is the universal feature Polite with the single language-specific value Depr. It is used to mark the rare 'derogatory' plural forms of some 'human masculine' nouns, e.g., *bliźniaki* 'twins' (Polite=Depr) vs. *bliźniacy* 'twins' (neutral). Such 'derogatory' forms behave morphosyntactically as 'animate masculine' nouns, so they are marked as SubGender=Masc2, even though they are systematically related to SubGender=Masc1 nouns. In $\text{UD}^{\text{PL}}_{\text{LFG}}$ only nominative and a single vocative Depr forms occur, although in theory they could also occur in some very specific accusative positions (Makowska and Saloni 2009). Polite=Depr is a very rare feature, it only occurs 17 times in $\text{UD}^{\text{PL}}_{\text{LFG}}$ (on forms of ten different lexemes).

Recall from Section 8.2.1 that also deverbal nouns, i.e., gerunds (their lemmata end in *-nie/-cie* in Polish), are assigned the coarse part of speech NOUN. Such tokens differ from run-of-the-mill nouns in having the VerbForm=Vnoun, and also bearing the features Aspect (Imp or Perf) and Polarity (Neg or Pos).

8.2.5 Adjectives (ADJ)

Adjectives inflect for case, number, gender and often degree, so typical adjectival tokens will have the features Case, Number, Gender (and SubGender for masculine forms; see Section 8.2.4 above) and Degree, the last one with the same values as in the case of adverbs (see Section 8.2.2). In the case of adjectives which do not synthetically inflect for degree, their Degree value is Pos.

There are four classes of tokens marked as ADJ which have a different repertoire of features. First, there are some (de)adjectival forms which only occur as dependents of prepositions in certain constructions, e.g., *niemiecku* in *po niemiecku* 'in German' or *daleka* in *z daleka* 'from far away'. Such tokens bear only one feature, PrepCase=Pre (compare the use of PrepCase with pronouns, Section 8.2.3). In $\text{UD}^{\text{PL}}_{\text{LFG}}$, there are 42 tokens marked this way (corresponding to 22 different lemmata).

Second, some (de)adjectival forms are only used in ad-adjectival positions within adjectival constructions with a hyphen, e.g., *czarno* in *czarno-biały* 'black-and-white' or *polsko* in *wyznanie polsko-katolickie*, lit. 'denomination Polish-catholic'. Such forms have the single feature Hyph=Yes in the FEATS field (and no other tokens are marked with the Hyph feature). There are 18 tokens (representing 15 lemmata) marked this way in $\text{UD}^{\text{PL}}_{\text{LFG}}$.

Third, some Polish adjectives have a short form, used in predicative constructions, e.g., *ciekaw* 'curious' (predicative only) vs. *ciekawy* 'curious' (either predicative or attributive). As in other Slavic languages, such short forms are marked with the multipurpose Variant=Short feature, and this is their only morphosyntactic feature. In fact, only one token – *ciekaw* – is marked in $\text{UD}^{\text{PL}}_{\text{LFG}}$ this way.

Finally, as mentioned in Section 8.2.1, also adjectival participles are marked as ADJ. Just as typical adjectives, they inflect for Case, Number and Gender, so they may also have the SubGender feature, but they do not bear the Degree feature. Also, they are marked as Vform=Part and, like gerunds, they bear the Aspect and Polarity features. Additionally, passive participles are marked as Voice=Pass, and active participles – as Voice=Act.

Note that some morphosyntactically adjectival forms are assigned the DET coarse part of speech, as discussed in Section 8.2.3.

8.2.6 Numerals (NUM)

In Polish, numeral forms inflect for case and gender, so they bear the features Case, Gender, and possibly SubGender, but they have a lexically specified number, whose value is always plural: Number=Plur. All numerals are also specified for NumType, with the following values:

- Card (i.e., cardinal) – most numerals,
- Frac (i.e., fractional) – numerals with the lemma PÓŁ 'half' (theoretically also ĆWIERĆ 'quarter', etc., but PÓŁ 'half' is the only fractional numeral in $\text{UD}^{\text{PL}}_{\text{LFG}}$).

Note that, apart from cardinal and fractional numerals, various forms which are traditionally treated as numeral are not assigned the NUM coarse part of speech; these include:

- ordinal numerals, e.g., DRUGI 'second' – they are morphosyntactic adjectives, so they are tagged as ADJ,
- words such as TRZYKROTNIE 'three times' – they are morphosyntactic adverbs, so they are tagged as ADV,
- words such as DWÓJKA 'two' – they are morphosyntactic nouns, so they are tagged as NOUN, etc.

Note also that some morphosyntactically numeral forms are assigned the DET coarse part of speech (with the feature NumType=Card), as discussed in Section 8.2.3.

All morphosyntactic numerals, also those marked as DET, also have the language-specific DepType feature. This is a syntactic feature, it resides in the MISC field. Its possible values are:

- Rec – this numeral token expects the noun it combines with to be in the genitive case,
- Congr – this numeral token agrees in case with the noun it combines with.

8.2.7 Prepositions (ADP)

All adpositions have the feature AdpType, with the following values:

- Post – postposition, only TEMU 'ago' in $\text{UD}^{\text{PL}}_{\text{LFG}}$,
- Prep – preposition, all other adpositions.

Some of the prepositions ending in a consonant also have a form with the additional vowel *e* at the end, e.g., *z* vs. *ze* 'from, with'. The two forms are distinguished via the multipurpose language-specific feature Variant, with the usual values: Short (e.g., *z*) and Long (e.g., *ze*).

Unlike in some other UD treebanks, there is no Case feature in the FEATS field of adposition, as it is not a morphological, but a syntactic (valency) feature. For this reason, this feature is present in the MISC field and takes six values in $\text{UD}^{\text{PL}}_{\text{LFG}}$ (all Polish cases apart from the vocative).

8.2.8 Coordinate and subordinate conjunctions (CCONJ and SCONJ)

Coordinate conjunctions are marked as CCONJ, subordinate conjunctions – or complementisers – as SCONJ. They form closed classes, but perhaps not as small as might be expected: there are 25 different forms of coordinate conjunctions (3282 tokens in $\text{UD}^{\text{PL}}_{\text{LFG}}$), and 28 different forms of subordinate conjunctions (1509 tokens). Note that preconjunctions are not distinguished from proper conjunctions at the morphosyntactic level – this distinction is made at the syntactic level, in dependency relation labels. In general, there are no features relevant to conjunctions, with the only exception – already mentioned in Section 8.2.3 – concerning SCONJ tokens of the form *co*, when they introduce a certain kind of relative clauses (which may contain a resumptive pronoun): there are six occurrences of such tokens in $\text{UD}^{\text{PL}}_{\text{LFG}}$, and they are marked as PronType=Rel.

8.2.9 Other parts of speech (PART, INTJ and PUNCT)

There are 78 different particles (5888 tokens) in $\text{UD}^{\text{PL}}_{\text{LFG}}$. The most frequent is the negative marker NIE, distinguished by the presence of the Polarity=Neg feature. There are also four question particles, CZY, CZYŻ, CZYŻBY and the dated AZALIŻ, distinguished by the presence of the PartType=Int universal feature with a language-specific value. (The question particle CZYŻ also bears the Emphatic=Yes feature.) Another class consists of adnumeral operators, i.e., non-inflecting words which attach to numerals, sometimes with the effect of making their meaning approximate; they often have the same form as existing prepositions, but they do not govern a specific case, e.g., OKOŁO 'about, around' or Z 'some, around'. Just like the preposition Z, the adnumeral operator Z has two forms: *z* and *ze*, distinguished via the Variant feature (Short for *z* and Long for *ze*). The other adnumeral operators, and all other particles, do not have any morphosyntactic features. Many of the other particles differ from other parts of speech by their ability to attach to diverse syntactic categories: nominal, adjectival, verbal, etc., e.g., TYLKO 'only' or ZWŁASZCZA 'especially'.

There are 42 occurrences of 17 different interjections (INTJ) in $\text{UD}^{\text{PL}}_{\text{LFG}}$. They do not have any morphosyntactic features.

Finally, the PUNCT coarse 'part of speech' is used for all punctuation marks (apart from those which are integral parts of word forms; see Section 8.1 above). Almost all have the PunctType feature. The only exceptions are the ellipsis characters, …, and dots which are not used as sentence-final periods: there seem to be no relevant universal values of PunctType for such punctuation marks (20 tokens altogether, compared to 25,820 punctuation marks in $\text{UD}^{\text{PL}}_{\text{LFG}}$). For other punctuation characters, the values of PunctType are:

- Peri – for the sentence-final full stop (.),
- Excl – for the exclamation mark (*!*),
- Qest (sic!) – for the question mark (*?*),
- Dash – for hyphens (-) and longer dashes (–),
- Comm – for commas (,),
- Semi – for semi-colons (;),
- Quot – for Polish and English-style quotation marks,
- Brck – for parentheses (only round parentheses occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$).

Whenever the value of PunctType is Quot or Brck, another feature is also present, PunctSide, with the following values:

- Ini – an opening parenthesis (() or quotation character (", „ or “),
- Fin – a closing parenthesis ()) or quotation character (" again or ").

8.3 Syntax

8.3.1 Nominal constructions

Let us start by considering the internal structure of broadly nominal (i.e., also numeral and prepositional) phrases, on the basis of example (8.1),[5] whose UD structure is given in Figure 8.1. As in the previous part of this monograph, the basic dependency tree is represented above the sentence, the enhanced dependency – below the sentence, and any differences between them are marked in red.

(8.1) Aresztowanego na 48 godzin mistrza zwolniono po
arrested.ACC.SG.M for 48.ACC hours.GEN master.ACC.SG.M released.IMPS after
poręczeniu majątkowym.
guarantee.LOC property.ADJ.LOC
'The master who was detained for 48 hours was released on bail bond.'

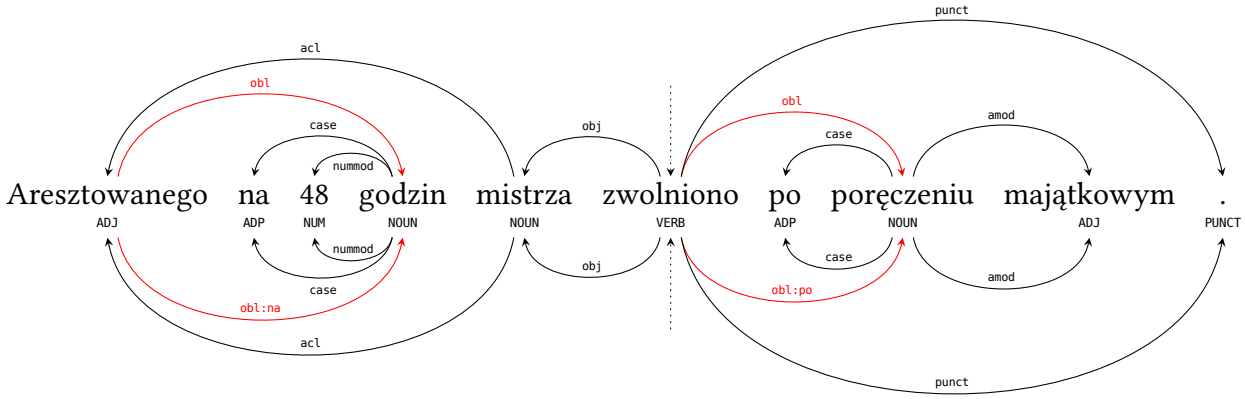


Figure 8.1: UD representation of (8.1)

In compliance with UD guidelines, semantic rather than syntactic or morphosyntactic criteria decide about the headedness of broadly understood nominal phrases.[6] Thus, in Figure 8.1, the prepositional phrases *na 48 godziny* 'for 48 hours' and *po poręczeniu majątkowym* 'after bail bond' are not headed by the preposition, but by the noun (*godziny* 'hours' and *poręczeniu* 'guarantee', respectively). Similarly, while there are good arguments for numerals as heads of

[5] Concerning morphosyntactic information in glosses, see footnote 1 on page 3.

[6] On headedness criteria, see Zwicky 1985, Hudson 1987 and Croft 1996, as well as papers in Corbett et al. 1993.

numeral phrases (see, e.g., Saloni and Świdziński 1985, Przepiórkowski 1999 and references therein), in UD they are dependents, as illustrated with *48 godzin* '48 hours'. As also shown in this figure, numeral dependents of nouns bear the `nummod` label, and adpositional dependents of nouns are marked as `case`.

Another relation label of dependents of nouns is `amod`, borne by typical adjectives; see *poręczeniu majątkowym* 'bail bond', lit. 'guarantee.NOUN property.ADJ' in Figure 8.1. However, this relation does not apply to adjectival participles, which – as recommended by Joakim Nivre (p.c.) – are treated here as reduced relative clauses, i.e., they are marked with the `acl` relation appropriate for clausal dependents of nouns. In the above example, the noun *mistrza* 'master' is modified by such a passive participial phrase, *aresztowanego na 48 godzin* 'arrested for 48 hours'. In the case of full relative clauses modifying nouns, the relation is subtyped to `acl:relcl`, as in Figure 8.2 for sentence (8.2).

(8.2) Żyje dzięki komuś, kto rozumiał to hasło.
lives.3SG thanks somebody.DAT.SG.M who.NOM.SG.M understood.3SG.M this.ACC motto.ACC
'(S)he lives thanks to somebody who understood this motto.'

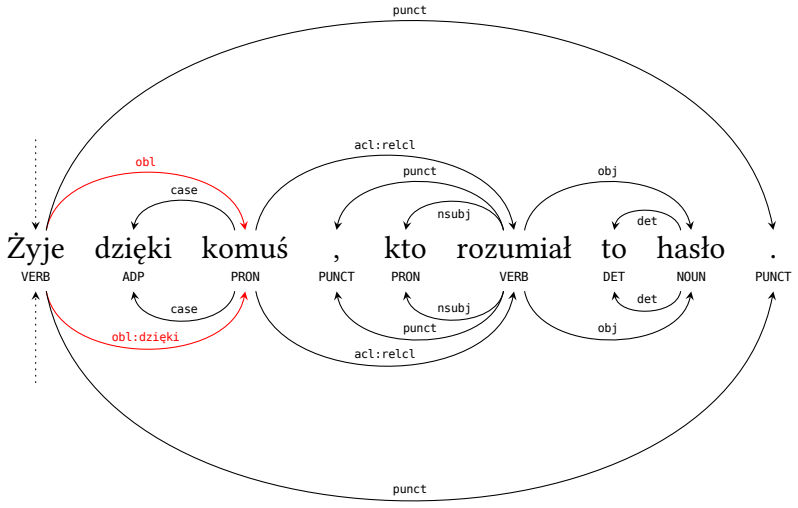


Figure 8.2: UD representation of (8.2)

The same figure also illustrates another relation to dependents of nouns, namely, the `det` relation to determiners.

There are also a few possible relations between nominals, the most typical being `nmod`, as in Figure 8.3 for sentence (8.3).

(8.3) To nasza ostatnia polemika z radnymi PO.
COP our.NOM.SG.F last.NOM.SG.F debate.NOM.SG.F with councillors.INS PO.GEN
'This is our last debate with PO councillors.'

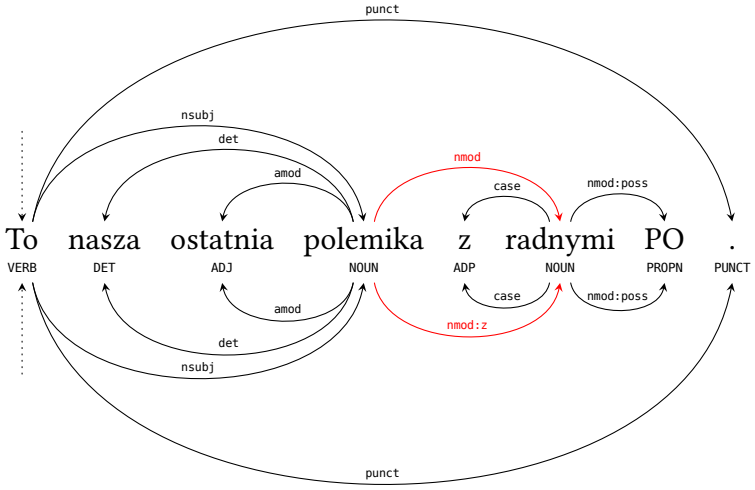


Figure 8.3: UD representation of (8.3)

Since the prepositional phrase *z radnymi PO* 'with Civic Platform councillors' (*PO* = *Platforma Obywatelska* 'Civic Platform') is headed by the noun *radnymi* 'councillors', the relation between *polemika* 'polemic, public debate' and this prepositional phrase is represented as a dependency between two nominals, hence the `nmod` relation, subtyped in enhanced dependencies with the lemma of the preposition. Such subtypes in the enhanced representation may also be complex prepositions, consisting of a number of tokens related with the `fixed` dependency (cf. the introduction to Section 8.2). In the case of such "words with spaces", the spaces are replaced with underscore characters, as illustrated in Figure 8.4, corresponding to sentence (8.4) – see the enhanced dependency `nmod:na_temat` there.

(8.4) Powodem były jego publiczne wypowiedzi na temat ludzi upośledzonych
reason.INS were his public.NOM statements.NOM on topic people.GEN disabled.GEN
fizycznie.
physically
'The reason was his public statements about physically disabled people.'

Apart from such language-specific subtypes of `nmod` in the enhanced representation, there is also one universal subtype of the this relation, `nmod:poss`, which is used in $\text{UD}^{\text{PL}}_{\text{LFG}}$ – both in basic trees and in enhanced graphs – to represent all kinds of relations expressed by the genitive case, not just the narrow possession relation. Hence, in Figure 8.3, the dependency between *radnymi* 'councillors' and the genitive *PO* 'Civic Platform' is labelled `nmod:poss`, and similarly for the dependency between *wypowiedzi* 'statements' and *jego* 'his' in Figure 8.4. Note, however, that not all possessive relations are labelled this way; the relation between a noun and its determiner is always labelled as `det`, even in the case of possessive determiners, as in Figure 8.3: see the relation between *nasza* 'our' and *polemika* 'polemic, public debate' there. The crucial difference between *nasza polemika* 'our debate' in this figure and *jego wypowiedzi* 'his statements' in Figure 8.4 is that *nasza* 'our' is tagged as `DET` and, hence, the dependency is `det`, while *jego* 'his' is tagged as `PRON`, so the dependency is the same as to any other possessive nominal element, i.e., `nmod:poss`.

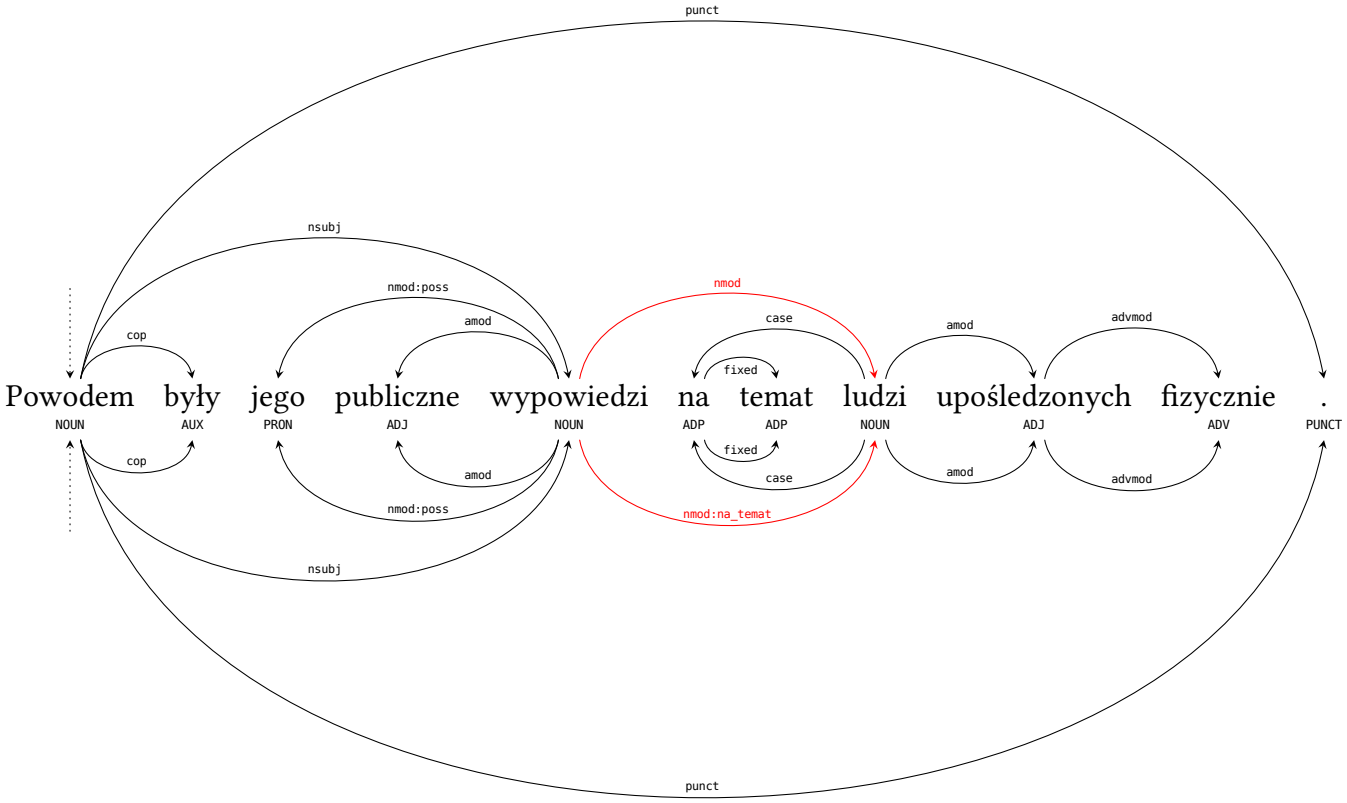


Figure 8.4: UD representation of (8.4)

There are two UD relations for appositions. The `appos` relation is used for 'flexible' appositions, where the order of the apposed constituents is relatively free. On the other hand, the `flat` relation is used for 'rigid' appositions, where the order is fixed, as is the case for names preceded by titles, etc. Both relations are illustrated with Figure 8.5 for sentence (8.5) (repeated from the previous part).

(8.5) Zdecydował o tym Lech Kaczyński, prezydent
decided.3SG.M about this Lech.NOM.SG.M Kaczyński.NOM.SG.M president.NOM.SG.M
RP.
RP.GEN
'It was decided by Lech Kaczyński, the president of the Republic of Poland.'

As illustrated in Figure 8.6 for sentence (8.6), longer apposition chains are – in compliance with UD guidelines – not represented as chains; instead, all non-initial elements of the apposition are dependents of the first element.

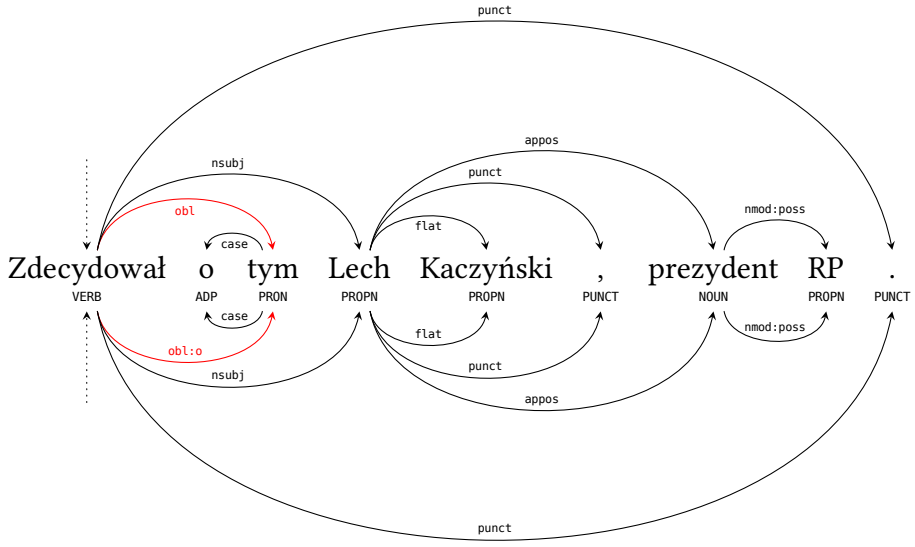


Figure 8.5: UD representation of (8.5)

(8.6) Po nim zabierze głos pan poseł Maciej
after him take.FUT.3SG voice.ACC mister.NOM.SG.M deputy.NOM.SG.M Maciej.NOM.SG.M
Manicki.
Manicki.NOM.SG.M
'After him, the floor will be given to Mr. Maciej Manicki, MP.'

