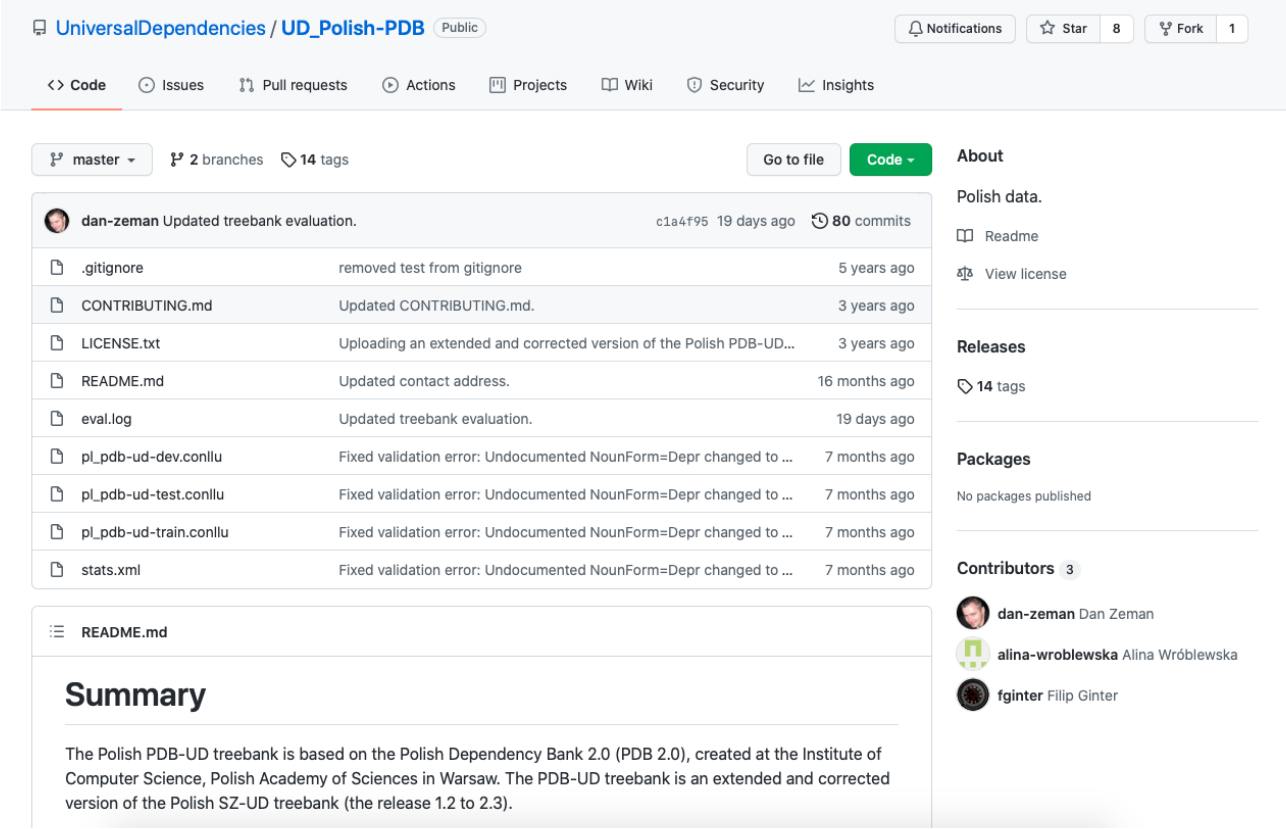
Universal Dependencies is an international initiative for developing a cross-linguistically consistent tree annotation schema and building a large multilingual collection of dependency treebanks annotated according to this schema. The largest Polish treebank in this collection is the UD_Polish-PDB (Wróblewska, 2018) treebank with 22K dependency graphs and 352K tokens. A large part of the sentences underlying the PDB-UD trees, i.e. 14K sentences, comes from NKJP. In addition to the morphosyntactic UD annotations, PDB-UD includes an annotation layer with Polish-specific part-of-speech tags that are derived from NKJP. It is also worth emphasising that the segmentation of PDB-UD is also consistent with the segmentation scheme developed for NKJP.



