From Lexical Functional Grammar to Enhanced Universal Dependencies

Adam Przepiórkowski
Institute of Philosophy
University of Warsaw *and*
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
adamp@ipipan.waw.pl

Agnieszka Patejuk
Faculty of Linguistics, Philology and Phonetics
University of Oxford *and*
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
aep@ipipan.waw.pl

Universal Dependencies (UD; Nivre et al. 2016) has recently become a *de facto* standard as a dependency representation used in Natural Language Processing (NLP). As perhaps most of syntactic processing in NLP involves dependency structures, it is safe to say that it is becoming a standard for syntactic processing at large. There are 122 treebanks for 71 languages in the July 2018 release 2.2 of UD, publicly available at http://universaldependencies.org/. New UD treebanks are often the result of converting corpora adhering to other annotation schemes – not only dependency-based, but also constituency-based.

Lexical Functional Grammar (LFG; Bresnan 1982, Dalrymple 2001, Bresnan et al. 2015) 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), in which two asyndetically coordinated verbs within a clausal subject share a number of dependents, the c-structure is given in (2) and the f-structure – in (3):[1]

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

The first aim of this paper is to describe a procedure of converting such LFG structures to dependency representations following the UD standard, specifically, its enhanced version 2. Conversion of LFG structures to dependency structures is not a new task, but – with the 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.

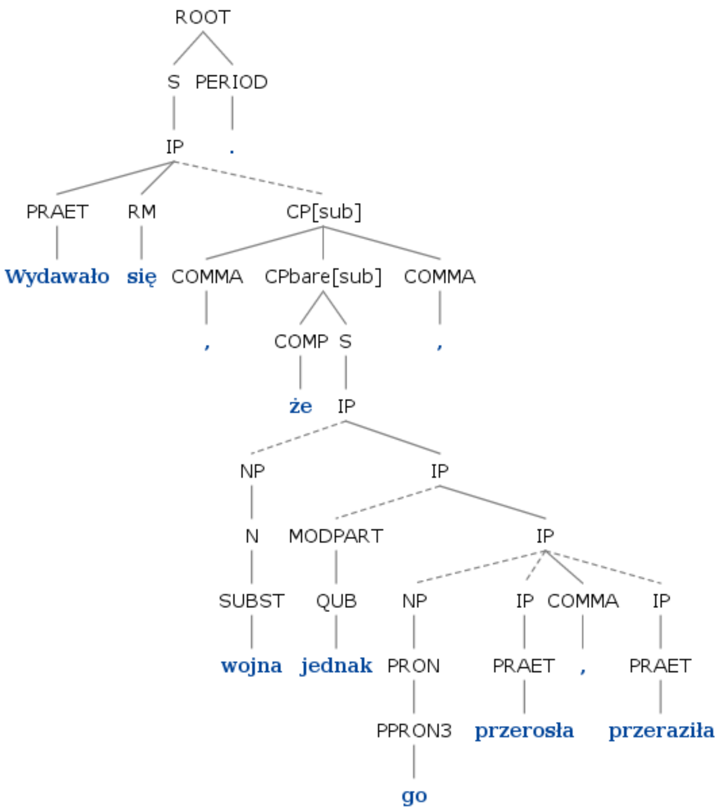
The current version 2 of Universal Dependencies assumes, apart from basic dependency trees, also *enhanced dependency structures*, which make it possible to represent phenomena beyond the scope of simple trees. For example, the result of converting the LFG structures (2)–(3) to UD is shown in (4) (with the basic tree displayed above the text and the enhanced structure – below the text, with the differences shown in red). The second aim of this paper is to examine to what extent rich information available in LFG structures is or may in principle be preserved in such enhanced UD representations.

The empirical basis for the conversion is a manually disambiguated LFG parsebank of Polish (Patejuk and Przepiórkowski 2014) consisting of over 17,000 sentences (almost 131,000 tokens). Since this is a parsebank, it only contains analyses successfully provided by the LFG parser of Polish (Patejuk and Przepiórkowski 2012b, 2015) and selected by human annotators as correct. While this constrains the number and kinds of constructions present in the corpus, the underlying LFG grammar of Polish is currently one of the largest implemented LFG grammars, and it includes a comprehensive analysis of various kinds of coordination and its interaction with other phenomena (Patejuk and Przepiórkowski 2012a), so there is no shortage of sentences which pose potential difficulties for the conversion.