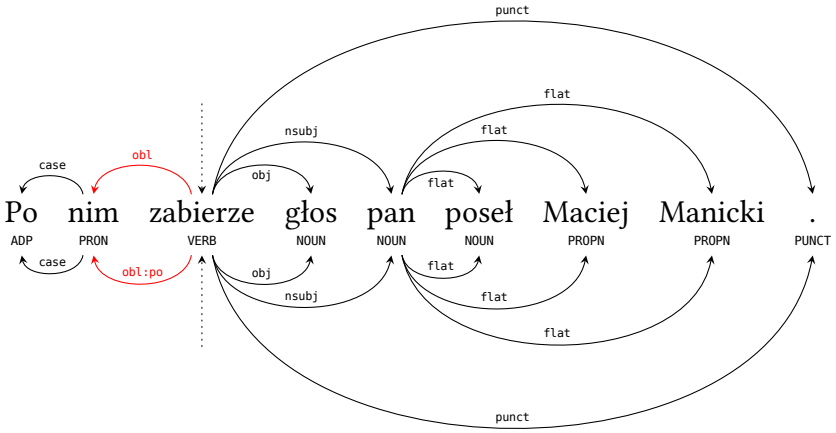


Figure 8.6: UD representation of (8.6)

There are two more UD relations which are immediately relevant for the nominal domain but are not used in $\text{UD}^{\text{PL}}_{\text{LFG}}$, namely, `clf`, useful for languages – such as Chinese – with highly grammaticalised classifier systems, and `compound`, used to represent nominal compounds (e.g., *phone book*) and particle verbs (e.g., *put up*) in languages such as English, as well as serial verbs in languages that display this phenomenon.

8.3.2 Verbal constructions

An important feature of UD is that it attempts to make no distinction between arguments and adjuncts (also called modifiers). We find this feature very attractive, as the reality of this dichotomy is highly questionable (Przepiórkowski 2016a, 2016b, 2017a, 2017b). So, for example, apart from subjects, direct objects and indirect objects, all broadly nominal (i.e., also prepositional and numeral) dependents of verbs are treated as oblique, without distinguishing them further into arguments and adjuncts.[7] In the three subsections below, we first describe nominal dependents of verbs, then we move to other kinds of dependents, and finally cover constructions involving auxiliaries and copulas.

Nominal dependents

In compliance with UD, we assume three kinds of nominal core arguments.

Subjects Nominal subjects are marked as `nsubj`, as in Figures 8.2–8.6 above. Prototypical subjects are bare nominative noun phrases, but note that not all nominative phrases are subjects – they may also be parts of some prepositional constructions, i.e., be marked as obliques (see below). In fact, since nominative nouns may be heads of copular constructions (see Section 8.3.2 below), they may have all kinds of incoming relations. Subjects in passive constructions, i.e., those corresponding to direct objects (see below) in the active voice, are marked as `nsubj:pass`.

The strongest test for subjecthood in Polish is agreement with finite verbs. A class of broadly nominal subjects that does not pass this test consists of typical numeral phrases, where the numeral requires its nominal companion to be in the genitive case.[8] Such typical numerals bear the accusative case in the subject position (Franks 1995; Przepiórkowski 1999, 2004a). Hence, as there is no nominative element within such subject numeral phrases, the verb occurs in the default 3rd person singular neuter form – in Polish, as in other Indo-European languages, verbs only agree with nominative subjects, and occur in the default form if there is no subject, the subject has no case feature at all (e.g., it is a clause) or it has case value different than nominative. This is exemplified with sentence (8.7) and its UD representation in Figure 8.7.

(8.7) Zginęło wtedy dwóch pilotów.
died.3SG.N then two.ACC.PL pilots.GEN.PL
'Two pilots were killed then.'

How do we know, then, that such numeral phrases are subjects? They satisfy all other subjecthood criteria in Polish, including control into adverbial participles, binding of anaphoric pronouns and the possibility to be coordinated with uncontroversial nominative subjects (Dziwirek 1994; Przepiórkowski 1999).

[7] But see Zeman 2017 for an attempt to re-introduce this dichotomy in UD, as well as Chapter 8.5 for a discussion of relevant differences between $\text{UD}^{\text{PL}}_{\text{LFG}}$ and $\text{UD}^{\text{PL}}_{\text{SZ}}$.

[8] Such typical numerals bear the `DepType=Rec` feature in the MISC field in the CoNLL-U representation, while other numerals, agreeing in case with the noun, are marked as `DepType=Congr` – see Section 8.2.6.

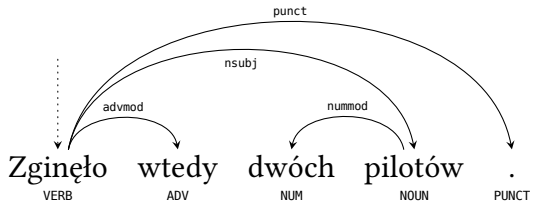


Figure 8.7: UD representation of (8.7)

Note that, theoretically, also prepositional phrases may be subjects in Polish (Jaworska 1986a, 1986b), although in $\text{UD}^{\text{PL}}_{\text{LFG}}$ there seem to be no sentences illustrating this possibility.

Direct objects In $\text{UD}^{\text{PL}}_{\text{LFG}}$, as in Polish linguistics in general (e.g., Gołąb et al. 1968: 132, Urbańczyk 1992: 62), direct object (Pol. *dopełnienie bliższe*) is understood as that dependent of the verb which becomes the subject under passivisation. Nominal direct objects are marked using the `obj` dependency label, as in Figures 8.1–8.2 and 8.6 above. These three figures illustrate the typical situation where the direct object is in the accusative case, but it may also occur in the genitive, when a higher verb is negated (see, e.g., Przepiórkowski 2000 and references therein), and – additionally – some verbs require their direct objects to occur in the instrumental or the genitive case.[9]

Conversely, not all accusative dependents of verbs are marked as direct objects, and that for two reasons. First, some verbs which combine with accusative complements do not passivise at all. Second, some accusative dependents of verbs are not complements, but adjuncts, e.g., durative modifiers.

Again, also prepositional phrases may in principle – but not actually in $\text{UD}^{\text{PL}}_{\text{LFG}}$ – be direct objects in Polish (Jaworska 1986b, 1986a).

Indirect objects There is no notion corresponding to indirect object which would be widely accepted in Polish linguistics. In the attempt to increase cross-lingual consistency, $\text{UD}^{\text{PL}}_{\text{LFG}}$ defines indirect objects, `iobj`, as any dative arguments, as in Figure 8.8.

(8.8) Asystent już podawał chirurgowi fartuch.
assistant.NOM.SG.M already gave.3SG.M surgeon.DAT gown.ACC
'The assistant was already giving the surgeon his gown.'

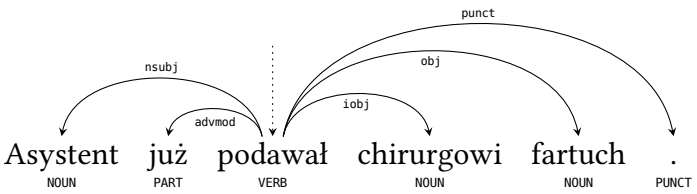


Figure 8.8: UD representation of (8.8)

[9] Perhaps also in the dative case (Zabrocki 1981: 124–125), although not in $\text{UD}^{\text{PL}}_{\text{LFG}}$.

This decision is somewhat controversial, as it re-introduces the argument–adjunct dichotomy: dative arguments are treated as indirect objects, while other – not subcategorised – dative dependents are treated as obliques (see below). For this reason, future releases of $\text{UD}^{\text{PL}}_{\text{LFG}}$ may redefine `iobj` or get rid of this relation altogether.

Oblique dependents Any other broadly understood nominal dependents of verbs are marked as obliques, as illustrated in Figures 8.1 (two oblique dependents, including a dependent of the passive participle), 8.2, and 8.5–8.6. In all these examples, the oblique dependents are actually prepositional phrases, so the enhanced dependency label is subtyped with the lemma of the preposition: `obl:na`, `obl:po` (in two of these examples), `obl:dzięki`, `obl:o`, etc. Analogically to the case of `nmod` subtypes (see Section 8.3.1), also complex (multi-token) prepositions may be used for subtyping `obl` (again, with the underscore in place of a space).

Also bare nominal dependents of verbs which do not happen to be subjects or (direct or indirect) objects are classified as `obl` (without any additional subtype in the enhanced representation), as in Figure 8.9. Apart from a prepositional oblique dependent, *na klatkę* 'to the hallway', there are two bare nominal obliques there (both would be traditionally classified as adjuncts): *czasami* 'sometimes', lit. 'times.INS', and *mi* 'me.DAT', which in this case is not (treated as) subcategorised by the verb, so it is not an indirect object.

(8.9)

Również	czasami	sąsiadka	coś	mi	wyniesie	na	klatkę.
also	times.INS	neighbour.NOM.SG.F	something.ACC	me.DAT	take_out.3SG	on	hallway

'Also, sometimes the neighbour will bring something for me to the hallway.'

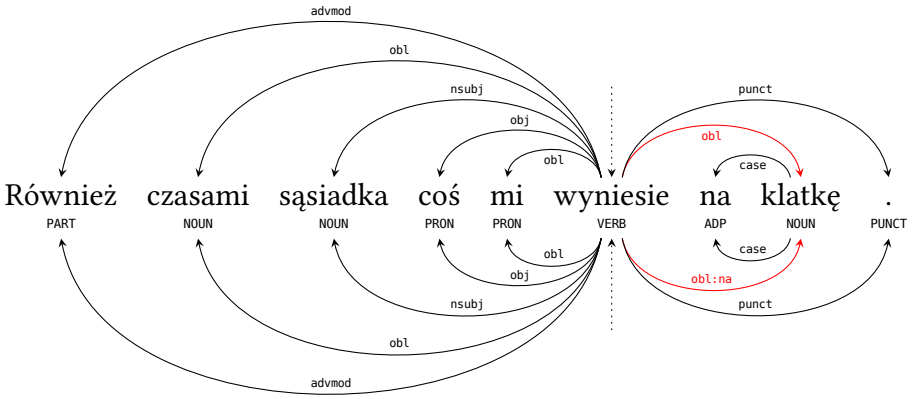


Figure 8.9: UD representation of (8.9)

Vocative dependents One special group of non-core nominal dependents is treated separately from other obliques, namely, vocative dependents, as in Figure 8.10.

(8.10) - Zamknij się, Kostek.
shut_up.IMP.2SG RM Kostek.NOM
'Shut up, Kostek.'

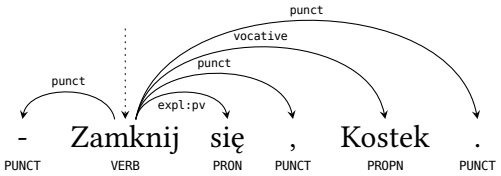


Figure 8.10: UD representation of (8.10)

All phrases in morphologically vocative case are treated as `vocative` dependents, but so are some nominative phrases used vocatively – this is the case in Figure 8.10, where the `vocative` dependent *Kostek* is in the nominative case.

There is one more UD relation which may be used for nominal dependents of verbs, but is not used in $\text{UD}^{\text{PL}}_{\text{LFG}}$, namely, `dislocated`, used for example for topics which are not direct dependents of verbs, such as *John* in *John, I really like him*. It is not clear whether constructions which would warrant the use of `dislocated` occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$.

Verbal dependents

There are four kinds of dependents of verbs which are mainly verbal or clausal.

Non-nominal subjects First, non-nominal subjects are marked as `csubj`. The UD standard also envisages `csubj:pass`, for non-nominal subjects in passive constructions, but such constructions do not occur in $\text{UD}^{\text{PL}}_{\text{LFG}}$. In Polish, there are two basic types of non-nominal subjects: clausal and infinitival. The first situation is illustrated by Figure 8.11 for sentence (8.11). There, the subject of a form of the verb ZDAWAĆ SIĘ 'seem' is a subordinate clause introduced by the complementiser ŻE.

(8.11) Zdaje się, że uczył w szkołach.
seems.3SG RM COMP taught.3SG.M in schools
'It seems that he taught at (various) schools.'

The other situation is illustrated in Figure 8.12. Here, a form of the verb UDAĆ SIĘ 'manage' takes two dependents: an indirect object, *mordercy* 'murderer', and an infinitival phrase, *zbiec* 'escape' (the additional enhanced `nsubj` edge will be discussed below).

(8.12) Mordercy udało się zbiec.
murderer.DAT.SG.M managed.3SG.N RM escape.INF
'The murderer managed to escape.'

In both cases, the `csubj` dependency is between two verbal forms. However, due to the standard UD analysis of copular constructions (see below), other parts of speech may occur on either side of the `csubj` dependency. (This remark also applies to other verbal dependencies discussed in this subsection.)

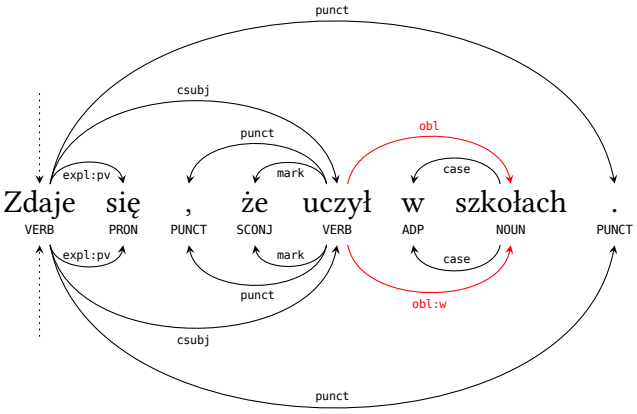


Figure 8.11: UD representation of (8.11)

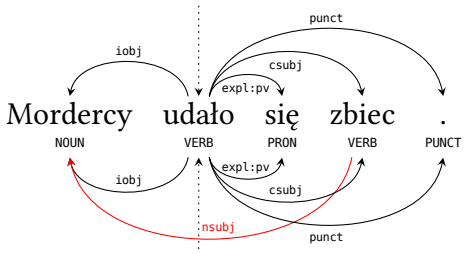


Figure 8.12: UD representation of (8.12)

Clausal (closed) arguments Second, non-subject clausal core dependents are marked as `ccomp`. The exact status of this relation is not clear at the time of writing (February 2018): the UD guidelines state that `ccomp` is a "clausal complement of a verb or adjective", i.e., it "is a dependent clause which is a core argument", adding vaguely that "it functions like an object of the verb" (`http://universaldependencies.org/u/dep/ccomp.html`). This would imply that any non-subject subcategorised clause is a `ccomp`. However, the program validating UD treebanks notices situations where a single verb has both an `obj` and a `ccomp` dependent and reports that each verb should have at most one object (see also `http://universaldependencies.org/svalidation.html`). For this reason, all subcategorised non-subject clauses are marked as `ccomp` in $\text{UD}^{\text{PL}}_{\text{LFG}}$, but those that are direct objects in the sense presented above (i.e., those that become subjects under passivisation) are subtyped to the language-specific relation `ccomp:obj`. The two kinds of `ccomp` dependents are illustrated in Figures 8.13–8.14 presenting dependency structures of sentences (8.13)–(8.14).

(8.13) Potem zapytała, skąd dzwoni.
later asked.3SG.F whence calls.3SG
'After that she asked where (s)he is calling from.'

(8.14) Przecież powiedziałem, że cię lubię.
but said.1SG.M COMP you.ACC like.1SG
'But I said I like you.'

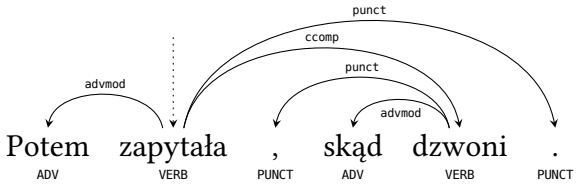


Figure 8.13: UD representation of (8.13)

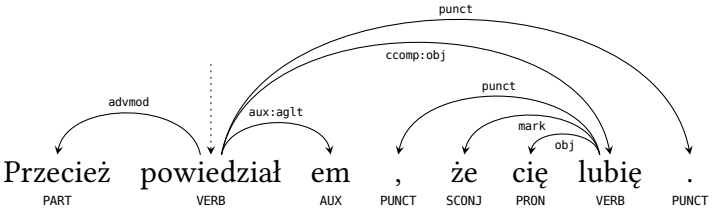


Figure 8.14: UD representation of (8.14)

Note that this understanding of `ccomp`, consistent with the main UD guidelines but inconsistent with the validating script, re-introduces the argument–adjunct dichotomy: only argument clauses are marked as `ccomp`, while – as discussed below – adjunct clauses are marked as `advcl`.

Infinitival (and other open) arguments Third, controlled infinitival phrases are marked as `xcomp`, as in Figure 8.15. As this example shows, control is understood broadly and it also includes raising (see, e.g., Landau 2013 and references therein).

(8.15) Historia zaczęła biec szybciej.
history.NOM.SG.F started.3SG.F run.INF faster
'History started to run faster.'

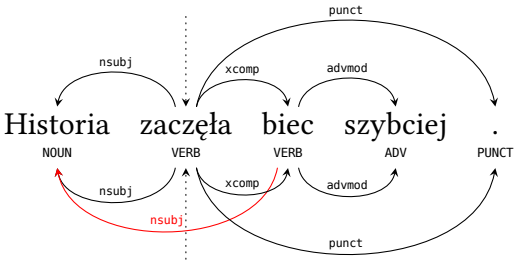


Figure 8.15: UD representation of (8.15)

Note that the enhanced structure contains now one more dependency, indicating the subject of the controlled verb: *historia* 'history' is not only the surface subject of the finite verb *zaczęła* 'started', but also an understood subject of the infinitival verb *biec* 'run'.

Again, just as in the case of some `ccomp` dependents, some `xcomp` dependents are direct objects in the sense defined above (referring to passivisation), so they are marked in $\text{UD}^{\text{PL}}_{\text{LFG}}$ with the language-specific `xcomp:obj` relation, as in Figure 8.16.

(8.16) Jednocześnie polecił zająć się milicjantem prowadzącym
simultaneously ordered.3SG.M handle.INF RM policeman.INS.SG.M leading.INS.SG.M
śledztwo.
investigation.ACC.SG.N
'At the same time, he ordered to take care of the policeman leading the investigation.'

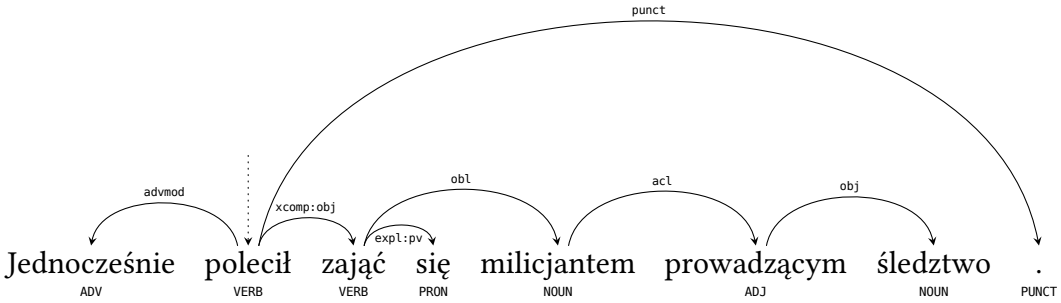


Figure 8.16: UD representation of (8.16)

(An additional enhanced edge is missing here only because the dependent of *polecił* 'ordered' which is understood as the subject of the infinitival *zająć się* 'handle, take care of', is not overtly realised in this sentence.)

In UD it is assumed that only non-subject dependents may bear the `xcomp` relation. This leads to somewhat inconsistent annotation, as there are also cases of subjects behaving like `xcomp` dependents in the sense that they are infinitival phrases whose own subjects are obligatorily understood as one of the dependents of the main verb.[10] In fact, Figure 8.12 above (page 190) illustrates exactly this phenomenon: the subject of the finite verb *udało się* 'managed' is an infinitival phrase, *zbiec* 'escape', whose subject must be co-referent with *mordercy* 'murderer', the indirect object of the finite verb. For this reason, the enhanced representation in Figure 8.12 includes an additional `nsubj` edge, just as Figure 8.15, involving `xcomp`, does.

Apart from infinitival dependents, `xcomp` also marks predicative complements (apart from copular constructions, see below), as in Figure 8.17 for example (8.17).

(8.17) Prezesem został Krzysztof Piotrowski.
chairman.INS.SG.M became.3SG.M Krzysztof.NOM.SG.M Piotrowski.NOM.SG.M
'Krzysztof Piotrowski became the chairman.'

Predicative complements, such as *prezesem* 'chairman' in this example, are assumed to be controlled in a similar way to infinitival complements, hence the additional enhanced `nsubj` relation also in this case.

Modifier clauses Finally, the fourth – modifying – kind of verbal dependents of verbs is marked as `advcl`. Figure 8.18 illustrates a typical use of this label: the main verb, *przeżył* 'survived', is modified by a subordinate clause, *bo udał, że jest martwy* 'because he pretended that he was dead'.

[10] Outside Polish, control into subjects is discussed, e.g., in Arka and Simpson 1998 (for Balinese).

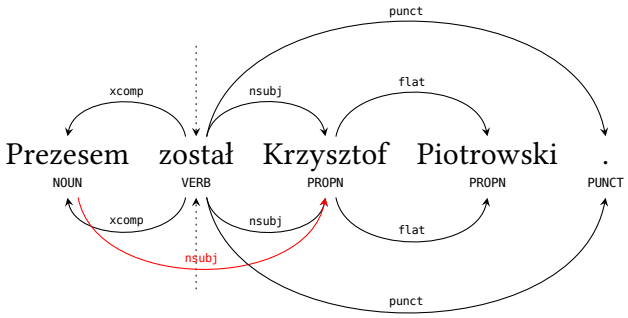


Figure 8.17: UD representation of (8.17)

(8.18) Przeżył, bo udał, że jest martwy.
survived.3SG.M because pretended.3SG.M COMP is.3SG dead.NOM.SG.M
'He survived because he pretended that he was dead.'

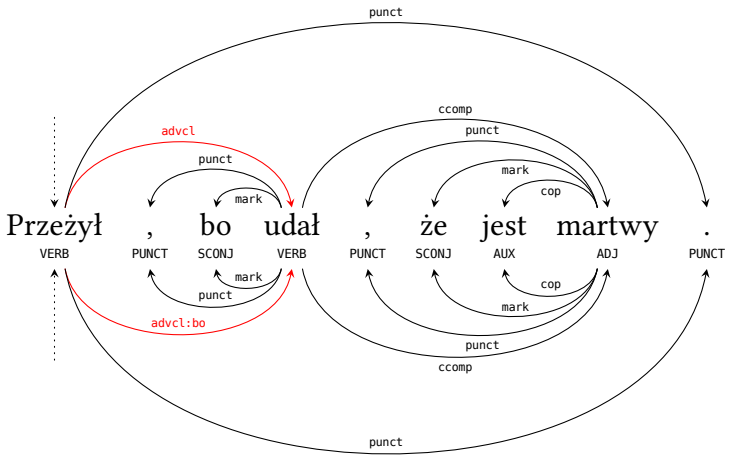


Figure 8.18: UD representation of (8.18)

Note that, just as in the case of `nmod` and `obl`, the `advcl` dependency is subtyped in the enhanced representation, in this case with the (lemma of the) subordinating conjunction attached to the label.

On the most prominent reading of example (8.18), all verbs are understood as sharing the (implicit) subject. This is not an instance of obligatory control, though: it is possible to imagine a scenario where somebody survived because somebody else pretended to be dead. The same `advcl` label is also used to mark dependencies involving such obligatory control; typical Polish examples involve adverbial participles (sometimes called converbs), as in Figure 8.19, where the subject of *zgadzając się* 'agreeing' must be understood as the subject of the main verb, *popełnił* 'made', i.e., as *Cimoszewicz*.

(8.19) Cimoszewicz popełnił błąd, zgadzając się kandydować.
Cimoszewicz.NOM.SG.M made.3SG.M mistake.ACC agreeing RM run.INF
'Agreeing to run (for presidency), Cimoszewicz made a mistake.'

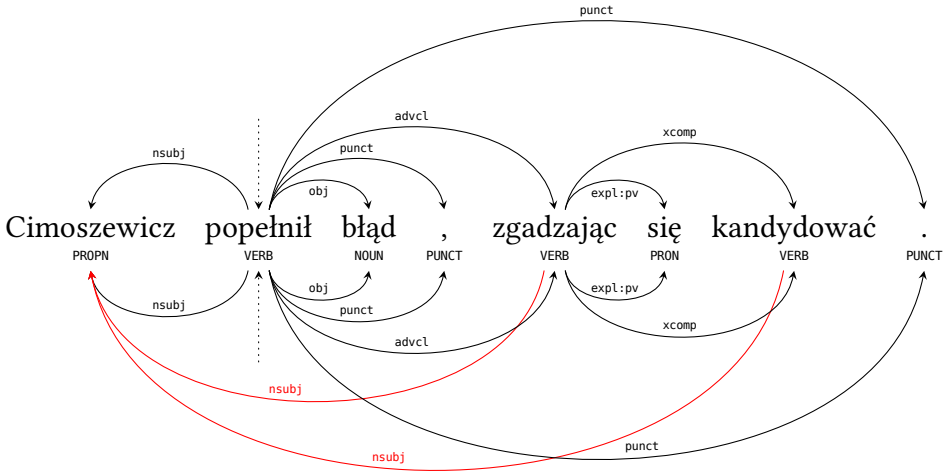


Figure 8.19: UD representation of (8.19)

Such cases of obligatory control to an adjunct are marked in $\text{UD}^{\text{PL}}_{\text{LFG}}$ with an additional enhanced dependency, see the `nsubj` from *zgadzając* to *Cimoszewicz*. Once the subject of *zgadzając* is identified, another enhanced dependency must be added, from *kandydować* 'run (for president)' to *Cimoszewicz*, because of the `xcomp` relation between *zgadzając* and *kandydować*.