dan-zeman Updated treebank evaluation. c1a4f95 19 days ago 80 commits

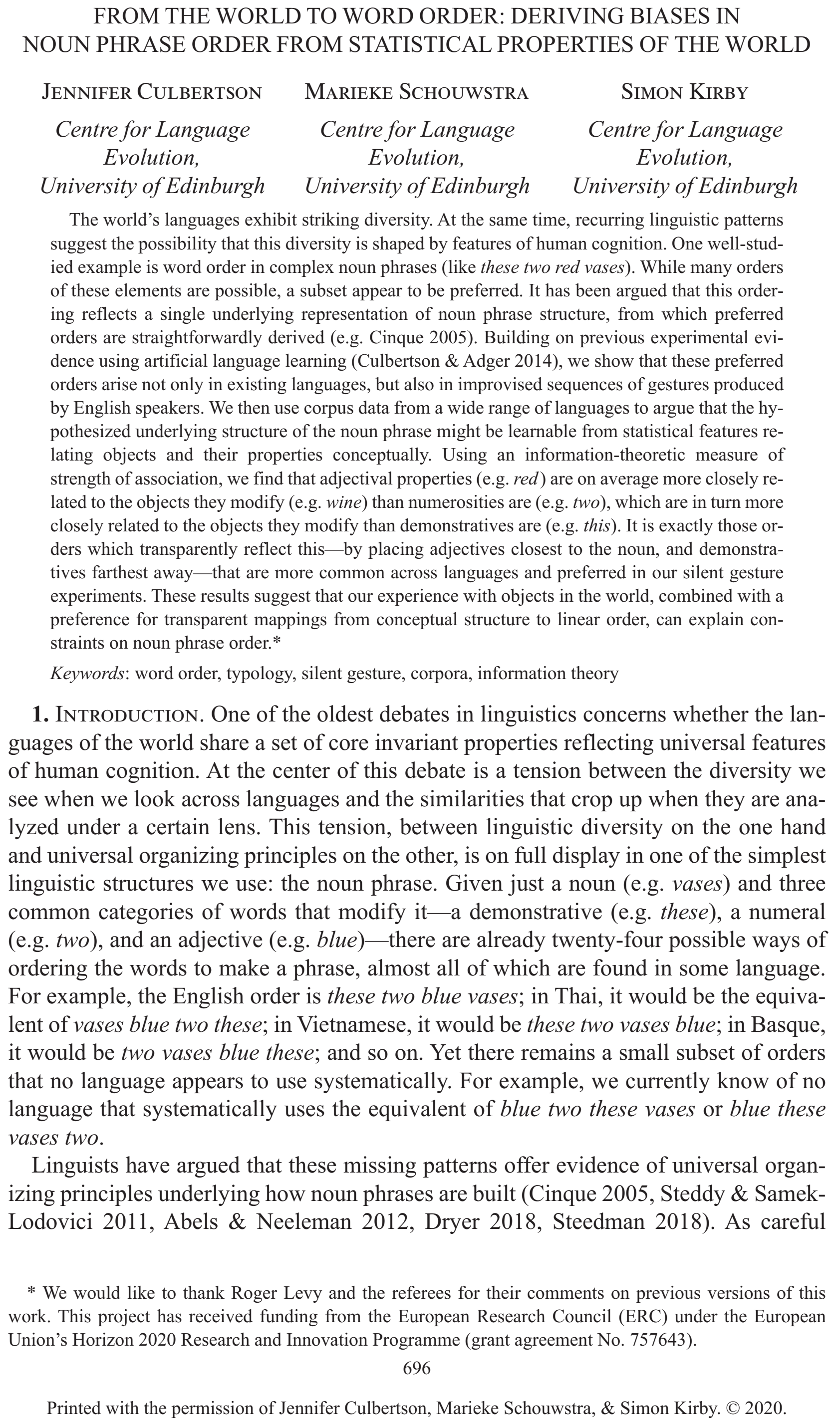
.gitignore	removed test from gitignore	5 years ago
CONTRIBUTING.md	Updated CONTRIBUTING.md.	3 years ago
LICENSE.txt	Uploading an extended and corrected version of the Polish PDB-UD...	3 years ago
README.md	Updated contact address.	16 months ago
eval.log	Updated treebank evaluation.	19 days ago
pl_pdb-ud-dev.conllu	Fixed validation error: Undocumented NounForm=Depr changed to ...	7 months ago
pl_pdb-ud-test.conllu	Fixed validation error: Undocumented NounForm=Depr changed to ...	7 months ago
pl_pdb-ud-train.conllu	Fixed validation error: Undocumented NounForm=Depr changed to ...	7 months ago
stats.xml	Fixed validation error: Undocumented NounForm=Depr changed to ...	7 months ago

Summary

The Polish PDB-UD treebank is based on the Polish Dependency Bank 2.0 (PDB 2.0), created at the Institute of Computer Science, Polish Academy of Sciences in Warsaw. The PDB-UD treebank is an extended and corrected version of the Polish SZ-UD treebank (the release 1.2 to 2.3).