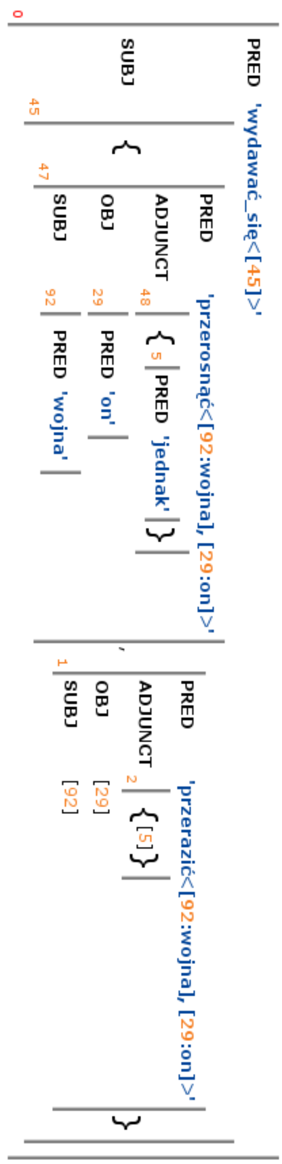
[1] RM in (1) stands for 'reflexive marker', which in this case is an inherent part of the verb *wydawało się* 'seemed'; other abbreviations are standard. LFG structures shown in (2)–(3) are visualisations produced by the INESS system (http://clarino.uib.no/iness/; Rosén et al. 2012), which hosts the Polish LFG structure bank, among other treebanks.

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 2–4
Santa Fe, New Mexico, USA, August 25-26, 2018.

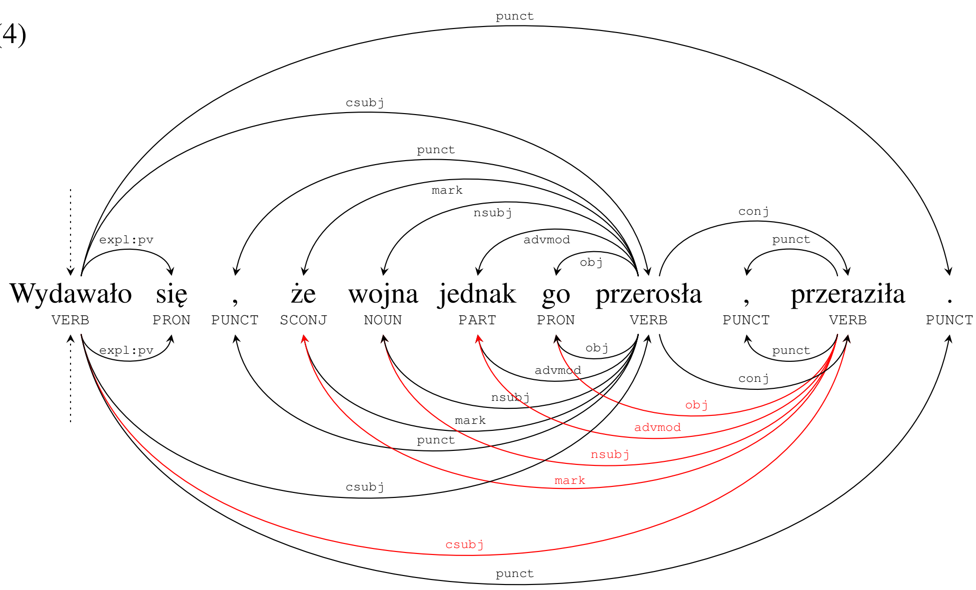
(2)


ROOT
S PERIOD
IP .
PRAET RM CP[sub]
Wydawało się COMMA CPbare[sub] COMMA
, ,
COMP S
że IP
NP IP
N MODPART IP
SUBST QUB NP IP COMMA IP
wojna jednak PRON PRAET , PRAET
PPRON3 przerosła przeraziła
go

(3)


PRED 'wydawać_się<[45]>'
SUBJ
PRED 'przerosnąć<[92:wojna], [29:on]>'
ADJUNCT PRED 'jednak'
OBJ PRED 'on'
SUBJ PRED 'wojna'
PRED 'przerazić<[92:wojna], [29:on]>'
ADJUNCT [5]
OBJ [29]
SUBJ [92]

(4)


Wydawało się , że wojna jednak go przerosła , przeraziła .
VERB PRON PUNCT SCONJ NOUN PART PRON VERB PUNCT VERB PUNCT
punct
csubj
expl:pv
mark
nsubj
advmod
obj
conj

References

Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Wiley-Blackwell, 2nd edition.

Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA.

Mary Dalrymple. 2001. *Lexical Functional Grammar*. Academic Press, San Diego, CA.

Paul Meurer. 2017. From LFG structures to dependency relations. In Victoria Rosén and Koenraad De Smedt, editors, *The Very Model of a Modern Linguist*, volume 8 of *Bergen Language and Linguistics Studies*, pages 183–201. University of Bergen Library, Bergen.

Joakim Nivre, 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 Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 1659–1666, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Agnieszka Patejuk and Adam Przepiórkowski. 2012a. A comprehensive analysis of constituent coordination for grammar engineering. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2191–2207, Mumbai, India.

Agnieszka Patejuk and Adam Przepiórkowski. 2012b. 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*, pages 3849–3852, Istanbul, Turkey. ELRA.

Agnieszka Patejuk and Adam Przepiórkowski. 2014. Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*, pages 113–126, Tübingen. Department of Linguistics (SfS), University of Tübingen.

Agnieszka Patejuk and Adam Przepiórkowski. 2015. Parallel development of linguistic resources: Towards a structure bank of Polish. *Prace Filologiczne*, LXV:255–270.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *LREC 2012 META-RESEARCH Workshop on Advanced Treebanking*, pages 22–29, Istanbul, Turkey. ELRA.