It is worth noting that, despite UD's attempt to avoid the argument–adjunct distinction (as explicitly stated on `http://universaldependencies.org/u/overview/syntax.html`), this dichotomy is preserved in the treatment of verbal dependents: controlled dependents are marked as `xcomp` when they are (non-subject) arguments, but as `advcl` when they are adjuncts. Similarly, clausal (non-subject) arguments are marked as `ccomp` (but see above for some remarks on the inconsistency of the current understanding of `ccomp`), but clausal adjuncts – as `advcl`.

Let us also comment on another questionable principle of current UD guidelines. Recall from the discussion of Figure 8.1 above that adjectival participles modifying nouns are treated as reduced relative clauses, i.e., marked as `acl`. However, such participles may also appear in sentences lacking an overt realisation of the noun they refer to. Typically, this occurs in cases of subject pro-drop, as in Figure 8.20.

(8.20) Nagle stanęła przygwożdżona do ziemi.
suddenly stopped.3SG.F nailed.NOM.SG.F to floor
'She suddenly stopped (as if) nailed to the floor.'

Here, the passive participle *przygwożdżona* 'nailed' is a dependent of the verb, *stanęła* 'stood, stopped', only because the subject is dropped. As a dependent of a verb, the participle is marked as `advcl`. Were the subject present, the participle would be its dependent and it would be marked as `acl`.

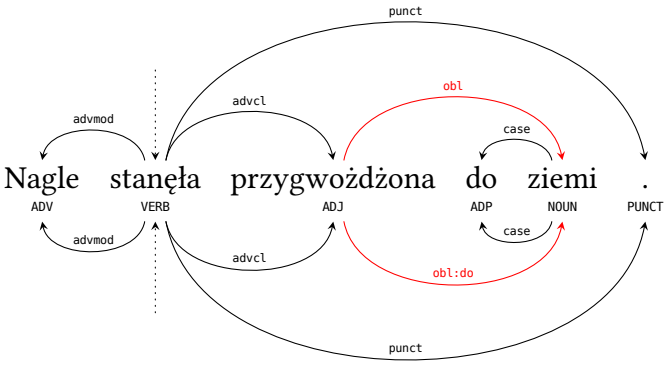


Figure 8.20: UD representation of (8.20)

Other dependents

Adverbial dependents Another major class of dependents of verbs consists of broadly understood adverbial dependents: not only those headed by tokens with coarse part of speech `ADV`, but also `PART` (particles) and `INTJ` (interjections). Such dependents are marked as `advmod`, as shown in Figure 8.21.

(8.21) – Ehm, nie wiem tak naprawdę, Komediancie.
ehm NEG know.1SG so really comedian.VOC.
'Well, Comedian, I don't really know.'

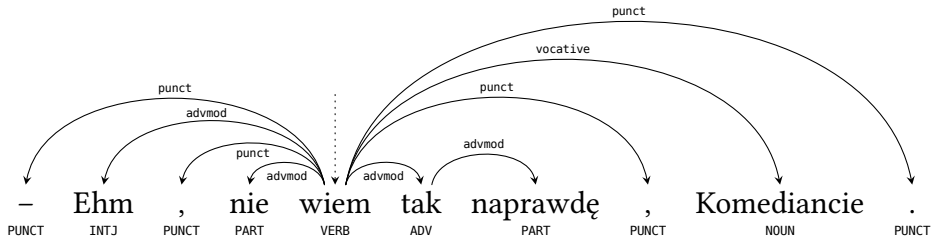


Figure 8.21: UD representation of (8.21)

Four tokens bear this relation there, of which three are dependents of the main verb, *wiem* 'know': the adverb *tak* 'so', the negative particle *nie* and the interjection *ehm* 'ehm, well'. Additionally, the particle *naprawdę* 'really' is analysed here as an `advmod` dependent of the adverb *tak*.

Functional words Two more relations indicate functional dependents of verbs. One is `mark`, used for subordinating conjunctions, as in Figures 8.11, 8.14 and 8.18 above. The last of these shows that `mark` is used both in argument subordinate clauses (i.e., `ccomp` dependents) and in adjunct subordinate clauses (`advcl` dependents).

The other relation is `expl`, which by itself is used in various UD treebanks to mark expletive pronouns. While $\text{UD}^{\text{PL}}_{\text{LFG}}$ does not use the bare `expl` dependency (arguably, there are no

expletive pronouns in Polish), it does use two universal subtypes of this relation employed to mark two different functions of the so-called reflexive pronoun *się*: inherent, where it is an integral part of the lemma (`expl:pv`), and impersonal, where it is used to form an impersonal construction (`expl:impers`). The inherent *się* is very frequent and it occurs above in Figures 8.10–8.12, 8.16 and 8.19. The impersonal *się* is much rarer and it is illustrated in Figure 8.22, which actually contains both subtypes of `expl`.

(8.22) Może nauczysz się wreszcie, jak się takie rzeczy załatwia.
perhaps teach.FUT.2SG RM finally how RM such.ACC things.ACC handle.3SG
'Perhaps you'll finally learn how one takes care of such things.'

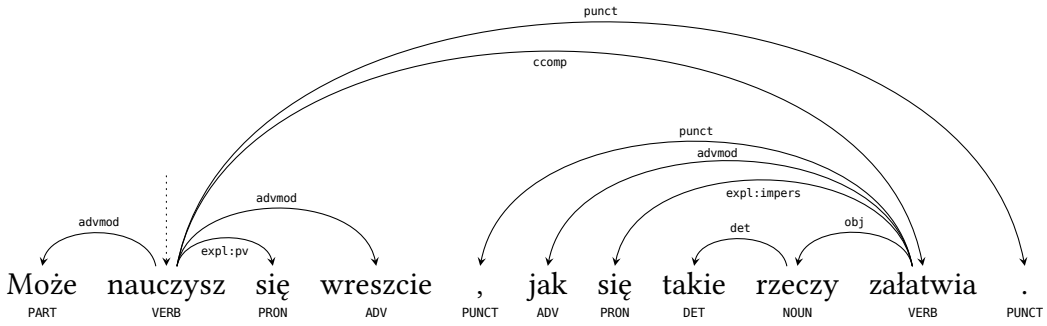


Figure 8.22: UD representation of (8.22)

A statistically insignificant weakness of this representation of different functions of *się* is that it is impossible to represent those rare situations where a single *się* has both functions.[11]

Auxiliaries and copulas

Unlike in many linguistic theories, including to some extent LFG, auxiliaries and copulas are treated in UD as dependents of the lexical verbs rather than as their heads. This is illustrated in Figure 8.23, where the auxiliary *będę* 'I will' is a dependent of the root verb, *krył* 'hide', and the copula *jestem* 'I am' is a dependent of the predicative noun phrase *człowiekiem prawicy* 'rightist', lit. 'man (of the) Right'.

(8.23) Nie będę krył, że jestem człowiekiem prawicy.
NEG will.1SG hide.SG.M COMP am.1SG man.INS Right.GEN
'I won't hide the fact that I am a rightist.'

Note that the `ccomp` dependency between the main verb, *krył*, and the subordinate clause actually targets a noun (rather than a verb). Similarly, also the `csubj` label may be used on a dependency between a verb and a nominal element (rather than between two verbs), as in Figures 8.24–8.25, where the subjects are copular constructions (a finite clause in Figure 8.24 and an infinitival phrase in Figure 8.25) headed by nominal words (a pronoun and a proper noun, respectively), and in Figures 8.26–8.27, where common nouns in copular constructions have verbal subjects (a finite clause in Figure 8.26 and an infinitival phrase in Figure 8.27).

[11] See the discussion of example (7.31) (page 158) in the previous part of this monograph and references therein.

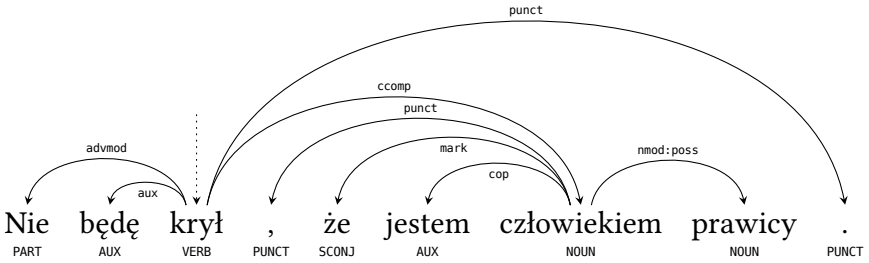


Figure 8.23: UD representation of (8.23)

(8.24) Pewnie ciekawi Cię, kim jest pani Kownacka?
perhaps interests.3SG you.ACC who.INS is.3SG Mrs.NOM Kownacka.NOM
'Perhaps you're curious who Mrs. Kownacka is?'

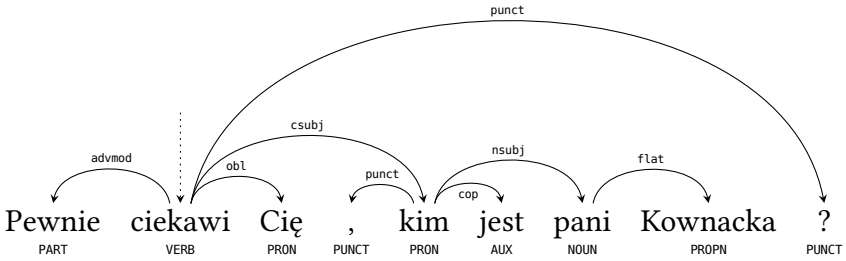


Figure 8.24: UD representation of (8.24)

(8.25) - Nie wystarczyło być na Syberii, aby otrzymać takie odznaczenie.
NEG sufficed.3SG.N be.INF on Siberia to receive.INF such.ACC distinction.ACC
'It was not enough to be in Siberia, to receive such a distinction.'

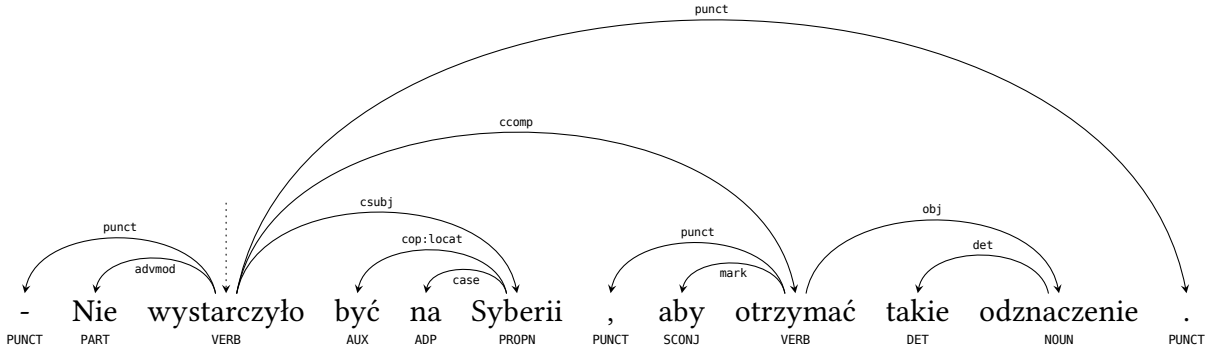


Figure 8.25: UD representation of (8.25)

(8.26) Chociaż faktem jest, że Hubal okropnie przeżył te
although fact.INS is.3SG COMP Hubal.3SG.M terribly experienced.NOM.SG.M these.ACC
zdarzenia.
events.ACC
'Although it's a fact that Hubal took these events terribly emotionally.'

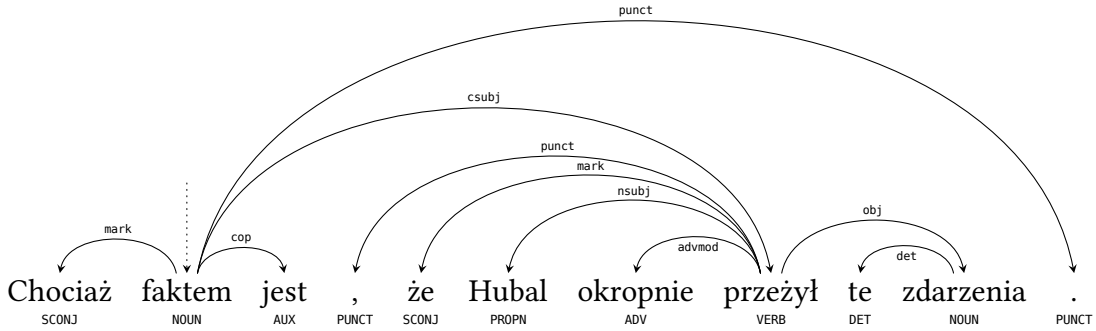


Figure 8.26: UD representation of (8.26)

(8.27) - Wymknęło ci się, że przesadą byłoby zginąć za komunizm.
slipped you.DAT RM COMP exaggeration.INS be.COND die.INF for communism
'– You let it slip that dying for communism would be too much.'

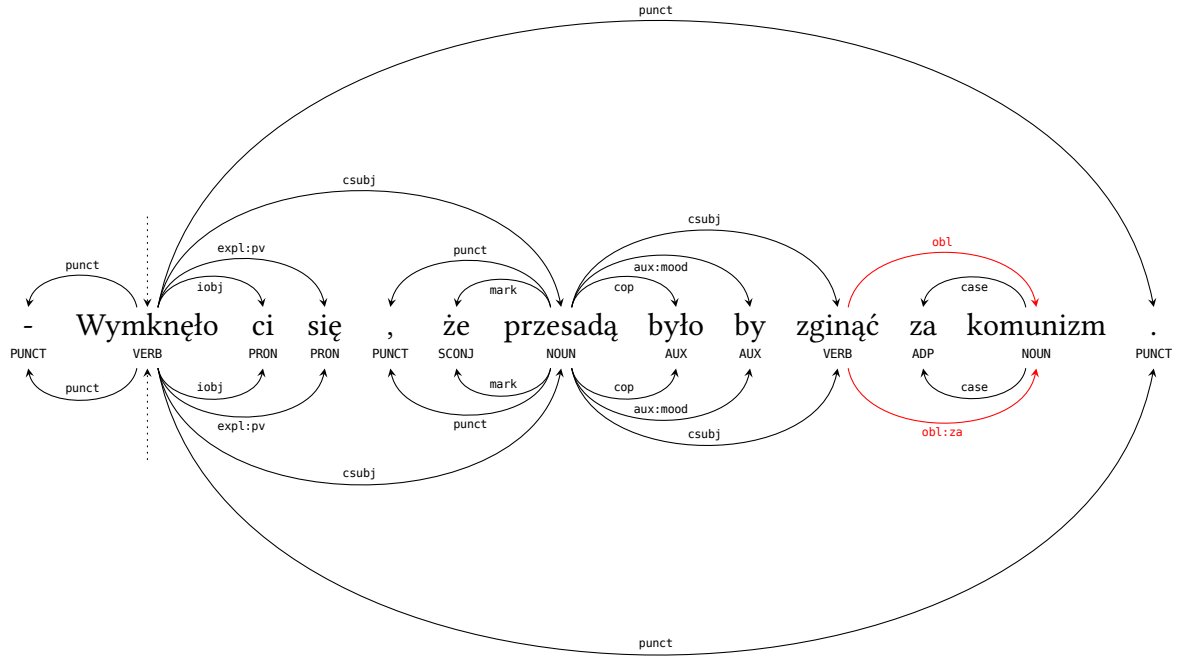


Figure 8.27: UD representation of (8.27)

8.3.3 Dependents of deverbal nouns and adjectives

In brief, deverbal nouns (gerunds, forms with lemmata ending in *-nie/-cie*) are treated as nouns but deverbal adjectives (active and passive participles) – as verbs for the purpose of establishing the labels of outgoing dependency relations. This contrast is illustrated by Figures 8.28–8.29.

(8.28) Na dzień obecny kontynuuję czytanie biografii Nerona.
on day present continue.1SG reading.GER.ACC biography.GEN.SG.F Nero.GEN.SG.M
'Today, I continue reading a biography of Nero.'

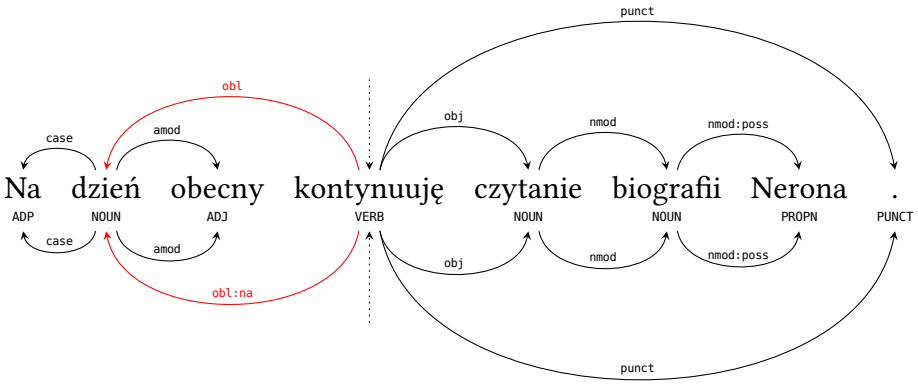


Figure 8.28: UD representation of (8.28)

(8.29) Wyobrażał sobie minę Każe-duba czytającego jego raport.
imagined.3SG.M REFL.DAT face.ACC Każe-dub.GEN reading.ADJ.PTCP.GEN his report.ACC
'He imagined the face of Każe-dub reading his report.'

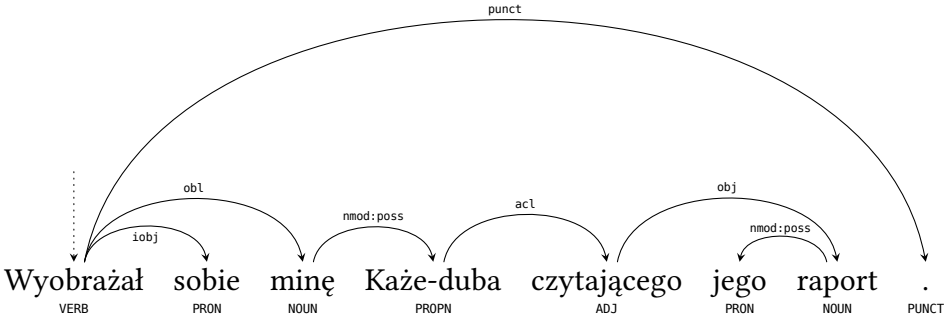


Figure 8.29: UD representation of (8.29)

In the former, the gerundial form of the verb CZYTAĆ 'read', i.e., *czytanie* combines with *biografii Nerona* 'Nero's biography'. This dependent corresponds to the verb's direct object, but is marked here as an `nmod`, because UD guidelines do not allow dependents of `NOUN`s to be objects. No such restrictions hold for `ADJ`ectives, so the analogous dependent of the active participle

czytającego, namely, *jego raport* 'his report', is marked as `obj`. We view this contrast as a clear inconsistency, one that directly follows from the current UD principles (nouns cannot have objects, etc.) and contingent decisions (across Slavic languages, gerund forms are marked as nouns).

8.3.4 Dependents of adjectives and adverbs

Typical dependents of (not deverbal) adjectives and adverbs are marked with the relation `advmod` (already introduced in the presentation of verbal constructions). For example, in Figure 8.4 above (page 184), this relation labels the dependency from the adjective *upośledzonych* 'disabled' to the adverb *fizycznie* 'physically'. Also adverbs and particles which are dependents of adverbs are marked with the `advmod` relation, as in Figure 8.21 (page 195) – see the relation between *tak* 'so' and *naprawdę* 'really'. However, in the case of nominal dependents of adjectives and adverbs, the dependency is `obl`, as in Figure 8.30, where the prepositional phrase *w dużym stopniu* 'to a large extent', headed by the noun *stopniu*, is an `obl` dependent of the adjective *dziedziczne* 'hereditary'.

(8.30) Uważano, że są one w dużym stopniu dziedziczne.
considered.IMPS COMP are they in large extent hereditary
'They were considered to be to a large extent hereditary.'

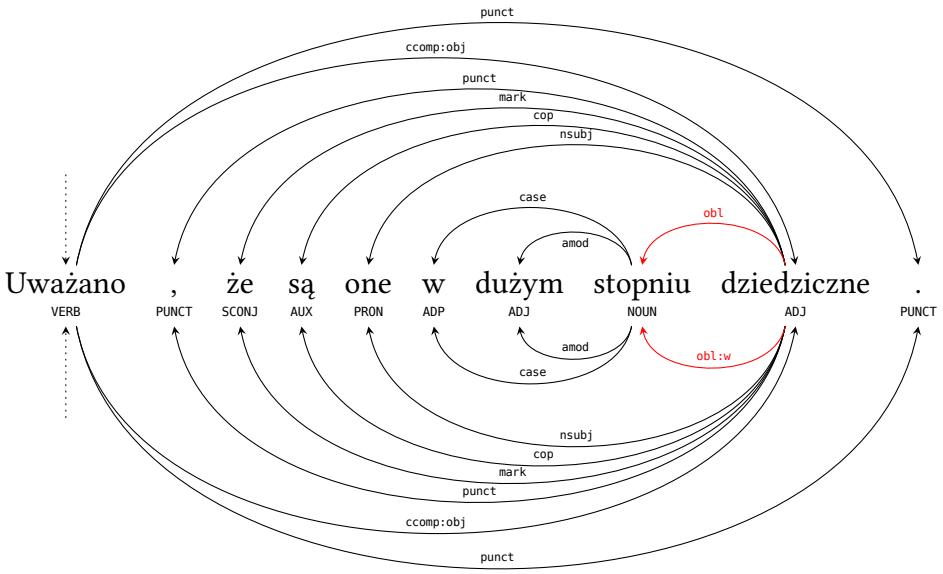


Figure 8.30: UD representation of (8.30)

8.3.5 Coordinate structures

In UD, coordinate structures are headed by the first conjunct, all other conjuncts are its dependents (with the `conj` dependency holding between them), and any overt conjunctions or preconjunctions are dependent on the immediately following conjunct (with the `cc` label in the case of conjunctions, `cc:preconj` in the case of preconjunctions, and `punct` in the case of commas acting as conjunctions). This is illustrated in Figure 8.31, in which two infinitival verbs are coordinated: *złagodzić* 'ease, relieve' and *znieść* 'eliminate'.

(8.31) Także w tym przypadku fototerapia może złagodzić lub znieść
also in this case phototherapy.NOM.SG.F may.3SG relieve.INF or eliminate.INF
całkowicie niekorzystne objawy.
completely unfavourable.ACC.PL.M symptoms.ACC.PL.M
'Also in this case, phototherapy may relieve or completely eliminate unfavourable symptoms.'

As also shown here, dependencies between the whole coordinate structure and other phrases, i.e., between the first conjunct (the head of coordination) and those other phrases, propagate in the enhanced representation to other conjuncts. Thus, since the coordinate structure is an open dependent of the finite verb *może* 'may', as indicated in the base tree by the `xcomp` edge from this finite verb to the first conjunct (*złagodzić*), there is also another `xcomp` edge in the enhanced representation, from this finite verb to the second conjunct (*znieść*). Conversely, since the conjuncts of the coordinate structure share a dependent, namely, the direct object *niekorzystne objawy* 'unfavourable symptoms', which is indicated in the base tree by the `obj` edge from the first conjunct, *złagodzić*, to the head of this direct object, *objawy*, there is also another `obj` edge in the enhanced representation, namely, between the second conjunct, *znieść*, and the head of the direct object, *objawy*. Note that the enhanced representation plays an important role here, i.e., it disambiguates between two readings consistent with the basic tree representation: one where *niekorzystne objawy* is the direct object of both infinitival verbs (as in this example), and another, where it is the direct object of the first conjunct only.

There is another dependent that is shared between the two conjuncts, namely, the subject *fototerapia* 'phototherapy'. Here coordination interacts with control: since the first conjunct, *złagodzić*, is (subject-)controlled by the finite verb, *może*, as indicated by the `xcomp` relation between them, there is an enhanced `nsubj` relation not only from the finite verb (this one is already present in the basic tree), but also from the infinitival verb *złagodzić*. But since this `xcomp` relation propagates to the second conjunct, *znieść*, also this second conjunct bears the `nsubj` relation to the main verb in the enhanced representation.