The second largest Polish treebank in the UD repository – UD_Polish-LFG (Patejuk and Przepiórkowski, 2018) – contains 17,000 dependency graphs (130,000 segments), which are exclusively based on sentences derived from NKJP.

Polish dependency treebanks from the UD repository are used in scientific research. In linguistic-cognitive research (Culbertson et al., 2020) published in the prestigious journal Language, the authors checked whether the structure of a noun phrase is determined by the properties of objects existing in the real world and whether hypothetical structures of noun phrases can be learned from the statistical features of these objects.



FROM THE WORLD TO WORD ORDER: DERIVING BIASES IN NOUN PHRASE ORDER FROM STATISTICAL PROPERTIES OF THE WORLD

JENNIFER CULBERTSON — *Centre for Language Evolution, University of Edinburgh*

MARIEKE SCHOUWSTRA — *Centre for Language Evolution, University of Edinburgh*

SIMON KIRBY — *Centre for Language Evolution, University of Edinburgh*

The world's languages exhibit striking diversity. At the same time, recurring linguistic patterns suggest the possibility that this diversity is shaped by features of human cognition. One well-studied example is word order in complex noun phrases (like *these two red vases*). While many orders of these elements are possible, a subset appear to be preferred. It has been argued that this ordering reflects a single underlying representation of noun phrase structure, from which preferred orders are straightforwardly derived (e.g. Cinque 2005). Building on previous experimental evidence using artificial language learning (Culbertson & Adger 2014), we show that these preferred orders arise not only in existing languages, but also in improvised sequences of gestures produced by English speakers. We then use corpus data from a wide range of languages to argue that the hypothesized underlying structure of the noun phrase might be learnable from statistical features relating objects and their properties conceptually. Using an information-theoretic measure of strength of association, we find that adjectival properties (e.g. *red*) are on average more closely related to the objects they modify (e.g. *wine*) than numerosities are (e.g. *two*), which are in turn more closely related to the objects they modify than demonstratives are (e.g. *this*). It is exactly those orders which transparently reflect this—by placing adjectives closest to the noun, and demonstratives farthest away—that are more common across languages and preferred in our silent gesture experiments. These results suggest that our experience with objects in the world, combined with a preference for transparent mappings from conceptual structure to linear order, can explain constraints on noun phrase order.*

Keywords: word order, typology, silent gesture, corpora, information theory

1. INTRODUCTION. One of the oldest debates in linguistics concerns whether the languages of the world share a set of core invariant properties reflecting universal features of human cognition. At the center of this debate is a tension between the diversity we see when we look across languages and the similarities that crop up when they are analyzed under a certain lens. This tension, between linguistic diversity on the one hand and universal organizing principles on the other, is on full display in one of the simplest linguistic structures we use: the noun phrase. Given just a noun (e.g. *vases*) and three common categories of words that modify it—a demonstrative (e.g. *these*), a numeral (e.g. *two*), and an adjective (e.g. *blue*)—there are already twenty-four possible ways of ordering the words to make a phrase, almost all of which are found in some language. For example, the English order is *these two blue vases*; in Thai, it would be the equivalent of *vases blue two these*; in Vietnamese, it would be *these two vases blue*; in Basque, it would be *two vases blue these*; and so on. Yet there remains a small subset of orders that no language appears to use systematically. For example, we currently know of no language that systematically uses the equivalent of *blue two these vases* or *blue these vases two*.