Adding incoming and outgoing enhanced dependencies to non-initial conjuncts does not always involve simple copying of the dependency label from the first conjunct. In the case of sentence (8.32), repeated from the previous part and involving asyndetic coordination, the finite verb *jest* 'is' plays a dual role: it is the passive auxiliary dependent of the first conjunct, the participial *zapięta pod szyję* 'buttoned up to the neck', but a regular copula dependent of the second conjunct, the simple adjectival *wysmukła jak kwiat* 'lean like a flower'; see Figure 8.32.

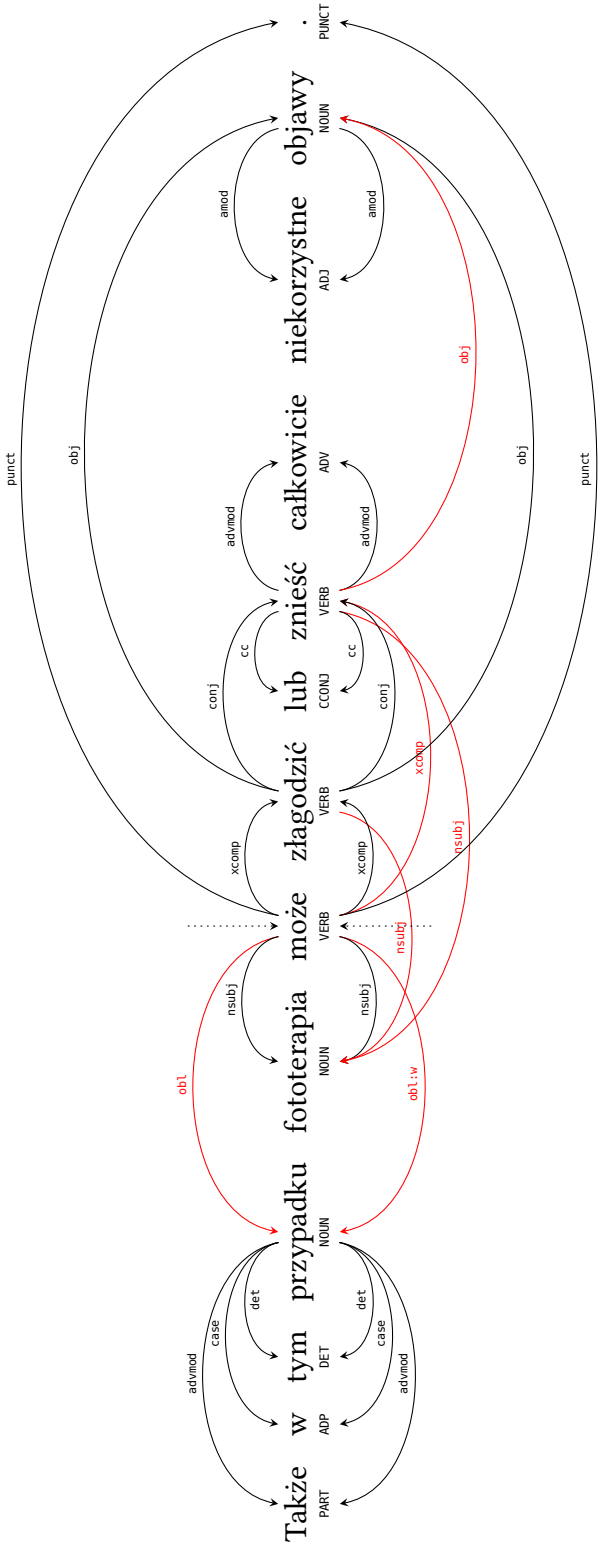


Figure 8.31: UD representation of (8.31)

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

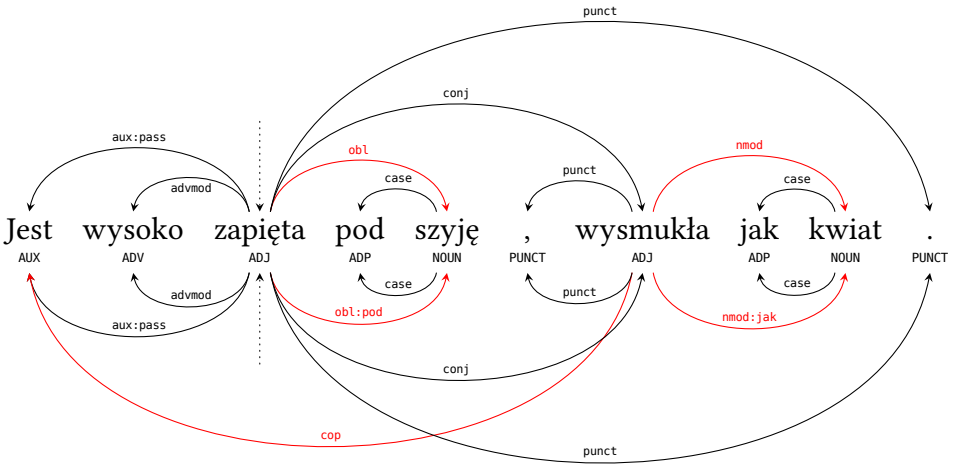


Figure 8.32: UD representation of (8.32)

While in the case just considered the shared dependent is the target of different dependency relations from different conjuncts, it is also possible – and in fact more frequent – for the shared governor to assign different dependency labels to the dependent conjuncts, as illustrated in Figure 8.33, corresponding to example (8.33).

(8.33) Nad każdym, nawet najkrótszym tekstem medytuję.
over each.INS even shortest.INS text.INS meditate.1SG
'I meditate over each – even the shortest – text.'

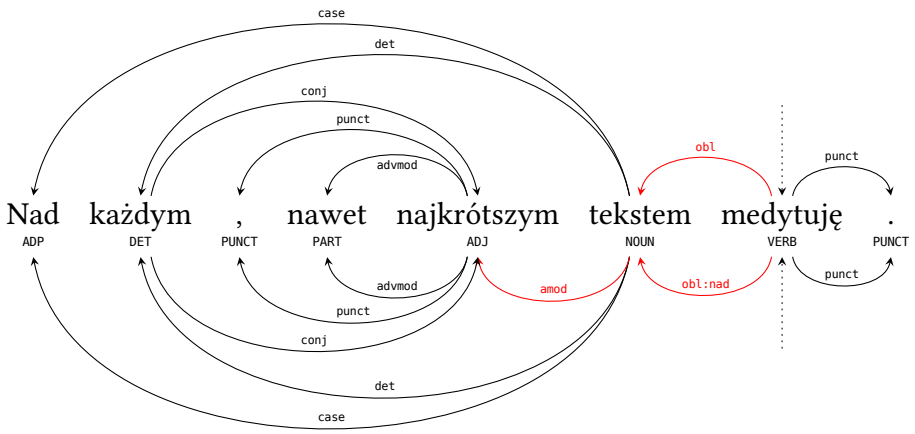


Figure 8.33: UD representation of (8.33)

The two morphosyntactically adjectival forms, *każdym* 'each' and *najkrótszym* 'shortest', are treated here as again asyndetically coordinated[12] and both modifying the noun *tekstem* 'text'. However, *każdym* is assigned the `DET` part of speech, so it bears the `det` dependency relation, and *najkrótszym* is a typical adjective, so the dependency added in the enhanced representation bears the `amod` label.

8.4 Underlying data

Texts in $\text{UD}^{\text{PL}}_{\text{LFG}}$ are ultimately drawn from two corpora: over 84% of utterances come from the *National Corpus of Polish* (`http://nkjp.pl/`; Przepiórkowski et al. 2011, 2012) and almost 16% – from the *Corpus of 1960s Polish* (`http://clip.ipipan.waw.pl/PL196x`; Kurcz et al. 1990; Bień and Woliński 2003; Ogrodniczuk 2003). Both corpora were manually lemmatised and morphosyntactically tagged, and these lemmata and tags are to a large extent preserved in $\text{UD}^{\text{PL}}_{\text{LFG}}$, in the LEMMA and XPOS fields of the CoNLL-U format.[13]

More directly, the sentences in $\text{UD}^{\text{PL}}_{\text{LFG}}$ come from the LFG structure bank described in the first part of this monograph. 19,597 sentences with their LFG syntactic structures form an input to the conversion described in the second part. Many of these are sentences with multiple possible LFG analyses, as well as accidental duplicates. After conversion only unique UD structures are retained, i.e., a sentence may appear in the corpus a couple of times only with different dependency annotations. As a result, the $\text{UD}^{\text{PL}}_{\text{LFG}}$ treebank contains 17,246 dependency representations (with 130,967 segments) for 17,190 different sentences.

These 17,246 trees were split into training, development and test subcorpora in two stages, in compliance with UD guidelines.[14] First, for each sentence, it was checked whether this sentence occurs in the $\text{UD}^{\text{PL}}_{\text{SZ}}$ treebank of Polish. If it occurred in the training corpus there, it was also assigned – with all its dependency structures, if there were more than one – to the training subcorpus of the $\text{UD}^{\text{PL}}_{\text{LFG}}$ treebank. Otherwise, if it was found in the $\text{UD}^{\text{PL}}_{\text{SZ}}$ development corpus, it was assigned to the $\text{UD}^{\text{PL}}_{\text{LFG}}$ development corpus. Otherwise, if it occurred in the $\text{UD}^{\text{PL}}_{\text{SZ}}$ test corpus, it was assigned to the $\text{UD}^{\text{PL}}_{\text{LFG}}$ test corpus. Altogether, 3502 (2594 + 439 + 469, respectively) dependency representations were pre-classified to the three subcorpora this way.

Second, the remaining sentences were randomly added to the development and test subcorpora until each of these subcorpora contained more than 20% of the whole corpus, in terms of both the number of dependency representations and the number of tokens. The rest of the sentences were added to the training corpus. This procedure resulted in the split summarised in Table 8.1.

About 42.1% of sentences represent the fiction genre, 39.1% – news, 7.4% – nonfiction, 7.3% – spoken, 3% – interactive Internet texts (forums, chatrooms, etc.), and there are also traces

[12] This analysis is controversial, but it serves well as an illustration of the general point made here.

[13] In fact, some tags were consistently mapped to new tags, e.g., in the case of numeral subjects of the governing (`DepType=Rec`) type, which had originally been tagged as nominative, but for the purpose of the LFG structure bank were reanalysed as accusative, a position justified, e.g., briefly in Franks 1995 and at length in Przepiórkowski 1999, 2004a.

[14] `http://universaldependencies.org/release_checklist.html#data-split`

Table 8.1: Subcorpora of UD^{PL}_{LFG}

	trees	**tokens**
training	13,744	104,750
development	1745	13,105
test	1727	13,112

of static Internet pages (0.8%), academic style (0.3%) and legal texts (0.1%). For each sentence, genre is explicitly given in a comment to this sentence. In the case of sentences derived from the National Corpus of Polish, this genre information is taken directly from the headers of appropriate texts; in the case of sentences from the Corpus of 1960s Polish, they were derived from two (of five) parts of the corpus, News and Fiction, and were classified accordingly.

8.5 Comparison to UD^{PL}_{SZ}

UD^{PL}_{LFG} is the first Polish UD treebank making use of enhanced dependencies. It is available since February 2018 and it is officially released as part of UD version 2.2. However, there is also another UD treebank of Polish, available since UD release 1.2 in November 2015, namely, UD^{PL}_{SZ}. That treebank is based on the *Składnica zależnościowa* treebank (http://zil.ipipan.waw.pl/Składnica; Wróblewska 2014) version 0.5, which is the result of automatic conversion from a constituency parsebank (Świdziński and Woliński 2010). *Składnica zależnościowa* was first converted – by Dan Zeman and colleagues – to the Prague dependency style and then to Universal Dependencies (HamleDT 3.0, 2015; Zeman et al. 2014).[15] We have not performed a systematic comparison of the two treebanks, but have – in the process of developing UD^{PL}_{LFG} – noticed various differences worth documenting. The rest of this section compares UD^{PL}_{LFG} released in UD v.2.2 (July 2018) with the UD^{PL}_{SZ} released in UD v.2.1 (November 2017).

8.5.1 Tokenisation

There are at least two tokenisation differences between the two treebanks. First, UD^{PL}_{SZ}, but not UD^{PL}_{LFG}, takes advantage of the possibility to represent sequences of tokens written without intervening spaces also as single tokens, as in *Straciłem równowagę*. 'I lost my balance', lit. 'lost.1SG.M balance.ACC.SG.F', where *Straciłem* 'lost.1SG.M' is a sequence of two tokens: *Stracił* 'lost.SG.M' and *em* 'AUX.1SG' (see Section 8.1 on such multi-token units in Polish). In the CoNLL-U representation of this sentence, there are five lines (apart from the comment lines) in UD^{PL}_{SZ}; (8.34) shows the first four columns and the final column (with missing material between them indicated by '…'):

[15] See the description of UD^{PL}_{SZ} at `https://github.com/UniversalDependencies/UD_Polish-SZ/blob/dev/README.md`.

(8.34)
```
1-2 Straciłem _          _      …  _
1   Stracił   stracić     VERB   …  _
2   em        być        AUX    …  _
3   równowagę równowaga   NOUN   …  SpaceAfter=No
4   .         .          PUNCT  …  _
```

On the other hand, the partial representation of the same sentence in $\text{UD}^{\text{PL}}_{\text{LFG}}$ is as in (8.35) – it differs not only in the lack of one line, but also in the more consistent – in our opinion – use of the `SpaceAfter=No` feature.

(8.35)
```
1   Stracił   stracić     VERB   …  SpaceAfter=No
2   em        być        AUX    …  _
3   równowagę równowaga   NOUN   …  SpaceAfter=No
4   .         .          PUNCT  …  _
```

In the case of Polish, both representations give exactly the same information and may be easily converted one to another.

The second – minor – difference is that $\text{UD}^{\text{PL}}_{\text{SZ}}$ does not indicate the lack of space between a preposition and the following short pronominal form, as in *doń* 'to him(/it/her)' (again, see Section 8.1) – neither via the `SpaceAfter=No` feature, nor via an additional line for such a multi-token unit. This error should be easy to correct in future releases of $\text{UD}^{\text{PL}}_{\text{SZ}}$.

8.5.2 Morphosyntax

There are various morphosyntactic differences between the two treebanks; some – discussed immediately below – stem from some controversial decisions taken by the developers of $\text{UD}^{\text{PL}}_{\text{SZ}}$, other are probably the result of lack of certain kinds of information in the input data converted to $\text{UD}^{\text{PL}}_{\text{SZ}}$, and still other are minor errors, which should be easy to correct in future editions of $\text{UD}^{\text{PL}}_{\text{SZ}}$.

Polish has five genders (Mańczak 1956),[16] including three masculine genders sometimes – misleadingly – called 'human masculine', 'animate masculine' and 'inanimate masculine'. There are good morphosyntactic tests making it possible to distinguish the three (sub)genders, without any recourse to semantic intuition. As discussed in Section 8.2.4, the correlation between the three masculine genders and the animacy feature is far from perfect. For this reason, the three masculine genders are distinguished in $\text{UD}^{\text{PL}}_{\text{LFG}}$ via the values of the `SubGender` feature. In $\text{UD}^{\text{PL}}_{\text{SZ}}$, however, the `Animacy` feature is employed to this end, with three possible values: `Hum` for 'human masculine', `Nhum` for 'animate masculine' and `Inan` for 'inanimate masculine'. This is highly misleading – the cursory inspection of the 150 lemmata whose forms are marked as 'animate masculine' `NOUN`s in $\text{UD}^{\text{PL}}_{\text{SZ}}$ suggests that perhaps only about half of them refer to animals. For example, considering such lemmata starting in T, only two out of seven are semantically animate:

[16] More on some accounts, e.g., nine according to Saloni 1976.

- TAROT – cards for divination,
- TENIS – 'tennis',
- TIR – a heavy vehicle,
- TRUP – 'corpse',
- TRZECI – 'third' (possibly an error in input data),
- TRZMIEL – 'bumblebee',
- TYGRYS – 'tiger'.

Only the last two are semantically animate.

A closely related problem stems from the lack of proper handling of 'derogatory' forms of 'human masculine' nouns in UD_{SZ}^{PL}, e.g., *profesory* 'professors (derogatory)' vs. the neutral *profesorowie*. Such forms behave morphosyntactically as if they were 'animate masculine', so they are marked as `Animacy=Nhum` in UD_{SZ}^{PL}, even though they are without exception semantically human masculine. (This problem is statistically insignificant, though, as it only concerns four tokens.) Recall that in UD_{LFG}^{PL} such derogatory forms are marked as `Polite=Depr`.

Another controversial decision – or perhaps simply a conversion error – is the annotation of impersonal *-no/-to* forms as adjectival passive participles in UD_{SZ}^{PL}, i.e., as tokens with the `ADJ` coarse part of speech and with `VerbForm=Part` and `Voice=Pass`, as well as, somewhat curiously, `Case=Nom`, `Gender=Neut` and `Number=Sing` among their features. Tokens such as *wyrzucano* 'one used to throw away' or *zdobyto* 'one conquered', are – uncontroversially – purely verbal, with no grammatical case, no clear values of number and gender, and they may be formed from verbs which do not passivise at all. In UD_{LFG}^{PL} they are treated as finite verbs with the distinguishing feature `Person=0` marking their morphologically impersonal status.

Three other differences probably stem from the lack of appropriate information in the data that was used to develop UD_{SZ}^{PL}. First, UD_{SZ}^{PL} does not distinguish between relative and interrogative uses of various forms of such (broadly understood) pronouns as KTO 'who', CO 'what' and KTÓRY 'which', marking them all as `PronType=Int,Rel`, i.e., as 'interrogative or relative'. In contrast, such pronouns are appropriately marked as interrogative or as relative in the LFG structure bank described in the first part of this monograph, i.e., they are disambiguated in UD_{LFG}^{PL}.

Second, the UD coarse part of speech tag `X`, "used for words that for some reason cannot be assigned a real part-of-speech category",[17] is used in UD_{SZ}^{PL} in two situations. One is easy to correct (as well as rare) and concerns predicative-only (short) adjectives – such forms should be tagged as `ADJ` and assigned the `Variant=Short` feature. The other concerns 273 tokens (with 46 different lemmata) of abbreviations. Such abbreviations are tagged with specific parts of speech (in morphosyntactic features) in UD_{LFG}^{PL}, but only as `X` in UD_{SZ}^{PL}.

Third, last and certainly least, UD_{SZ}^{PL} does not distinguish between prepositions and postpositions, marking them all as `AdpType=Prep`. But as there is only one clear exception to the generalisation that Polish adpositions are always prepositions, namely, the postposition TEMU 'ago', this only affects 28 tokens representing this lemma.

[17] See `http://universaldependencies.org/u/pos/X.html`, accessed on 1 March 2018.

8.5.3 Syntax

The fundamental difference between $\text{UD}^{\text{PL}}_{\text{SZ}}$ and $\text{UD}^{\text{PL}}_{\text{LFG}}$ is the presence of enhanced dependencies in the latter. The intensive use of secondary edges in $\text{UD}^{\text{PL}}_{\text{LFG}}$ makes it possible to express many syntactic relations absent in $\text{UD}^{\text{PL}}_{\text{SZ}}$, including grammatical control and sharing of dependents between conjuncts in coordinate structures.

Apart from this, probably the biggest conceptual difference between the two UD treebanks of Polish concerns the **argument–adjunct distinction**, as well as the definition of direct and indirect objects. $\text{UD}^{\text{PL}}_{\text{LFG}}$ attempts to follow the general UD philosophy of not trying to distinguish arguments from adjuncts:

> The UD taxonomy is centered around the fairly clear distinction between core arguments (subjects, objects, clausal complements) versus other dependents. It does not make a distinction between adjuncts (general modifiers) versus oblique arguments (arguments said to be selected by a head but not expressed as a core argument).[18]

We strongly believe that the argument–adjunct distinction is untenable (Patejuk and Przepiórkowski 2016; Przepiórkowski 2016a, 2016b, 2017a, 2017b), so nominal core arguments (subjects, direct and indirect objects) are defined in $\text{UD}^{\text{PL}}_{\text{LFG}}$ in a narrow and rather traditional way, with the effect that many broadly nominal – both bare nominal and prepositional – dependents which would traditionally be classified as complements (i.e., hence, arguments) are not distinguished from traditional nominal adjuncts.

On the other hand, $\text{UD}^{\text{PL}}_{\text{SZ}}$ reintroduces the argument–adjunct distinction: apart from defining objects in a very broad way (see below), it also splits the oblique dependents into arguments, marked as `obl:arg`, and adjuncts, marked as `obl` (without any explicit subtype). The proposal to re-introduce argument–adjunct distinction into UD is explicitly presented in Zeman 2017.

The related important difference is the definition of **direct objects**, marked as `obj`. In $\text{UD}^{\text{PL}}_{\text{LFG}}$, direct object is defined in a precise and at the same time traditional (e.g., Gołąb et al. 1968: 132, Urbańczyk 1992: 62) way as that dependent of a verb which is realised as the subject in passive occurrences of this verb. On the other hand, in $\text{UD}^{\text{PL}}_{\text{SZ}}$ the label `obj` is used for all (non-subject) bare nominal arguments, whether they passivise or not. Given that there are also bare nominal adjuncts in Polish, this definition of direct objects again presupposes the argument–adjunct distinction. Also, $\text{UD}^{\text{PL}}_{\text{SZ}}$ treats subcategorised clauses, marked as `ccomp`, as direct objects. Since there is a ban on two direct object dependents of a single verb, the situation where one verb has a `ccomp` dependent and an `obj` dependent is not allowed – as discussed immediately below, the direct object is re-analysed as an indirect object.

Also the definitions of **indirect objects**, `iobj`, differ in the two treebanks, although neither is optimal. In $\text{UD}^{\text{PL}}_{\text{LFG}}$, indirect objects are defined as subcategorised bare dative (non-passivisable) dependents; the subcategorisation requirement re-introduces – albeit in a very limited way – the argument–adjunct dichotomy. While such limited references to this dichotomy are present

[18] See `http://universaldependencies.org/u/overview/syntax.html`, accessed on 2 March 2018.

also elsewhere in the UD standard, this goes against the spirit of UD and should be changed in future editions of UD_{LFG}^{PL}; since traditional Polish grammars do not recognise the class of indirect objects, perhaps all `iobj` labels should simply be replaced by `obl` labels. The definition of indirect objects in UD_{SZ}^{PL} is even more questionable: if there are two candidates for the direct object dependency, only one is assigned the `obj` label. In particular, if a subcategorised clause is one of the two candidates, it is assigned the status of direct object, and the bare accusative dependent receives the `iobj` label. This leads to some annotations which are in direct conflict with any linguistic definition of direct objects. For example, in (8.36) (sentence `train-s2613` in UD_{SZ}^{PL}), the verb *spytało* 'asked' combines with the numeral subject *kilka osób* 'several people', the accusative nominal *mnie* 'me' and the subordinate clause *czy jestem…* 'whether I am…'; since the subordinate clause is subcategorised, it is marked as `ccomp`, but that means that *mnie* 'me' must be marked as indirect object, `iobj`, even though it becomes the subject under passivisation and it occurs in the accusative case, so it is a prototypical direct object.

(8.36) Kilka osób spytało mnie, czy jestem dzięki feminizmowi
several people asked me.ACC.SG whether am.1SG thanks feminism.DAT
szczęśliwsza.
happier.NOM.SG.F
'Some people have asked me whether feminism made me happier.'

This leads to obvious inconsistencies, as in other sentences, lacking such subordinate clause dependents, similar accusative dependents are correctly marked as direct objects, as is the case with *ją* 'her' in (8.37) (sentence `train-s2739` in UD_{SZ}^{PL}):

(8.37) Chciał ją spytać o wiele rzeczy.
wanted her.ACC ask.INF about many things
'He wanted to ask her about many things.'

Note that *mnie* 'me' in (8.36) and *ją* 'her' in (8.37) bear exactly the same semantic role with respect to the two forms of the verb SPYTAĆ 'ask' and have the same grammatical properties (passivisability, grammatical case, etc.), so this is a clear case of intra-linguistic annotation inconsistency.

A somewhat related difference concerns the so-called reflexive marker SIĘ, whose two uses are distinguished in UD_{SZ}^{PL}: broadly anaphoric, in which case it is marked as `obj` (or `iobj`, if there is a better candidate for the `obj` label; see above), and inherent, in which case it is marked as `expl:pv`. It is not clear how many different SIĘ elements should be assumed – various linguistic works make different assumptions here – but it is clear that at least one more function of SIĘ should be carefully distinguished, namely, impersonal (see, e.g., the second *się* in example (8.22) on page 196 above). In UD_{SZ}^{PL}, they are lumped together with inherent uses, while in UD_{LFG}^{PL} they are assigned the `expl:impers` label.

The two treebanks differ also in the representation of two valency features: the grammatical case required by prepositions and the conditions on the nominal accompanying a numeral. In UD_{LFG}^{PL}, since both are syntactic properties of particular forms which may or may not surface in a given sentence (as adpositions and numerals may in some constructions appear without

Table 8.2: Quantitative comparison of $\text{UD}^{\text{PL}}_{\text{SZ}}$ and $\text{UD}^{\text{PL}}_{\text{LFG}}$

	$\textbf{UD}^{\text{PL}}_{\textbf{SZ}}$	$\textbf{UD}^{\text{PL}}_{\textbf{LFG}}$
sentences (running)	8227	17,246
sentences (different)	8139	17,190
tokens (running)	84,316	130,967
lemmata (different)	13,688	15,797

the normally required nominal phrases), these valency features are uniformly represented in the MISC field: the case required by an adposition as the value of the `Case` feature (to be distinguished from the inflectional `Case` feature in the FEATS field), and the information about the combinatory potential of a numeral as the value of the language-specific `DepType` feature: `Rec` if the numeral subcategorises for a genitive nominal, and `Congr` if the numeral and the nominal agree in case. In contrast, the two valency features are represented differently in $\text{UD}^{\text{PL}}_{\text{SZ}}$. The required case information is represented the same way as information about inflection case value of a given token, i.e., via the FEATS `Case` feature, and the information that a numeral governs the genitive case is represented as subtypes of dependency relations: `nummod:gov` or `det:numgov`.[19]

8.5.4 Underlying data

The ultimate source of texts and original morphosyntactic information in $\text{UD}^{\text{PL}}_{\text{SZ}}$ is the 1-million-word manually annotated subcorpus of the *National Corpus of Polish*, which is also the source of almost 85% of texts in $\text{UD}^{\text{PL}}_{\text{LFG}}$. This means that the values of the XPOS field are taken from the same tagset, but – given that some morphosyntactic analyses were modified in $\text{UD}^{\text{PL}}_{\text{LFG}}$ – not that they would necessarily be identical for the same sentence in the two treebanks. For example, typical numerals in the subject position are marked as nominative in $\text{UD}^{\text{PL}}_{\text{SZ}}$ but as accusative in $\text{UD}^{\text{PL}}_{\text{LFG}}$, in accordance with the analysis of such subjects in Przepiórkowski 1999, 2004a (see also fn. 13 on page 204).

The sizes of the two treebanks are compared in Table 8.2. $\text{UD}^{\text{PL}}_{\text{LFG}}$ is much larger: it contains 17,246 running sentences (17,190 types; duplicate sentences have different analyses), compared to 8227 running sentences in $\text{UD}^{\text{PL}}_{\text{SZ}}$ (8139 types; duplicate sentences may have the same analyses). In terms of running tokens, the respective numbers are 130,967 ($\text{UD}^{\text{PL}}_{\text{LFG}}$) vs. 84,316 ($\text{UD}^{\text{PL}}_{\text{SZ}}$), which implies that $\text{UD}^{\text{PL}}_{\text{SZ}}$ sentences are longer on the average. $\text{UD}^{\text{PL}}_{\text{LFG}}$ is also a little richer lexically (which is to be expected, given the bigger size).

[19] In practice, `nummod:gov` labels are missing in the 2.1 release of $\text{UD}^{\text{PL}}_{\text{SZ}}$, so cardinal numerals which require the genitive case are not marked as such.

Coda

Chapter 9

Lost in Translation?

One aim of this monograph has been to describe two linguistically-informed language resources for Polish: an LFG structure bank (in Part I) and an enhanced UD treebank (in Part III). Another aim has been to present the procedure of translating LFG syntactic structures into UD dependency representations (in Part II). As is well known, dependency trees are less expressive than functional structures of LFG. One reason is that they do not make it possible to represent shared dependents, e.g., the fact that a dependent of a higher verb is at the same time the subject of the lower verb in control or raising constructions. For example, in (9.1), whose f-structure is given in Figure 9.1, *Poczta* '(Polish) Post', is the subject not only of the two conjoined finite verbs, *zmniejsza* 'reduces' and *powinna* 'should', but also of the controlled verbs *zacząć* 'start' and *przynosić* 'bring, make'. This is directly expressed in the f-structure (see the multiple occurrences of the substructure 156 in Figure 9.1), as well as in the LFG-like dependency representation in Figure 9.2, which is not a dependency tree, but a more complex graph. On the other hand, this information is lost in the basic UD tree in the upper part of Figure 9.3.

(9.1) Poczta zmniejsza swój deficyt i już w 1997 r. powinna
post.NOM.SG.F reduces.3SG.F self's deficit and already in 1997 year should.3SG.F
zacząć przynosić zyski.
start.INF bring.INF profits.ACC
'Post reduces its deficit and it should start to make profit already in 1997.'

However, such information is easy to represent in the enhanced UD graph, which does not have to be a tree. Hence, in the case at hand, the information that four verbs share the same subject is not lost in the (enhanced) UD representation. The natural question is then, to what extent – if any – information is lost in the translation from syntactic structures assumed in LFG to enhanced Universal Dependencies.

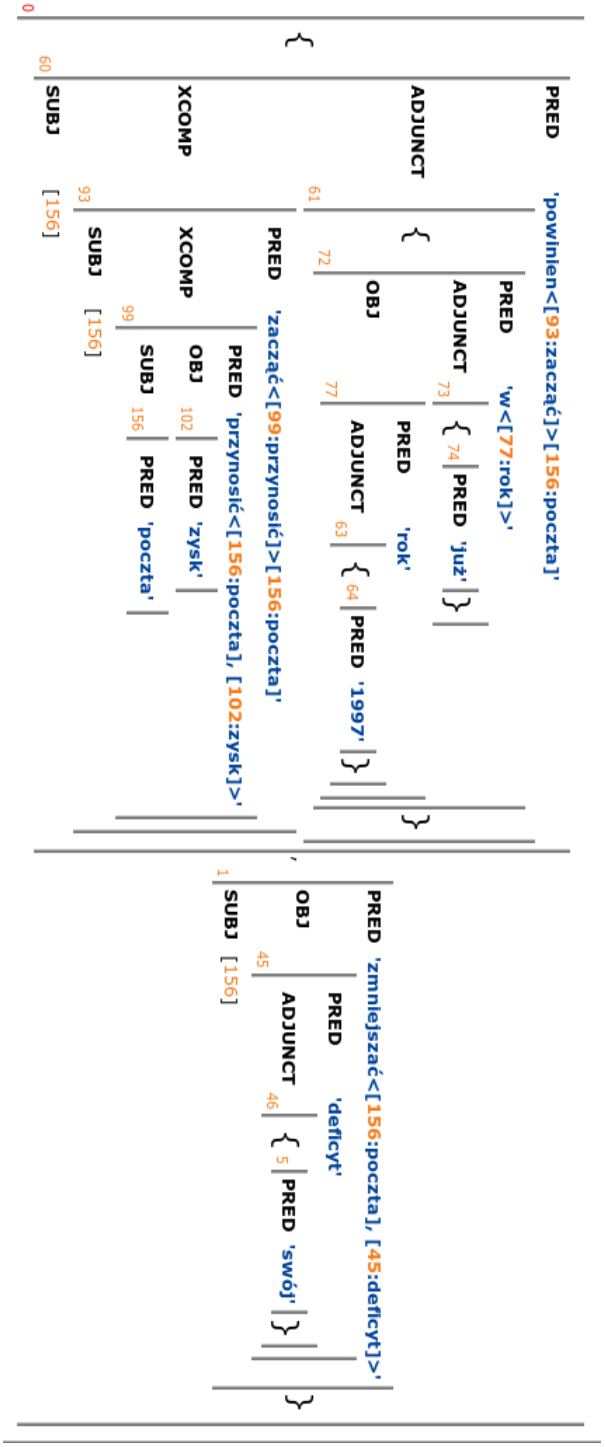


Figure 9.1: Schematic f-structure of (9.1)

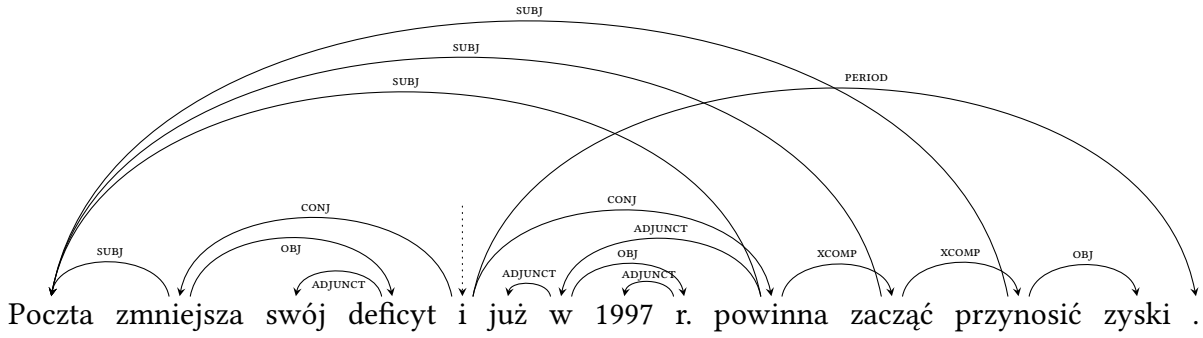


Figure 9.2: Initial dependency representation of (9.1)

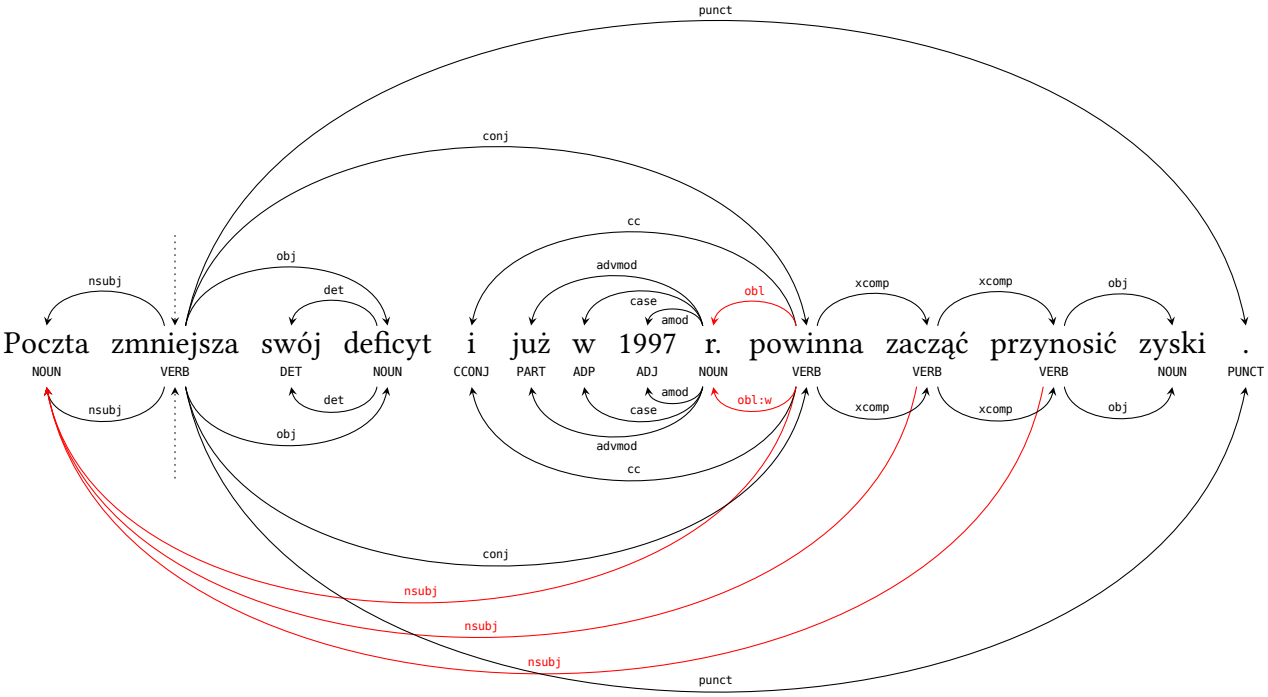
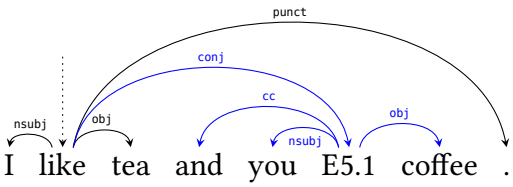


Figure 9.3: Final dependency representation of (9.1)

9.1 Empty dependents not allowed

Clear loss of information results from the fact that UD does not make it possible to represent *pro*-dropped dependents. This is not a matter of a general ban on null nodes in dependency representations: enhanced UD allows for the possibility to represent elided *predicates*, as in the following example from the UD guidelines:[1]

(9.2)



Here, *E5.1* is an artificial token, added to the input sentence *in lieu* of the elided verb *like*. However, similar addition of tokens standing for *pro*-dropped *dependents* is currently prohibited, with the effect that information is lost in the conversion of some of the examples given above.

Consider again example (7.3), repeated below, and its simplified f-structure in Figure 7.9, repeated below as Figure 9.4.

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

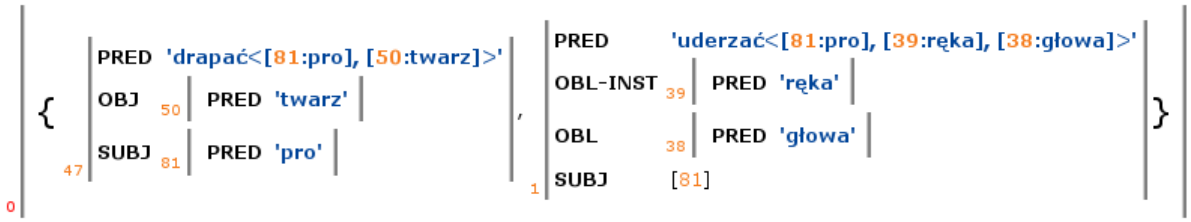


Figure 9.4: Schematic f-structure of (7.3)

Polish is a rampantly *pro*-drop language, and in this sentence the *pro*-dropped subject is shared between the two finite verbs (see the substructure with index 81 in Figure 9.4). That is, the same person is understood to have done the pounding and the scratching. In contrast, the UD representation in Figure 9.5 misses this information – it is underspecified as to whether the same person performed the two actions. The same problem occurs in many other examples discussed in this monograph.

A related problem is that, in the case of the *pro*-drop of the controller, information is lost about the reference of the subject of the controlled verb. In the absence of *pro*-drop, this information is given explicitly in the enhanced representation; for example, in the UD representation in Figure 9.3, the controlled verbs are those with the incoming `xcomp` dependency – i.e., *zacząć* 'start' and *przynosić* 'bring, make' – and their subjects are marked by the `nsubj` enhanced

[1] See `http://universaldependencies.org/u/overview/enhanced-syntax.html#ellipsis`; the dependencies in blue are present only in the enhanced representation.

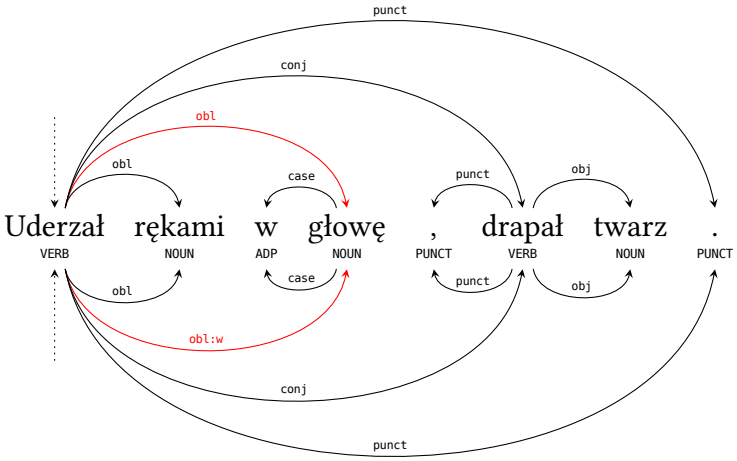


Figure 9.5: UD representation of (7.3)

dependencies to *Poczta* 'Post'. Consider, however, example (9.3), involving the control verb *kazał* 'ordered'.

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

Two arguments of this verb are *pro*-dropped: the subject and the dative argument which controls the subject of the infinitival *odsyłać* 'send back'. This information is explicitly represented in the f-structure in Figure 9.6. In particular, the SUBJect of the controlled verb, i.e., the substructure with index 25, is the same as the dative argument of the main verb, i.e., as the value of the OBJ-TH attribute there. Unfortunately, there is currently no way to represent this information in the UD structure – see Figure 9.7.

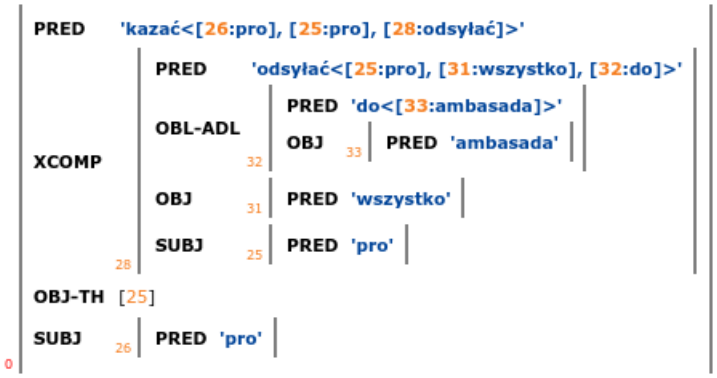


Figure 9.6: Schematic f-structure of (9.3)

Another problem stemming from the lack of any representation of *pro*-dropped dependents concerns the representation of non-core (not subcategorised, not required) secondary predicates, e.g., *pierwszy* 'first' in (7.29), repeated below, and *osłupiały* 'transfixed, shocked' in (9.4):[2]

[2]An analogous problem occurs in the case of (subcategorised, required) predicative complements.

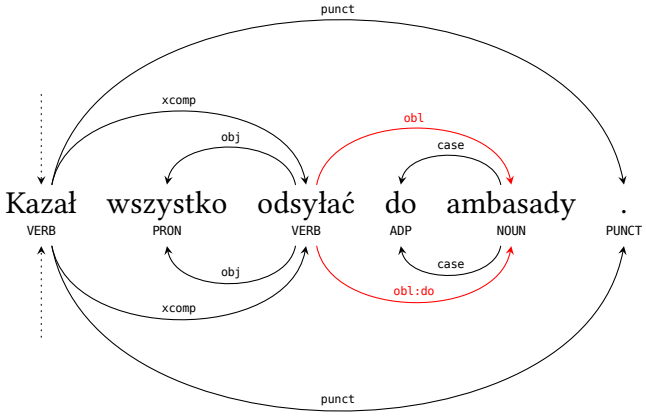


Figure 9.7: UD representation of (9.3)

(7.29) Król zaatakował pierwszy.
king.NOM.SG.M attacked.3SG.M first.NOM.SG.M
'The king attacked (as) first.'

(9.4) Przez chwilę stał osłupiały.
for while stood.3SG.M transfixed.NOM.SG.M
'He stood transfixed for a while.'

Such non-core secondary predicates are `acl` dependents of the nouns they predicate of, as shown in Figure 9.8. However, such an overt target of predication is missing in (9.4), in which case the secondary predicate should be an `advcl` dependent of the verb that governs the *pro*-dropped argument, as shown in Figure 9.9. This not only results in rather different representations of the same phenomenon, but also representations such as Figure 9.9 are in the general case underspecified as to which of the potentially *pro*-dropped dependents of the verb the predicate refers to.[3]

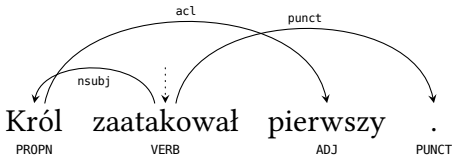


Figure 9.8: UD representation of (7.29)

While the prohibition on explicit representation of *pro*-dropped dependents is probably the most important source of information loss in the conversion procedure described above, we do not see it as a fundamental problem with UD representation: once this arbitrary prohibition is lifted, the problems described in this section should disappear.

[3]On the other hand, in most – but not all – instances the case value of the secondary predicate should make this clear.

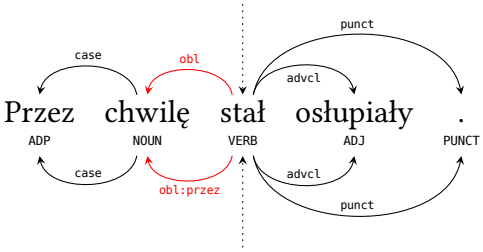


Figure 9.9: UD representation of (9.4)

9.2 Multiple dependencies between same tokens not allowed

A statistically insignificant problem, but one that did occur in the conversion process, is that it is illegal at the moment, even in enhanced dependencies, to have two different edges from token A to token B. The need for such a representation arises in those – admittedly very rare – cases where the multi-functional reflexive marker *się* plays two roles at the same time (Patejuk and Przepiórkowski 2015a), e.g., being a marker of an inherently reflexive verb (`expl:pv`) and being a part of an impersonal construction (`expl:impers`). A treebank example exhibiting this problem is (7.31), repeated below. As discussed in Section 7.2.6, the first *się*, in *uczestniczyło się* 'one participated', is purely impersonal, and the second *się*, in *modliło się* 'one prayed', is impersonal and also an inherent part of the verb MODLIĆ SIĘ 'pray', so it should bear two relations to *modliło*: `expl:impers` and `expl:pv`.

(7.31) W Laskach w liturgii uczestniczyło się przez cały dzień i modliło
in Laski in liturgy participated.3SG.N RM.IMPS for whole day and prayed.3SG.N
się wszędzie.
RM.INH.IMPS everywhere
'In Laski, one would take part in the liturgy for the whole day and one would pray everywhere.'

It seems that the ban on multiple edges could be lifted in the enhanced UD without any ill consequences.

9.3 Embedded coordination

A problem known to the UD community[4] is that there is no way to distinguish between embedded coordination, with the first conjunct itself being a coordinate structure, and flat coordination. There are about a dozen sentences in the Polish UD treebank described here where this is a potential problem, including the following:

[4] `http://universaldependencies.org/u/dep/conj.html`

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

In the LFG structure bank which is the input to the conversion procedure, this sentence is represented as a coordinate structure with the conjunction *ale* 'but'. The linearly first conjunct is also a coordinate structure, with comma acting as the conjunction – see the f-structure in (9.10). This embedding of coordination cannot be directly represented in UD – see Figure 9.11, which does not distinguish between flat ternary coordination and such binary coordination embedded within binary coordination.[5]

In practice, however, this is not a serious problem, as the right structure can usually – at least in the dozen or so cases in the current treebank – be inferred from the linear placement and kind of conjunctions. For example, a strictly binary contrastive conjunction *ale* is used in (9.5), so Figure 9.11 cannot represent flat ternary coordination – it must represent embedded coordination.

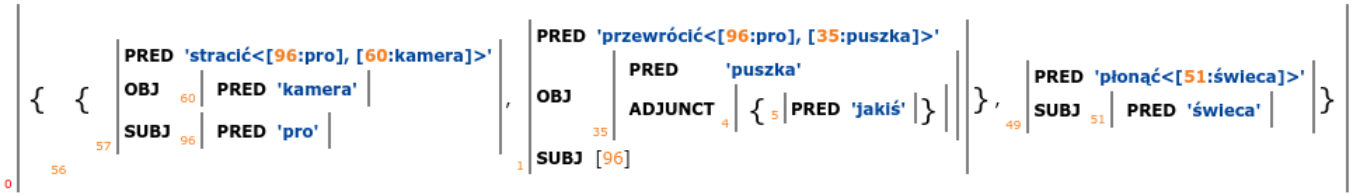


Figure 9.10: Schematic f-structure of (9.5)

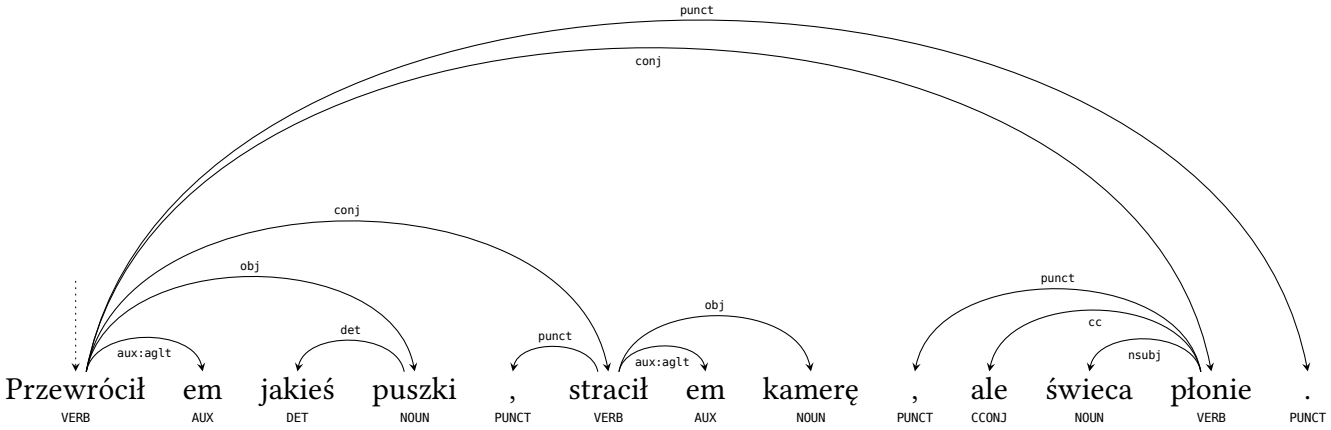


Figure 9.11: UD representation of (9.5)

[5]Only the basic tree is shown here, as the enhanced dependency representation is identical to this basic tree.