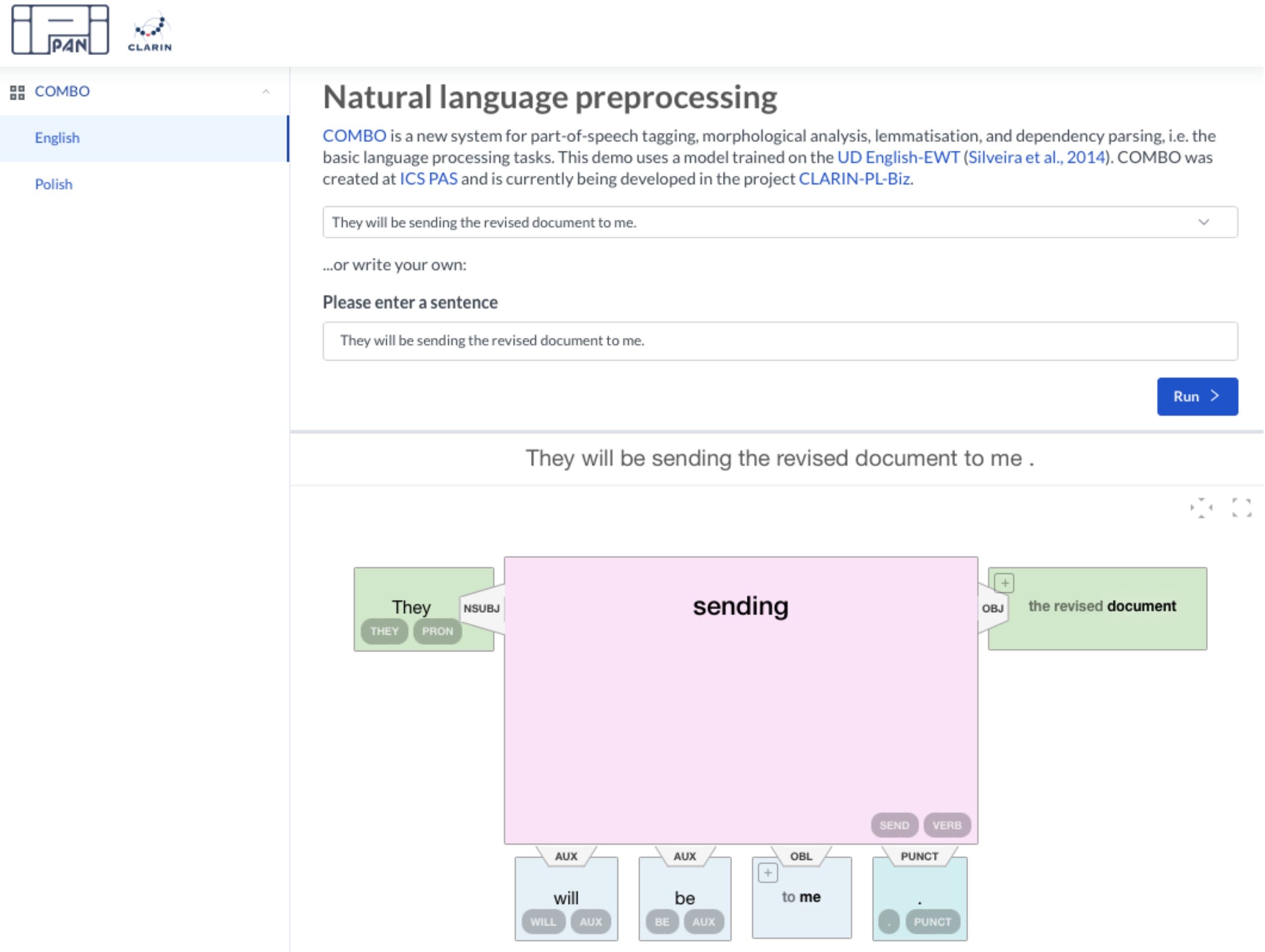
Linguists have argued that these missing patterns offer evidence of universal organizing principles underlying how noun phrases are built (Cinque 2005, Steddy & Samek-Lodovici 2011, Abels & Neeleman 2012, Dryer 2018, Steedman 2018). As careful

* We would like to thank Roger Levy and the referees for their comments on previous versions of this work. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 757643).

In the fields of natural language processing and computational linguistics, dependency treebanks are mainly used to train models that enable morphosyntactic analysis of natural languages. These models are used in state-of-the-art natural language preprocessing systems:

- COMBO (Klimaszewski and Wróblewska, 2021): *polish-herbert-large*|*polish-herbert-base*,
- spaCy (Honnibal et al., 2020): *pl_core_news_sm*|*md*|*lg*,
- spaCy v3: *pl_core_news_sm*|*md*|*lg*,
- UDPipe (Straka, 2018): *polish-pdb-ud-2.5-191206*,
- Stanza (Qi et al., 2020): *pl*.

COMBO-based morphosyntactic models for Polish (trained on UD_Polish-PDB) and English (trained on UD_Polish-EWT) are publicly available for research purposes in the web application.



Natural language preprocessing

COMBO is a new system for part-of-speech tagging, morphological analysis, lemmatisation, and dependency parsing, i.e. the basic language processing tasks. This demo uses a model trained on the UD English-EWT (Silveira et al., 2014). COMBO was created at ICS PAS and is currently being developed in the project CLARIN-PL-Biz.

They will be sending the revised document to me.

...or write your own:

Please enter a sentence

They will be sending the revised document to me.

Run >

They will be sending the revised document to me .