9.4 Insufficient information in dependency labels

Much information is also lost because UD dependency labels are less informative than LFG attributes. For example, while LFG distinguishes between different kinds of oblique arguments (e.g., only in the f-structures given above: OBL, OBL-STR, OBL-INST, OBL-ADL, etc.), and distinguishes them from adjuncts, UD treats all such obliques and adjuncts alike, and marks them as `obl`. However, it is easy to extend UD in a way that makes representing such information possible. To this end, the mechanism of subtypes – already alluded to above (e.g., the relations `expl:pv` and `expl:impers` are subtypes of the general `expl`(etive) relation) – may be used. In fact, Zeman 2017 proposes to distinguish oblique arguments from adjuncts by subtyping the former to `obl:arg`, and similar subtypes may be used, e.g., to represent adlative oblique arguments as, say, `obl:adl`, etc.

The same mechanism may be used to re-introduce many other kinds of information currently lost in translation, including:

- the distinction between control 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 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 eventuality and constituent negation (Przepiórkowski and Patejuk 2015), e.g., via the subtypes `advmod:eneg` and `advmod:cneg`,
- the distinction between semantic and non-semantic prepositions, e.g., by subtyping the `case` relation in the former to `case:sem`; etc.

9.5 Summary

The exercise described in Part II of this monograph demonstrates that it is relatively easy to convert an LFG treebank into a full-blown enhanced UD representation. As discussed in this concluding chapter, surprisingly little information is lost in the conversion from LFG to enhanced UD and most of the loss is not caused by any fundamental issues with UD, but rather due to the contingent – and easily rectified – decision of the UD developers not to represent certain kinds of information, such as *pro*-dropped dependents or subtypes of oblique dependents. A more basic problem concerns the representation of coordination which theoretically does not make it possible to distinguish between flat coordinate structures and certain embedded structures, but – as discussed in Section 9.3 – this problem is negligible in practice.[6]

In more general terms, we hope that the work described in this monograph has shown that Universal Dependencies is not only a utilitarian standardisation effort, but also a framework that may be of some interest to theoretical linguists. However, the converse is also true: Universal Dependencies should benefit from attempts – such as the one described in Part II – of

[6] See also Przepiórkowski and Patejuk 2018 for a discussion of some more fundamental problems related to the UD approach to arguments and adjuncts.

translating into UD linguistically advanced treebanks exemplified by the LFG structure bank presented in Part I. In our opinion, such conversion exercises help identify strong and weak areas of UD and may suggest ways of further development of the standard. We hope that the work presented in this monograph will inspire other researchers and will promote further dialogue between theoretical linguists and developers of language tools and resources.

Appendices

Appendix A

Legacy tagset

This appendix summarises the tagset used in the XPOS field of the CoNLL-U representation of UD structures. This is a slightly constrained version of the National Corpus of Polish (`http://nkjp.pl/`; Przepiórkowski et al. 2011, 2012), which itself is a relatively minor modification of the IPI PAN tagset (Woliński and Przepiórkowski 2001; Przepiórkowski and Woliński 2003a, 2003b) used earlier in the IPI PAN Corpus of Polish (Przepiórkowski 2004b). The full NKJP tagset is described on-line at `http://nkjp.pl/poliqarp/help/en.html`, among other places.

The following grammatical categories – and their values – are assumed in the tagset, illustrated with Polish forms bearing these values:

Number		
singular	sg	*oko*
plural	pl	*oczy*
Case		
nominative	nom	*woda*
genitive	gen	*wody*
dative	dat	*wodzie*
accusative	acc	*wodę*
instrumental	inst	*wodą*
locative	loc	*wodzie*
vocative	voc	*wodo*
Gender		
'human masculine' (virile)	m1	*papież, kto, wujostwo*
'animate masculine'	m2	*baranek, walc, babsztyl*
'inanimate masculine'	m3	*stół*
feminine	f	*stuła*
neuter	n	*dziecko, okno, co, skrzypce, spodnie*

Person		
first	pri	*bredzę, my*
second	sec	*bredzisz, wy*
third	ter	*bredzi, oni*
Degree		
positive	pos	*cudny*
comparative	comp	*cudniejszy*
superlative	sup	*najcudniejszy*
Aspect		
imperfective	imperf	*iść*
perfective	perf	*zajść*
Negation		
affirmative	aff	*pisanie, czytanego*
negative	neg	*niepisanie, nieczytanego*
Accentability		
accented (strong)	akc	*jego, niego, tobie*
non-accented (weak)	nakc	*go, -ń, ci*
Post-prepositionality		
post-prepositional	praep	*niego, -ń*
non-post-prepositional	npraep	*jego, go*
Accommodability		
agreeing	congr	*dwaj, pięcioma*
governing	rec	*dwóch, dwu, pięciorgiem*
Agglutination		
non-agglutinative	nagl	*niósł*
agglutinative	agl	*niosł-*
Vocalicity		
vocalic	wok	*-em, ze*
non-vocalic	nwok	*-m, z*

The following table lists the grammatical classes (very fine-grained parts of speech) assumed in the tagset, together with information about the grammatical categories appropriate for each class; $\oplus$ indicates that lexemes of a given class typically inflect for this category (e.g., nouns inflect for number and case), and $\odot$ – that lexemes of this class have this category set lexically (e.g., each noun has – but does not inflect for – gender).

		number	case	gender	person	degree	aspect	negation	accentability	post-prepositionality	accommodability	agglutination	vocalicity
noun	subst	⊕	⊕	⊙									
derogatory form	depr	⊙	⊕	⊙									
main numeral	num	⊙	⊕	⊕							⊕		
collective numeral	numcol	⊙	⊕	⊙							⊕		
adjective	adj	⊕	⊕	⊕		⊕							
ad-adj. adjective	adja												
post-prep. adjective	adjp												
pred. adjective	adjc												
adverb	adv					⊕							
pronoun (non-3rd person)	ppron12	⊙	⊕	⊕	⊙				⊕				
pronoun (3rd person)	ppron3	⊕	⊕	⊕	⊙				⊕	⊕			
pronoun SIEBIE	siebie		⊕										
non-past form	fin	⊕			⊕		⊙						
future BYĆ	bedzie	⊕			⊕		⊙						
agglut. BYĆ	aglt	⊕			⊕		⊙						⊕
l-participle	praet	⊕		⊕			⊙					⊕	
imperative form	impt	⊕			⊕		⊙						
impersonal form	imps						⊙						
infinitive	inf						⊙						
adv. contemp. prtcp.	pcon						⊙						
adv. anter. prtcp.	pant						⊙						
gerund	ger	⊕	⊕	⊙			⊙	⊕					
adj. act. prtcp.	pact	⊕	⊕	⊕			⊙	⊕					
adj. pass. prtcp.	ppas	⊕	⊕	⊕			⊙	⊕					
winien-like verb	winien	⊕		⊕			⊙						
predicative	pred												
preposition	prep		⊙										
coord. conjunction	conj												
subord. conjunction	conj												
particle-adverb	qub												
interjection	interj												
punctuation	interp												

Appendix B

LFG syntactic representation in TigerXML

This appendix contains the complete XML representation of the two syntactic LFG structures (constituency and functional) of the running example of Chapter 4, repeated below for convenience:

(4.1) Mężczyzna nie zdążył ich otworzyć.
man.NOM.SG.M NEG managed.3SG.M them.GEN open.INF
'The man didn't manage to open them on time.'

```
<?xml version='1.0' encoding='UTF-8'?>
<subcorpus name="NKJP1M_1305000000506_morph_1-p_morph_1.40-s-dis@1"
           sentence="Mężczyzna nie zdążył ich otworzyć.">
  <s id="NKJP1M_1305000000506_morph_1-p_morph_1.40-s-dis@1">
    <graph root="c_578">
      <terminals>
        <t id="c_5" lemma="mężczyzna" tag="+subst:sg:nom:m1" val="--" word="Mężczyzna"/>
        <t id="c_31" lemma="nie" tag="+qub" val="--" word="nie"/>
        <t id="c_37" lemma="zdążyć" tag="+praet:sg:m1:perf" val="--" word="zdążył"/>
        <t id="c_42" lemma="on" tag="+ppron3:pl:gen:m3:ter:akc:npraep" val="--" word="ich"/>
        <t id="c_55" lemma="otworzyć" tag="+inf:perf" val="--" word="otworzyć"/>
        <t id="c_56" lemma="--" tag="--" val="--" word="."/>
        <t id="f_0_PRED" lemma="--" tag="--" val="zdążyć" word="--"/>
        <t id="f_0_NEG" lemma="--" tag="--" val="+" word="--"/>
        <t id="f_6_TENSE" lemma="--" tag="--" val="past" word="--"/>
        <t id="f_2_PRED" lemma="--" tag="--" val="mężczyzna" word="--"/>
        <t id="f_2_GEND" lemma="--" tag="--" val="m1" word="--"/>
        <t id="f_7_PRED" lemma="--" tag="--" val="otworzyć" word="--"/>
        <t id="f_2_CASE" lemma="--" tag="--" val="nom" word="--"/>
        <t id="f_6_ASPECT" lemma="--" tag="--" val="perf" word="--"/>
        <t id="f_1__CAT" lemma="--" tag="--" val="praet" word="--"/>
        <t id="f_2_PERS" lemma="--" tag="--" val="3" word="--"/>
        <t id="f_2_NUM" lemma="--" tag="--" val="sg" word="--"/>
        <t id="f_6_MOOD" lemma="--" tag="--" val="indicative" word="--"/>
        <t id="f_9_CASE" lemma="--" tag="--" val="gen" word="--"/>
        <t id="f_9_PERS" lemma="--" tag="--" val="3" word="--"/>
        <t id="f_3__CAT" lemma="--" tag="--" val="subst" word="--"/>
        <t id="f_9_NUM" lemma="--" tag="--" val="pl" word="--"/>
```

```
    <t id="f_8__CAT" lemma="--" tag="--" val="inf" word="--"/>
    <t id="f_4_NSYN" lemma="--" tag="--" val="common" word="--"/>
    <t id="f_9_GEND" lemma="--" tag="--" val="m3" word="--"/>
    <t id="f_13_ASPECT" lemma="--" tag="--" val="perf" word="--"/>
    <t id="f_9_PRED" lemma="--" tag="--" val="on" word="--"/>
    <t id="f_11_NSYN" lemma="--" tag="--" val="pronoun" word="--"/>
    <t id="f_10__PPREP" lemma="--" tag="--" val="npraep" word="--"/>
    <t id="f_10__CAT" lemma="--" tag="--" val="pron" word="--"/>
    <t id="f_10__ACC" lemma="--" tag="--" val="akc" word="--"/>
    <t id="f_5_COMMON" lemma="--" tag="--" val="count" word="--"/>
    <t id="f_12_COMMON" lemma="--" tag="--" val="count" word="--"/>
  </terminals>
  <nonterminals>
    <nt id="c_578" cat="ROOT">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_69" label="--"/>
      <edge idref="c_570" label="--"/>
    </nt>
    <nt id="c_570" cat="S">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_564" label="--"/>
    </nt>
    <nt id="c_564" cat="IP">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_445" label="--"/>
      <edge idref="c_181" label="--"/>
    </nt>
    <nt id="c_181" cat="NP">
      <edge idref="f_2" label="f::"/>
      <edge idref="c_179" label="--"/>
    </nt>
    <nt id="c_179" cat="N">
      <edge idref="f_2" label="f::"/>
      <edge idref="c_178" label="--"/>
    </nt>
    <nt id="c_178" cat="SUBST">
      <edge idref="f_2" label="f::"/>
      <edge idref="c_5" label="--"/>
    </nt>
    <nt id="c_445" cat="IP">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_675" label="--"/>
      <edge idref="c_227" label="--"/>
      <edge idref="c_251" label="--"/>
    </nt>
    <nt id="c_227" cat="NEG">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_31" label="--"/>
    </nt>
    <nt id="c_251" cat="PRAET">
      <edge idref="f_0" label="f::"/>
      <edge idref="c_37" label="--"/>
    </nt>
    <nt id="c_675" cat="IP">
      <edge idref="f_7" label="f::"/>
```

```
  <edge idref="c_351" label="--"/>
  <edge idref="c_299" label="--"/>
</nt>
<nt id="c_299" cat="NP">
  <edge idref="f_9" label="f::"/>
  <edge idref="c_298" label="--"/>
</nt>
<nt id="c_298" cat="PRON">
  <edge idref="f_9" label="f::"/>
  <edge idref="c_297" label="--"/>
</nt>
<nt id="c_297" cat="PPRON3">
  <edge idref="f_9" label="f::"/>
  <edge idref="c_42" label="--"/>
</nt>
<nt id="c_351" cat="IP">
  <edge idref="f_7" label="f::"/>
  <edge idref="c_349" label="--"/>
</nt>
<nt id="c_349" cat="INF">
  <edge idref="f_7" label="f::"/>
  <edge idref="c_55" label="--"/>
</nt>
<nt id="c_69" cat="PERIOD">
  <edge idref="f_0" label="f::"/>
  <edge idref="c_56" label="--"/>
</nt>
<nt id="f_0" cat="_TOP">
  <edge idref="f_6" label="TNS-ASP"/>
  <edge idref="f_7" label="XCOMP"/>
  <edge idref="f_1" label="CHECK"/>
  <edge idref="f_0_PRED" label="PRED"/>
  <edge idref="f_0_NEG" label="NEG"/>
  <edge idref="f_2" label="SUBJ"/>
</nt>
<nt id="f_6" cat="--">
  <edge idref="f_6_MOOD" label="MOOD"/>
  <edge idref="f_6_ASPECT" label="ASPECT"/>
  <edge idref="f_6_TENSE" label="TENSE"/>
</nt>
<nt id="f_1" cat="--">
  <edge idref="f_1__CAT" label="_CAT"/>
</nt>
<nt id="f_2" cat="--">
  <edge idref="f_2_NUM" label="NUM"/>
  <edge idref="f_2_PERS" label="PERS"/>
  <edge idref="f_4" label="NTYPE"/>
  <edge idref="f_3" label="CHECK"/>
  <edge idref="f_2_PRED" label="PRED"/>
  <edge idref="f_2_GEND" label="GEND"/>
  <edge idref="f_2_CASE" label="CASE"/>
</nt>
<nt id="f_7" cat="--">
  <edge idref="f_8" label="CHECK"/>
  <edge idref="f_7_PRED" label="PRED"/>
```

```
        <edge idref="f_9" label="OBJ"/>
        <edge idref="f_13" label="TNS-ASP"/>
        <edge idref="f_2" label="SUBJ"/>
      </nt>
      <nt id="f_9" cat="--">
        <edge idref="f_9_CASE" label="CASE"/>
        <edge idref="f_10" label="CHECK"/>
        <edge idref="f_9_GEND" label="GEND"/>
        <edge idref="f_9_PERS" label="PERS"/>
        <edge idref="f_9_NUM" label="NUM"/>
        <edge idref="f_9_PRED" label="PRED"/>
        <edge idref="f_11" label="NTYPE"/>
      </nt>
      <nt id="f_13" cat="--">
        <edge idref="f_13_ASPECT" label="ASPECT"/>
      </nt>
      <nt id="f_4" cat="--">
        <edge idref="f_5" label="NSEM"/>
        <edge idref="f_4_NSYN" label="NSYN"/>
      </nt>
      <nt id="f_3" cat="--">
        <edge idref="f_3__CAT" label="_CAT"/>
      </nt>
      <nt id="f_8" cat="--">
        <edge idref="f_8__CAT" label="_CAT"/>
      </nt>
      <nt id="f_11" cat="--">
        <edge idref="f_11_NSYN" label="NSYN"/>
        <edge idref="f_12" label="NSEM"/>
      </nt>
      <nt id="f_10" cat="--">
        <edge idref="f_10__CAT" label="_CAT"/>
        <edge idref="f_10__PPREP" label="_PPREP"/>
        <edge idref="f_10__ACC" label="_ACC"/>
      </nt>
      <nt id="f_5" cat="--">
        <edge idref="f_5_COMMON" label="COMMON"/>
      </nt>
      <nt id="f_12" cat="--">
        <edge idref="f_12_COMMON" label="COMMON"/>
      </nt>
    </nonterminals>
  </graph>
 </s>
</subcorpus>
```

Appendix C

UD representations of conversion examples

This appendix presents final UD representations of those examples discussed in the conversion part of this monograph, in Chapter 7, which were not given such final representations there.

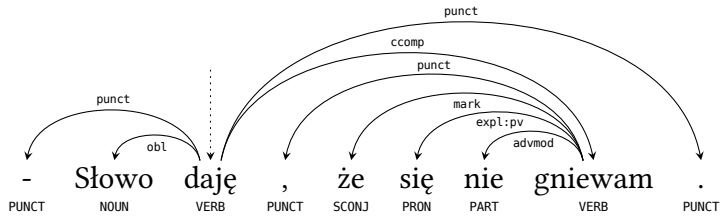


Figure C.1: UD representation of (7.2)

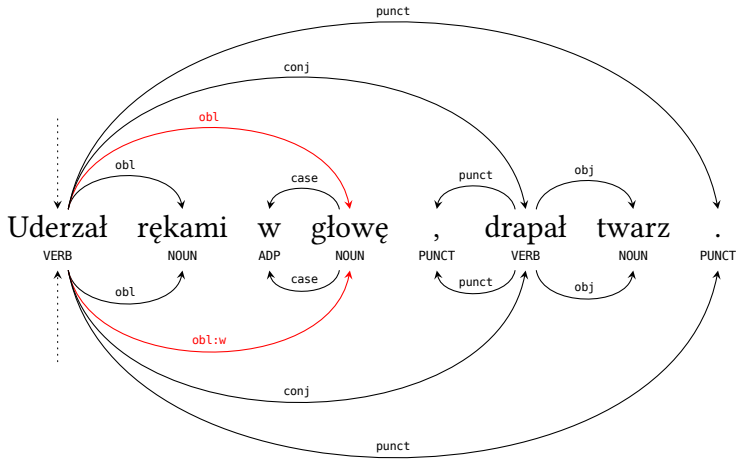


Figure C.2: UD representation of (7.3)

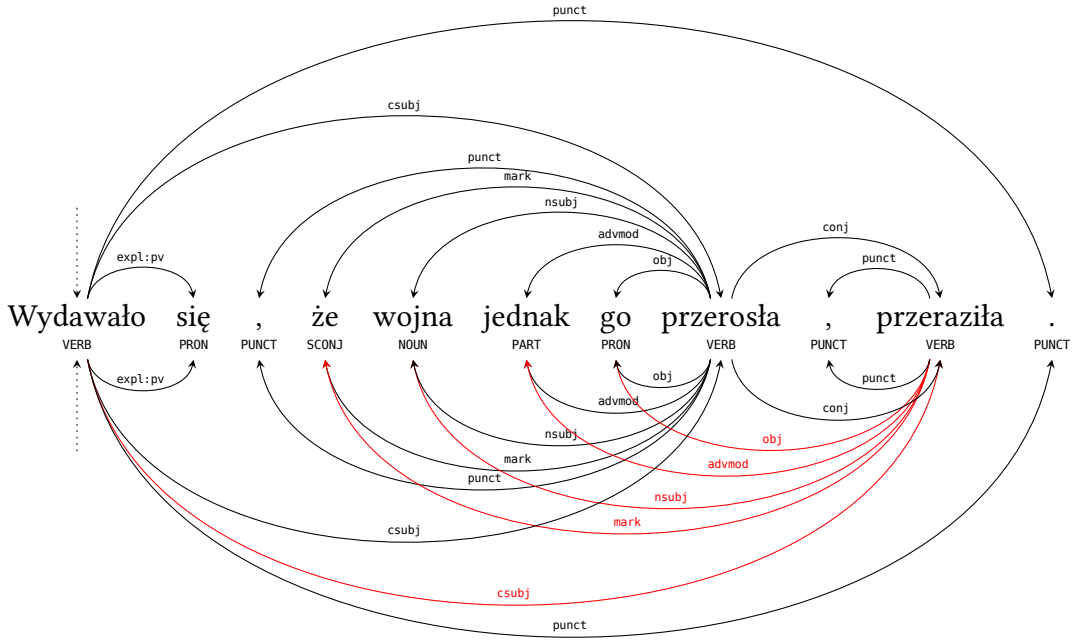


Figure C.3: UD representation of (7.4)

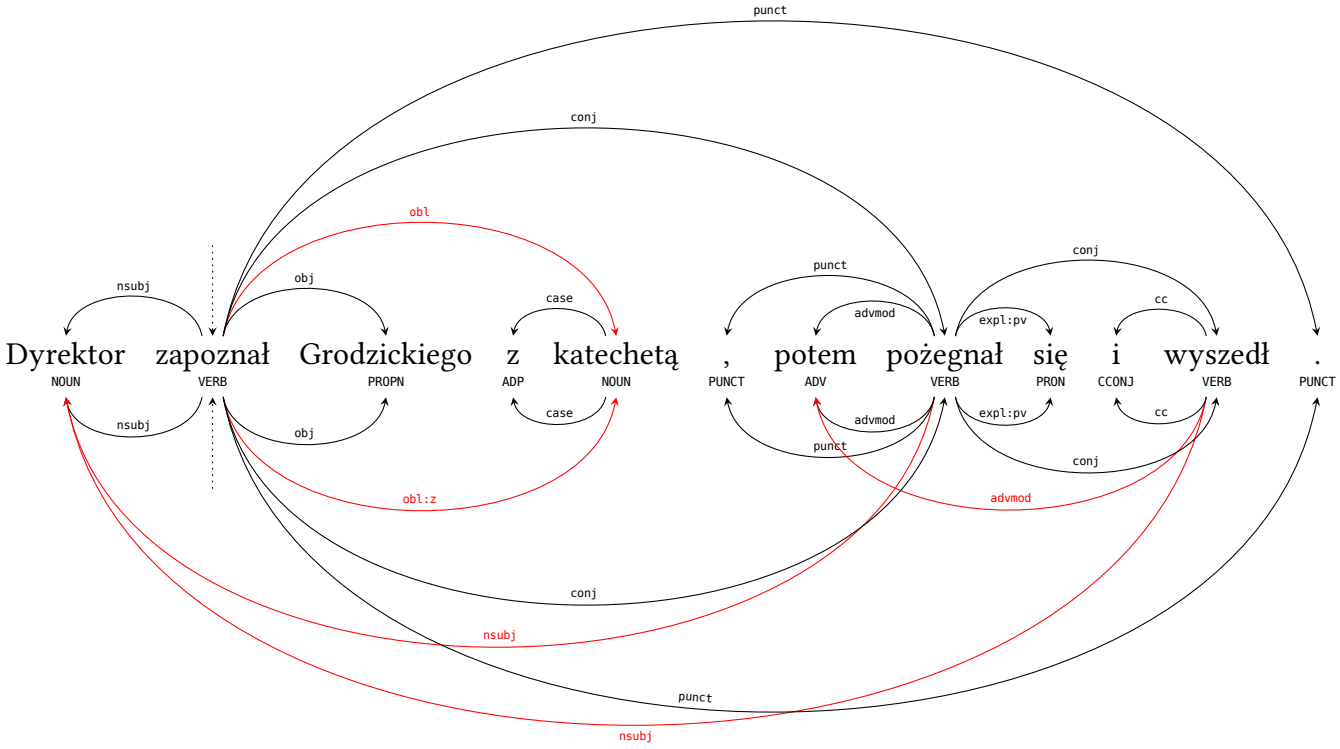


Figure C.4: UD representation of (7.6)

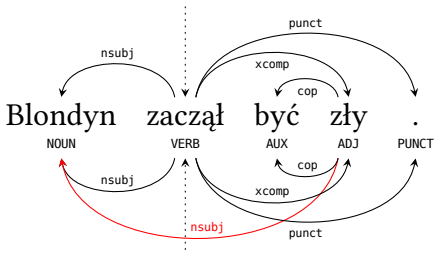


Figure C.5: UD representation of (7.7)

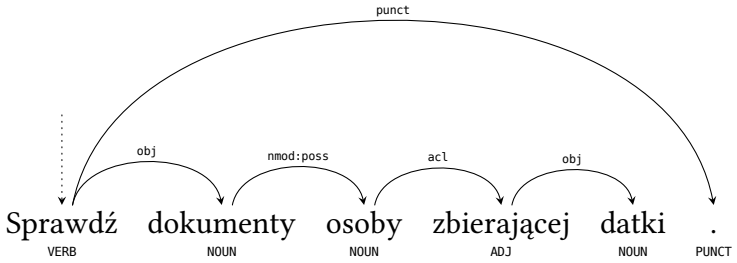


Figure C.6: UD representation of (7.8)

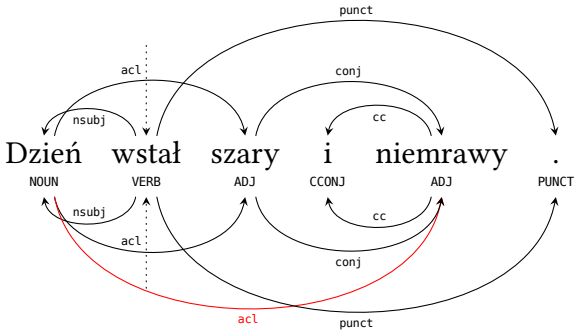


Figure C.7: UD representation of (7.9)

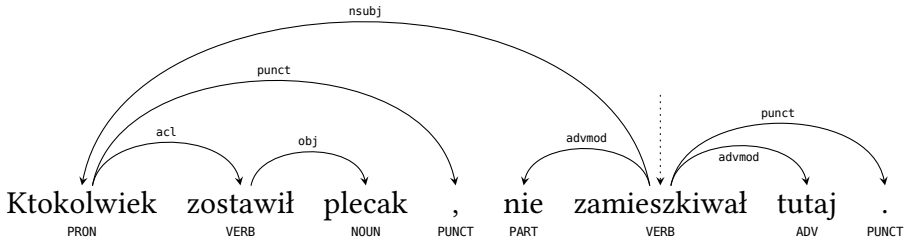


Figure C.8: UD representation of (7.10)

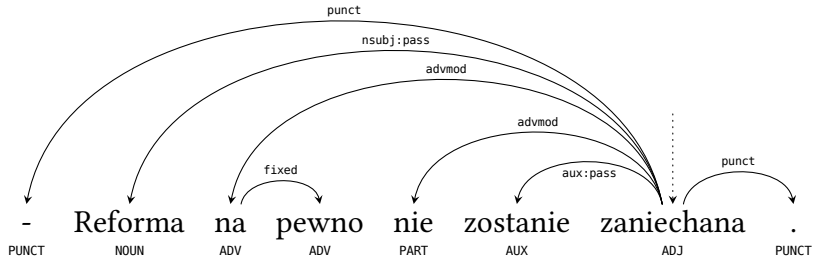


Figure C.9: UD representation of (7.11)

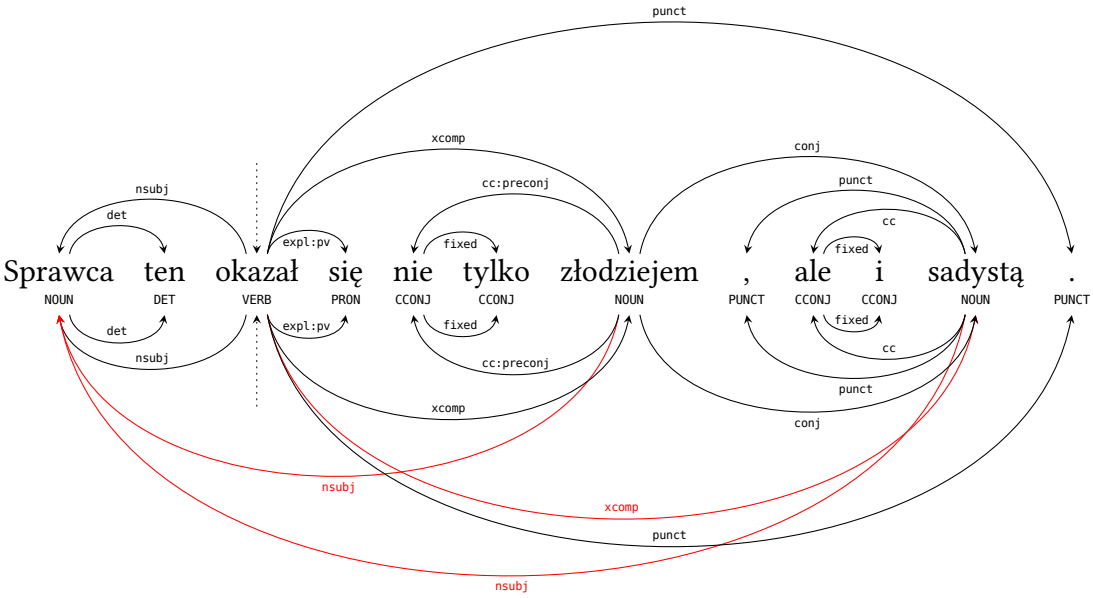


Figure C.10: UD representation of (7.12)

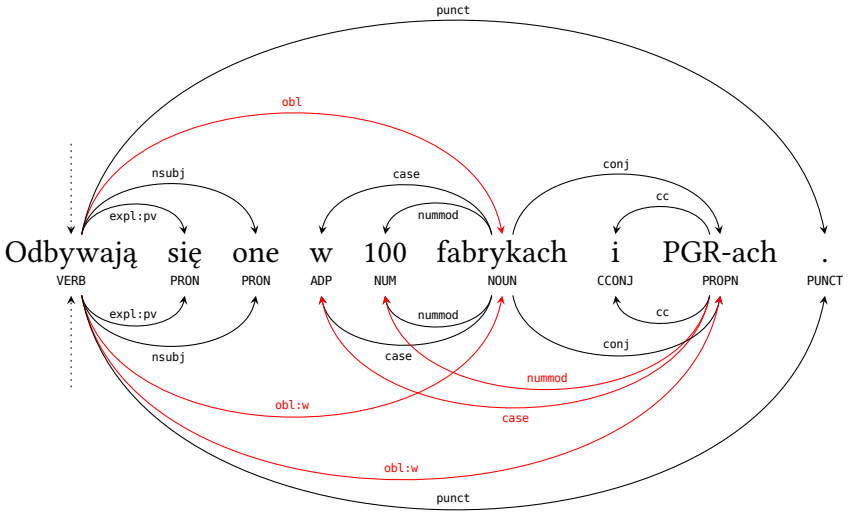


Figure C.11: UD representation of (7.13)

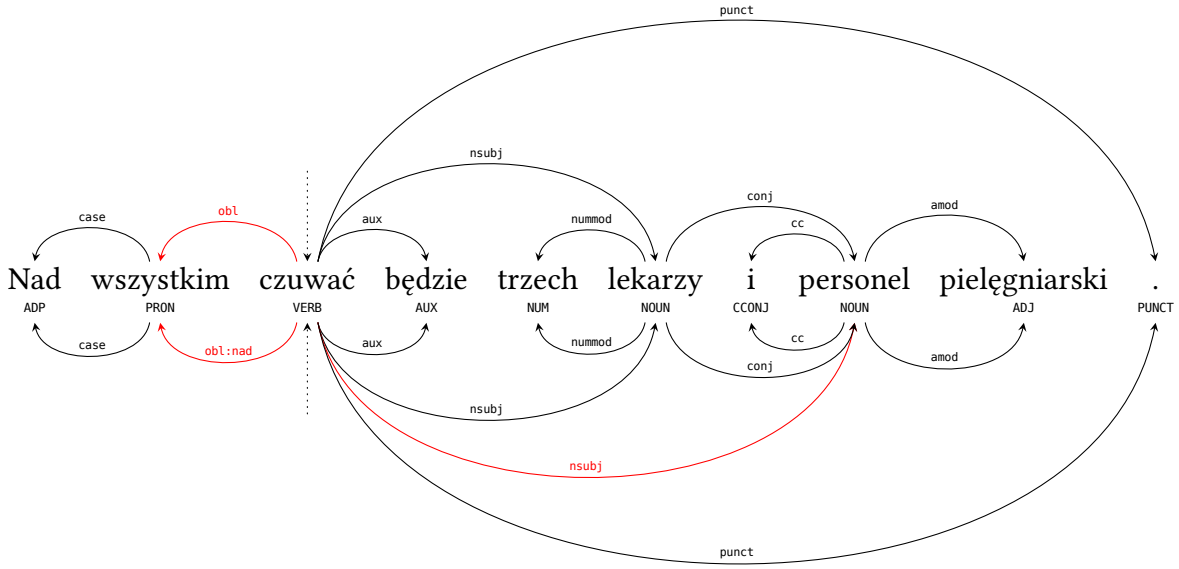


Figure C.12: UD representation of (7.14)

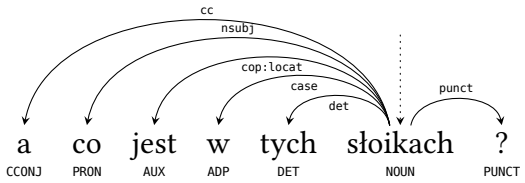


Figure C.13: UD representation of (7.16)

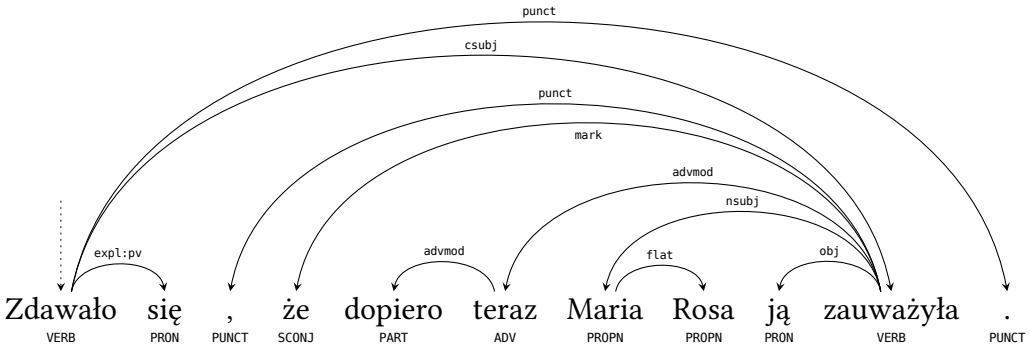


Figure C.14: UD representation of (7.17)

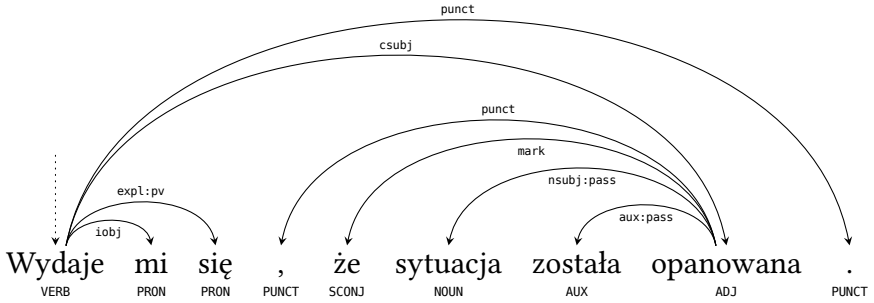


Figure C.15: UD representation of (7.18)

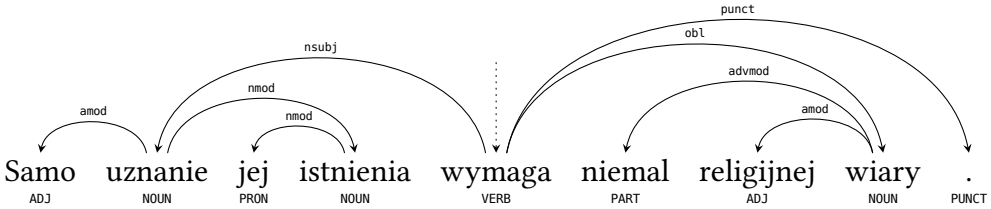


Figure C.16: UD representation of (7.19)

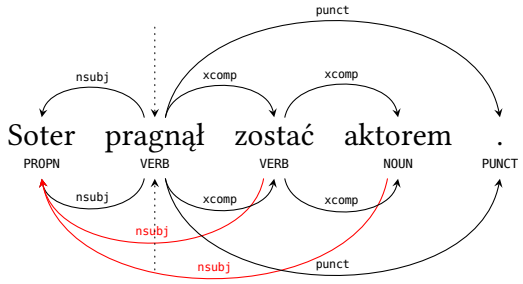


Figure C.17: UD representation of (7.20)

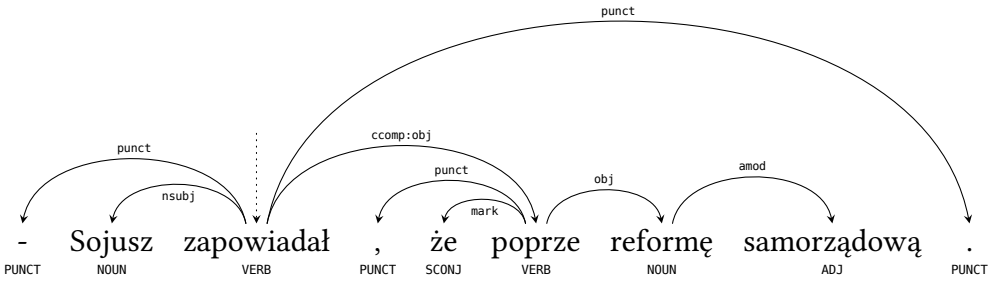


Figure C.18: UD representation of (7.21)

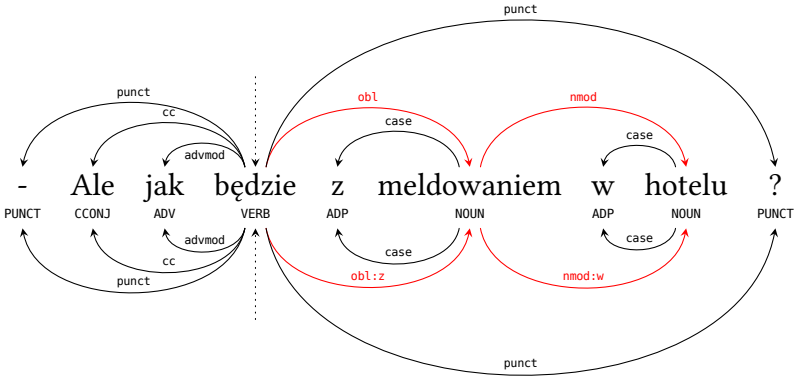


Figure C.19: UD representation of (7.22)

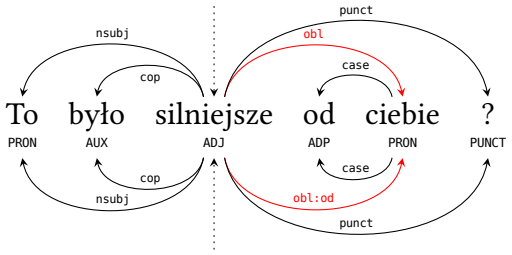


Figure C.20: UD representation of (7.23)

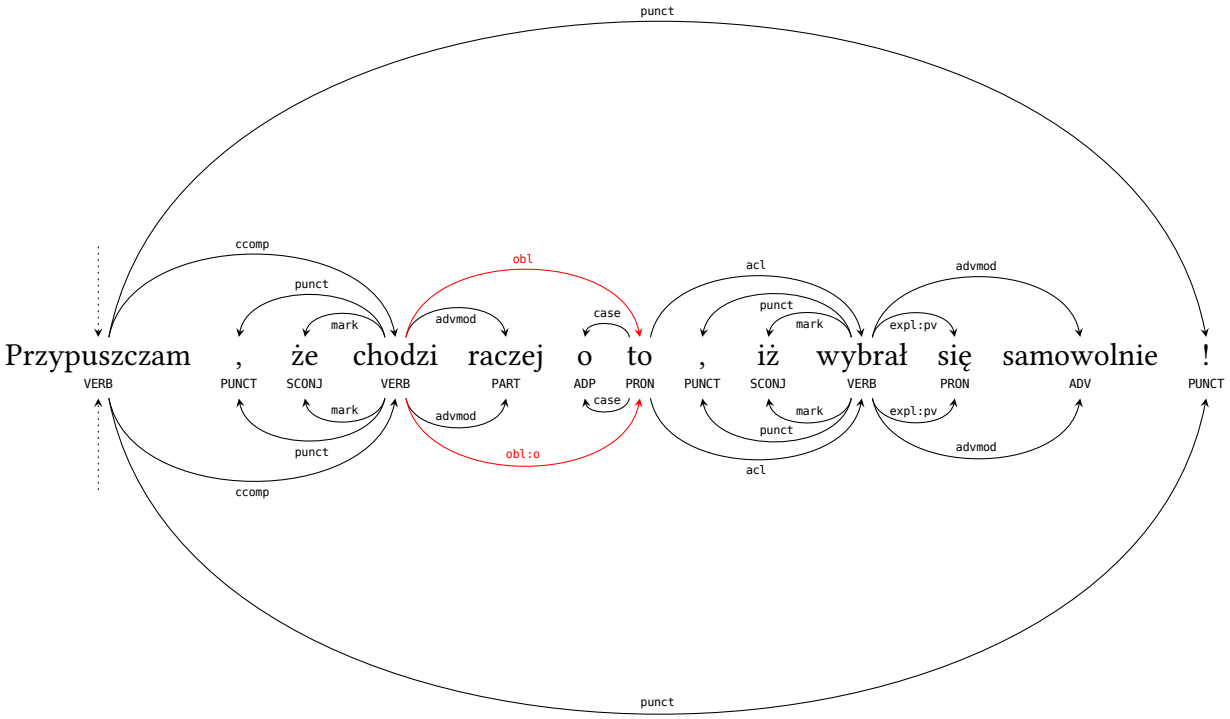


Figure C.21: UD representation of (7.24)

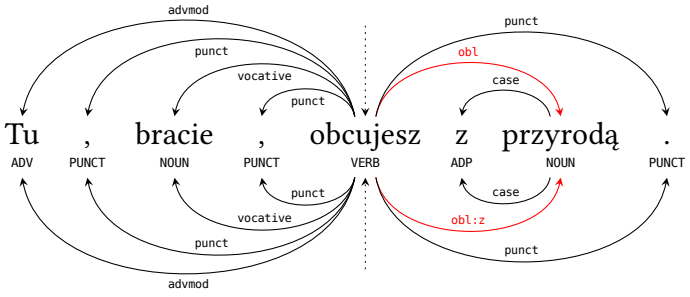


Figure C.22: UD representation of (7.25)

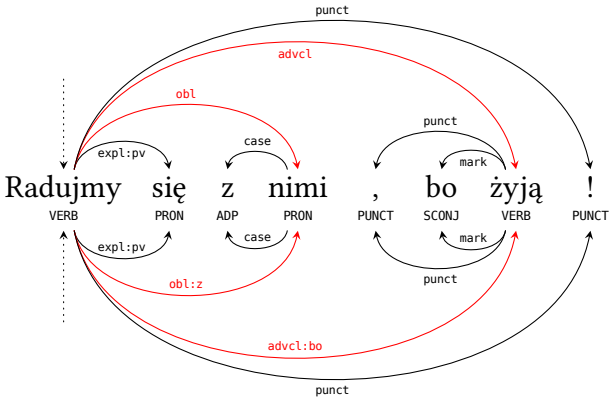


Figure C.23: UD representation of (7.26)

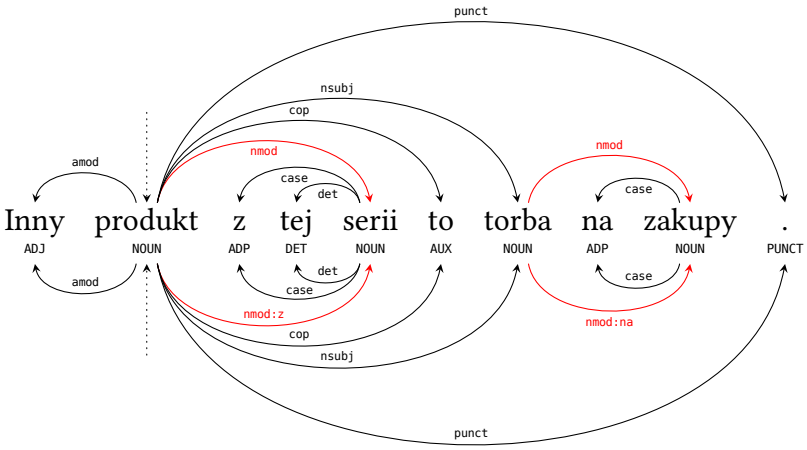


Figure C.24: UD representation of (7.27)

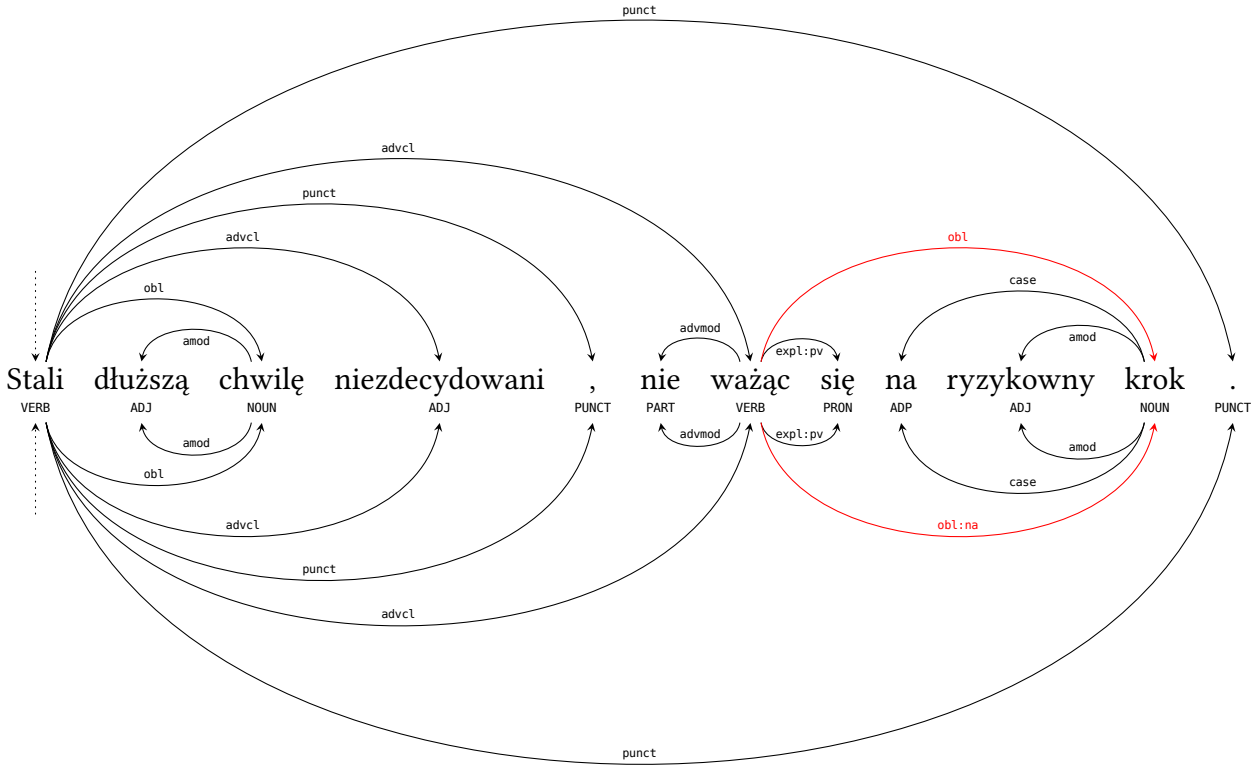


Figure C.25: UD representation of (7.28)

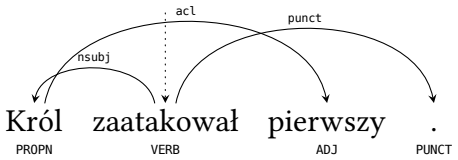


Figure C.26: UD representation of (7.29)

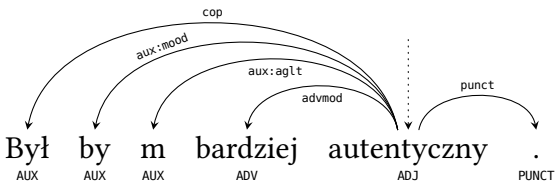


Figure C.27: UD representation of (7.30)

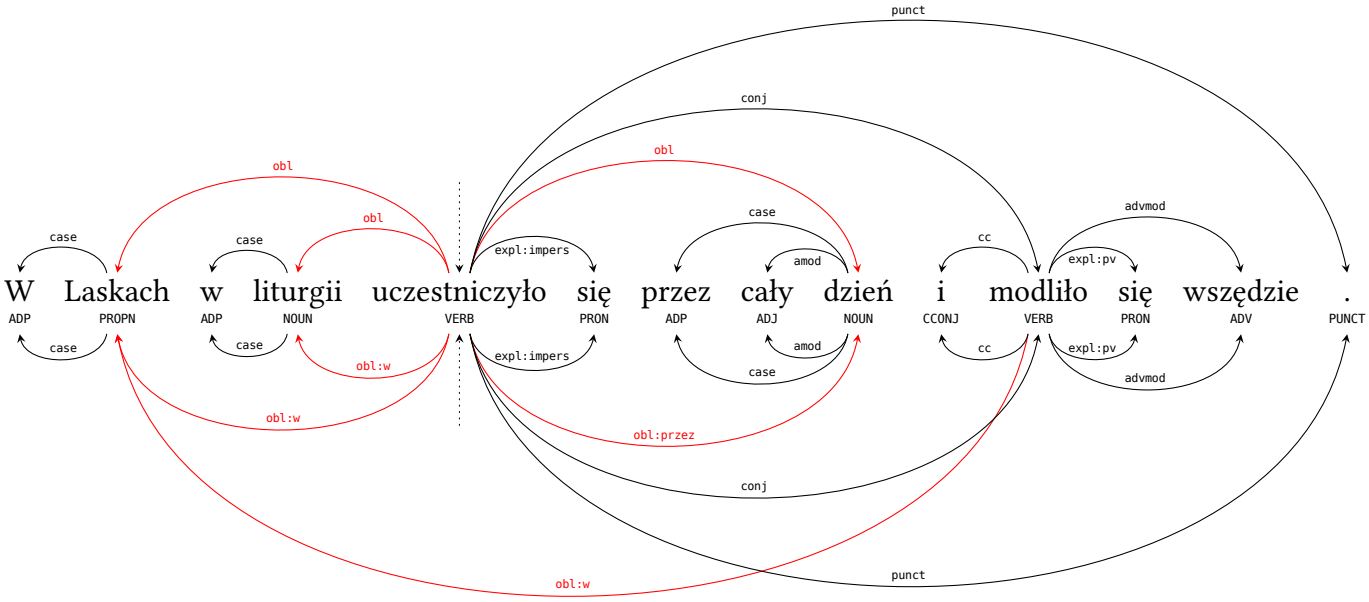


Figure C.28: UD representation of (7.31)

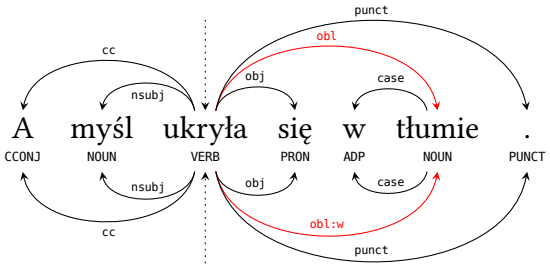


Figure C.29: UD representation of (7.32)

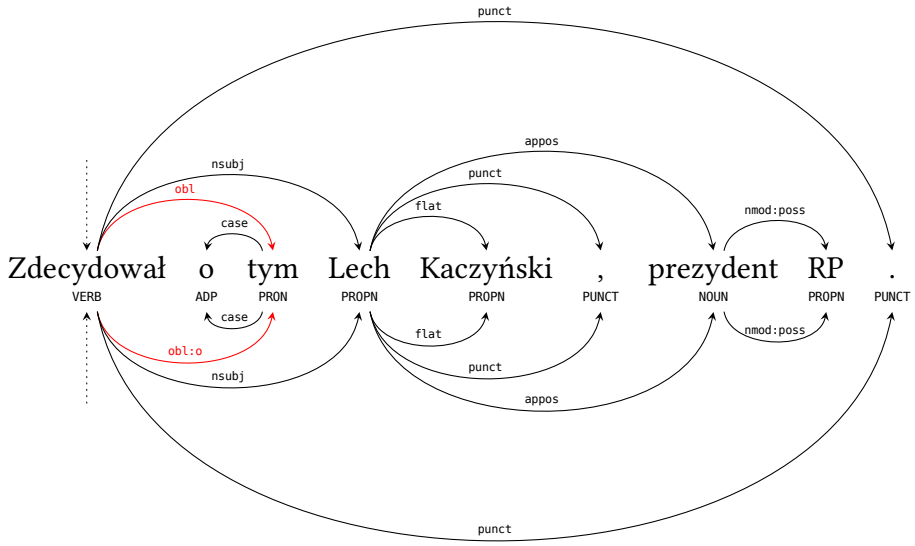


Figure C.30: UD representation of (7.33)

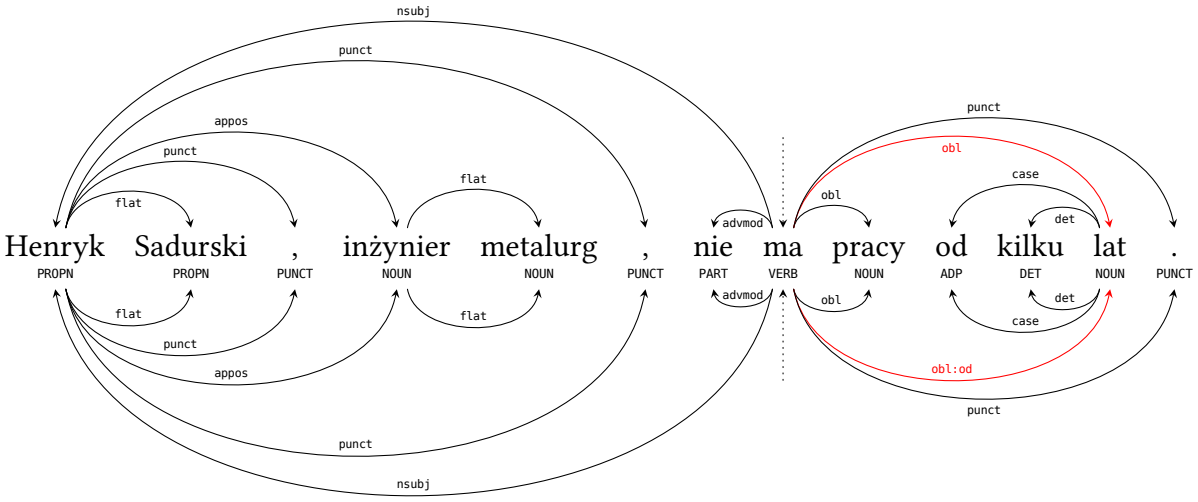


Figure C.31: UD representation of (7.34)

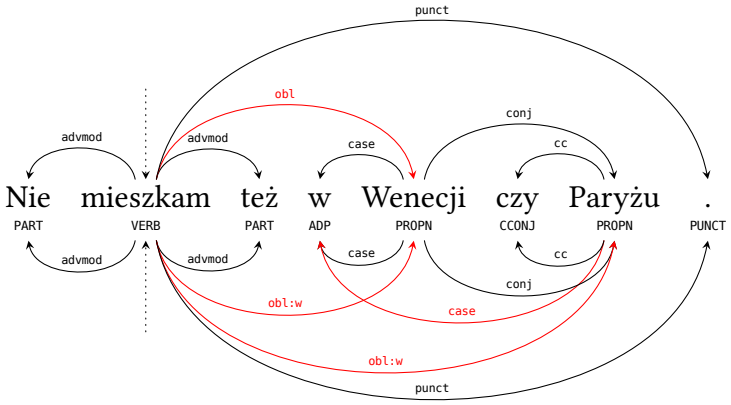


Figure C.32: UD representation of (7.35)

Bibliography

Andrews, Avery D. 2007. "Input and Glue in OT-LFG". In *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan*, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 319–340. Stanford, CA: CSLI Publications.

Arka, I Wayan, and Jane Simpson. 1998. "Control and Complex Arguments in Balinese". In *The Proceedings of the LFG'98 Conference*, ed. by Miriam Butt and Tracy Holloway King. University of Queensland, Brisbane: CSLI Publications.

Babby, Leonard [H.] 1980a. *Existential Sentences and Negation in Russian*. Ann Arbor, MI: Karoma Publishers.

– . 1980b. "The Syntax of Surface Case Marking". In *Cornell Working Papers in Linguistics*, ed. by Wayne Harbert and Julia Herschensohn, 1:1–32. Department of Modern Languages / Linguistics, Cornell University.

Bańko, Mirosław, ed. 2000. *Inny słownik języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Bień, Janusz S., and Zygmunt Saloni. 1982. "Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)". *Prace Filologiczne* XXXI:31–45.

Bień, Janusz S., and Marcin Woliński. 2003. "Wzbogacony korpus *Słownika frekwencyjnego polszczyzny współczesnej*". In *Prace językoznawcze dedykowane Profesor Jadwidze Sambor*, ed. by Jadwiga Linde-Usiekniewicz and Romuald Huszcza, 6–10. Warsaw: Uniwersytet Warszawski, Wydział Polonistyki.

Bondaruk, Anna. 2013. *Copular Clauses in English and Polish*. Lublin: Wydawnictwo KUL.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. "The TIGER Treebank". In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, ed. by Erhard Hinrichs and Kiril Simov. Sozopol.

Bresnan, Joan, ed. 1982. *The Mental Representation of Grammatical Relations*. Cambridge, MA: The MIT Press.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-Functional Syntax*. 2nd. Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Buchholz, Sabine, and Erwin Marsi. 2006. "CoNLL-X shared task on Multilingual Dependency Parsing". In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, 149–164.

Çetinoğlu, Özlem, Jennifer Foster, Joakim Nivre, Deirdre Hogan, Aoife Cahill, and Josef van Genabith. 2010. "LFG without C-structures". In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT 9)*, ed. by Markus Dickinson, Kaili Müürisep, and Marco Passarotti, 43–54. Tartu, Estonia.

Corbett, Greville G., Norman M. Fraser, and Scott McGlashan, eds. 1993. *Heads in Grammatical Theory*. Cambridge: Cambridge University Press.

Croft, William. 1996. "What's a Head?" In *Phrase Structure and the Lexicon*, ed. by Johan Rooryck and Laurie Zaring, 35–75. Dordrecht: Springer.

Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2011. *XLE Documentation*. `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego, CA: Academic Press.

Dalrymple, Mary, John Lowe, and Louise Mycock. 2018. *Lexical Functional Grammar*. Second edition, forthcoming. Oxford University Press.

Danielewiczowa, Magdalena. 2015. "Polskie *sobie* – wyzwanie rzucone lingwiście". In *U prostoru lingvističke slavistike*, ed. by L. Popović, D. Vojvodić, and M. Nomachi, 323–342. Belgrade: Filološki Fakultet Universitetu Beogradu.

Dziwirek, Katarzyna. 1990. "Default Agreement in Polish". In *Grammatical Relations: A Cross-Theoretical Perspective*, ed. by Katarzyna Dziwirek, Patrick Farrell, and Errapel Mejías-Bikandi. Stanford, CA: CSLI Publications.

– . 1994. *Polish Subjects*. New York: Garland.

Franks, Steven. 1995. *Parameters of Slavic Morphosyntax*. New York: Oxford University Press.

Gołąb, Zbigniew, Adam Heinz, and Kazimierz Polański. 1968. *Słownik terminologii językoznawczej*. Wydawnictwo Naukowe PWN.

Hajnicz, Elżbieta, Agnieszka Patejuk, Adam Przepiórkowski, and Marcin Woliński. 2016. "Walenty: słownik walencyjny języka polskiego z bogatym komponentem frazeologicznym". In *Výzkum slovesné valence ve slovanských zemích*, ed. by Karolina Skwarska and Elżbieta Kaczmarska, 71–102. Prague: Slovanský ústav AV ČR.

Hudson, Richard. 1987. "Zwicky on Heads". *Journal of Linguistics* 23 (1): 109–132.

Jaworska, Ewa. 1986a. "Aspects of the Syntax of Prepositions and Prepositional Phrases in English and Polish". PH.D. Thesis, University of Oxford.

– . 1986b. "Prepositional Phrases as Subjects and Objects". *Journal of Linguistics* 22:355–374.

Kallas, Krystyna. 1986. "Syntaktyczna charakterystyka wielofunkcyjnego *JAK*". *Polonica* XII:127–143.

– . 1995. "O konstrukcjach z przyimkiem *niż*". In *Wyrażenia funkcyjne w systemie i tekście*, ed. by Maciej Grochowski, 99–110. Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Kaplan, Ronald M. 1995. "The Formal Architecture of Lexical-Functional Grammar". In *Formal Issues in Lexical-Functional Grammar*, ed. by Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, 7–27. CSLI Lecture Notes 47. Stanford, CA: CSLI Publications.

König, Esther, Wolfgang Lezius, and Holger Voormann. 2003. *TIGERSearch 2.1: User's Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Kupść, Anna. 1999. "Haplology of the Polish Reflexive Marker". In *Slavic in Head-Driven Phrase Structure Grammar*, ed. by Robert D. Borsley and Adam Przepiórkowski, 91–124. Stanford, CA: CSLI Publications.

Kupść, Anna, and Adam Przepiórkowski. 2002. "Morphological Aspects of Verbal Negation in Polish". In *Current Approaches to Formal Slavic Linguistics: Proceedings of the Second European Conference on Formal Description of Slavic Languages, Potsdam, 1997*, ed. by Peter Kosta and Jens Frasek, 337–346. Frankfurt am Main: Peter Lang.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran, and Jerzy Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. Cracow: Wydawnictwo Instytutu Języka Polskiego PAN.

Landau, Idan. 2013. *Control in Generative Grammar: A Research Companion*. Cambridge: Cambridge University Press.

Makowska, Danuta, and Zygmunt Saloni. 2009. "Polska konstrukcja „iść w sołdaty" a kategoria deprecjatywności w języku polskim". *Pamiętnik Literacki C* 100 (1): 145–158.

Mańczak, Witold. 1956. "Ile jest rodzajów w polskim?" *Język Polski* XXXVI (2): 116–121.

Meurer, Paul. 2017. "From LFG Structures to Dependency Relations". In *The Very Model of a Modern Linguist*, ed. by Victoria Rosén and Koenraad De Smedt, 8:183–201. Bergen Language and Linguistics Studies. Bergen: University of Bergen Library. doi:`http://dx.doi.org/10.15845/bells.v8i1`.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1659–1666. Portorož, Slovenia: ELRA, European Language Resources Association (ELRA).

Ogrodniczuk, Maciej. 2003. "Nowa edycja wzbogaconego korpusu słownika frekwencyjnego". In *Językoznawstwo w Polsce. Stan i perspektywy*, ed. by Stanisław Gajda, 181–190. Opole: Komitet Językoznawstwa, Polska Akademia Nauk oraz Instytut Filologii Polskiej, Uniwersytet Opolski.

Øvrelid, Lilja, Jonas Kuhn, and Kathrin Spreyer. 2009. "Cross-Framework Parser Stacking for Data-Driven Dependency Parsing". *TAL* 50 (3): 109–138.

Patejuk, Agnieszka. 2015. "Unlike coordination in Polish: an LFG account". PH.D. dissertation, Institute of Polish Language, Polish Academy of Sciences.

– . 2018. "Incorporating Conjunctions in Polish". In *The Proceedings of the LFG'18 Conference*, ed. by Miriam Butt and Tracy Holloway King. Forthcoming. Stanford, CA: CSLI Publications.

Patejuk, Agnieszka, and Adam Przepiórkowski. 2012. "Towards an LFG Parser for Polish: An Exercise in Parasitic Grammar Development". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, 3849–3852. Istanbul, Turkey: ELRA.

– . 2014a. "In Favour of the Raising Analysis of Passivisation". In *The Proceedings of the LFG'14 Conference*, ed. by Miriam Butt and Tracy Holloway King, 461–481. Stanford, CA: CSLI Publications.

– . 2014b. "Structural Case Assignment to Objects in Polish". In *The Proceedings of the LFG'14 Conference*, ed. by Miriam Butt and Tracy Holloway King, 429–447. Stanford, CA: CSLI Publications.

– . 2014c. "Synergistic Development of Grammatical Resources: A Valence Dictionary, an LFG Grammar, and an LFG Structure Bank for Polish". In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*, ed. by Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, 113–126. Tübingen: Department of Linguistics (SfS), University of Tübingen.

– . 2015a. "An LFG Analysis of the so-called Reflexive Marker in Polish". In *The Proceedings of the LFG'15 Conference*, ed. by Miriam Butt and Tracy Holloway King, 270–288. Stanford, CA: CSLI Publications.

– . 2015b. "Parallel Development of Linguistic Resources: Towards a Structure Bank of Polish". *Prace Filologiczne* LXV:255–270.

– . 2016. "Reducing Grammatical Functions in Lexical Functional Grammar". In *The Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller, 541–559. Stanford, CA: CSLI Publications.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. Chicago, IL: Chicago University Press / CSLI Publications.

Popel, Martin, David Mareček, Jan Štěpánek, Daniel Zeman, and Zděněk Žabokrtský. 2013. "Coordination Structures in Dependency Treebanks". In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 517–527. Sofia, Bulgaria.

Przepiórkowski, Adam. 1999. "Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach". PH.D. dissertation, Universität Tübingen.

– . 2000. "Long Distance Genitive of Negation in Polish". *Journal of Slavic Linguistics* 8:151–189.

– . 2004a. "O wartości przypadka podmiotów liczebnikowych". *Biuletyn Polskiego Towarzystwa Językoznawczego* LX:133–143.

– . 2004b. *The IPI PAN Corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

– . 2009. "A comparison of two morphosyntactic tagsets of Polish". In *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, ed. by Violetta Koseska-Toszewa, Ludmila Dimitrova, and Roman Roszko, 138–144. Warsaw.

– . 2016a. “Against the Argument–Adjunct Distinction in Functional Generative Description”. *The Prague Bulletin of Mathematical Linguistics* 106:5–20.

– . 2016b. “How *not* to Distinguish Arguments from Adjuncts in LFG”. In *The Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, ed. by Doug Arnold, Miriam Butt, Berthold Crysmann, Tracy Holloway King, and Stefan Müller, 560–580. Stanford, CA: CSLI Publications.

– . 2017a. *Argumenty i modyfikatory w gramatyce i w słowniku*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

– . 2017b. “On the Argument–Adjunct Distinction in the Polish *Semantic Syntax* Tradition”. *Cognitive Studies / Études Cognitives* 17:1–10. doi:https://doi.org/10.11649/cs.1344.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, eds. 2012. *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik. 2011. “National Corpus of Polish”. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, ed. by Zygmunt Vetulani, 259–263. Poznań, Poland.

Przepiórkowski, Adam, Elżbieta Hajnicz, Anna Andrzejczuk, Agnieszka Patejuk, and Marcin Woliński. 2017. “Walenty: gruntowny składniowo-semantyczny słownik walencyjny języka polskiego”. *Język Polski* XCVII (1): 30–47.

Przepiórkowski, Adam, Anna Kupść, Małgorzata Marciniak, and Agnieszka Mykowiecka. 2002. *Formalny opis języka polskiego: Teoria i implementacja*. Warsaw: Akademicka Oficyna Wydawnicza EXIT.

Przepiórkowski, Adam, and Grzegorz Murzynowski. 2011. “Manual annotation of the National Corpus of Polish with Anotatornia”. In *Explorations across Languages and Corpora: PALC 2009*, ed. by Stanisław Goźdź-Roszkowski, 95–103. Frankfurt am Main: Peter Lang.

Przepiórkowski, Adam, and Agnieszka Patejuk. 2012a. “On case assignment and the coordination of unlikes: The limits of distributive features”. In *The Proceedings of the LFG’12 Conference*, ed. by Miriam Butt and Tracy Holloway King, 479–489. Stanford, CA: CSLI Publications.

– . 2012b. “The puzzle of case agreement between numeral phrases and predicative adjectives in Polish”. In *The Proceedings of the LFG’12 Conference*, ed. by Miriam Butt and Tracy Holloway King, 490–502. Stanford, CA: CSLI Publications.

– . 2015. “Two Representations of Negation in LFG: Evidence from Polish”. In *The Proceedings of the LFG’15 Conference*, ed. by Miriam Butt and Tracy Holloway King, 322–336. Stanford, CA: CSLI Publications.

– . 2018. “Arguments and Adjuncts in Universal Dependencies”. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe, NM.

Przepiórkowski, Adam, Filip Skwarski, Elżbieta Hajnicz, Agnieszka Patejuk, Marek Świdziński, and Marcin Woliński. 2014. “Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego”. *Polonica* XXXIII:159–178.

Przepiórkowski, Adam, and Marcin Woliński. 2003a. “A Flexemic Tagset for Polish”. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, 33–40. Budapest.

– . 2003b. “The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish”. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, 109–116.

Reinhart, Tanya, and Eric Reuland. 1991. “Anaphors and logophors: an argument structure perspective”. In *Long-Distance Anaphora*, ed. by Jan Koster and Eric Reuland, 283–334. Cambridge: Cambridge University Press.

– . 1993. “Reflexivity”. *Linguistic Inquiry* 24:657–720.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. “An Open Infrastructure for Advanced Treebanking”. In *LREC 2012 META-RESEARCH Workshop on Advanced Treebanking*, 22–29. Istanbul, Turkey: ELRA.

Rosén, Victoria, Paul Meurer, and Koenraad De Smedt. 2007. “Designing and Implementing Discriminants for LFG Grammars”. In *The Proceedings of the LFG’07 Conference*, ed. by Miriam Butt and Tracy Holloway King, 397–417. University of Stanford, California, USA: CSLI Publications.

Rouveret, Alain, and Jean-Roger Vergnaud. 1980. “Specifying Reference to the Subject: French causatives and conditions on representations”. *Linguistic Inquiry* 11 (1): 97–202.

Saloni, Zygmunt. 1974. “Klasyfikacja gramatyczna leksemów polskich”. *Język Polski* LIV:3–13, 93–101.

– . 1976. “Kategoria rodzaju we współczesnym języku polskim”. In *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, ed. by Roman Laskowski, 14:43–78. Prace Instytutu Języka Polskiego. Wrocław: Ossolineum.

Saloni, Zygmunt, and Marek Świdziński. 1985. *Składnia współczesnego języka polskiego*. 2nd (changed). Warsaw: Wydawnictwo Naukowe PWN.

Sulger, Sebastian, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoğlu, I Wayan Arka, and Meladel Mistica. 2013. “ParGramBank: The ParGram Parallel Treebank”. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 550–560. Sofia, Bulgaria: Association for Computational Linguistics.

Świdziński, Marek. 1992. *Gramatyka formalna języka polskiego*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Świdziński, Marek, and Marcin Woliński. 2010. “Towards a Bank of Constituent Parse Trees for Polish”. In *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, ed. by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, 197–204. Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag.

Urbańczyk, Stanisław, ed. 1992. *Encyklopedia języka polskiego*. Wrocław: Ossolineum.

Warren, D. H. D., and Fernando C. N. Pereira. 1980. “Definite Clause Grammars for Language Analysis — A Survey of the Formalism and a Comparison with Augmented Transition Networks”. *Artificial Intelligence* 13:231–278.

Witkoś, Jacek. 1993. “Some Aspects of Phrasal Movement in English and Polish”. PH.D. Thesis, Adam Mickiewicz University.

– . 1995. “Wh-Extraction from Clausal Complements in Polish: A Minimality/Locality Account”. *Folia Linguistica* XXIX (3/4): 223–264.

Woliński, Marcin. 2004. “Komputerowa weryfikacja gramatyki Świdzińskiego”. PH.D. dissertation, Institute of Computer Science, Polish Academy of Sciences.

– . 2006. “Morfeusz — a Practical Tool for the Morphological Analysis of Polish”. In *Intelligent Information Processing and Web Mining*, ed. by Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, 503–512. Advances in Soft Computing. Berlin: Springer-Verlag.

– . 2014. “Morfeusz Reloaded”. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1106–1111. Reykjavík, Iceland: ELRA.

Woliński, Marcin, Katarzyna Głowińska, and Marek Świdziński. 2011. “A Preliminary Version of Składnica—a Treebank of Polish”. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, ed. by Zygmunt Vetulani, 299–303. Poznań, Poland.

Woliński, Marcin, and Adam Przepiórkowski. 2001. *Projekt anotacji morfosynktaktycznej korpusu języka polskiego*. IPI PAN Research Report 938. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Wróblewska, Alina. 2014. “Polish Dependency Parser Trained on an Automatically Induced Dependency Bank”. PH.D. dissertation, Institute of Computer Science, Polish Academy of Sciences.

Zabrocki, Tadeusz. 1981. *Lexical Rules of Semantic Interpretation: Control and NP Movement in English and Polish*. Filologia Angielska 14. Poznań: Adam Mickiewicz University.

Zeman, Daniel. 2017. “Core Arguments in Universal Dependencies”. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing 2017)*, ed. by Simonetta Montemagni and Joakim Nivre, 287–296. Pisa, Italy.

Zeman, Daniel, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. “HamleDT: Harmonized Multi-Language Dependency Treebank”. *Language Resources and Evaluation* 48 (4): 601–637.

Zwicky, Arnold M. 1985. “Heads”. *Journal of Linguistics* 21 (1): 1–29.