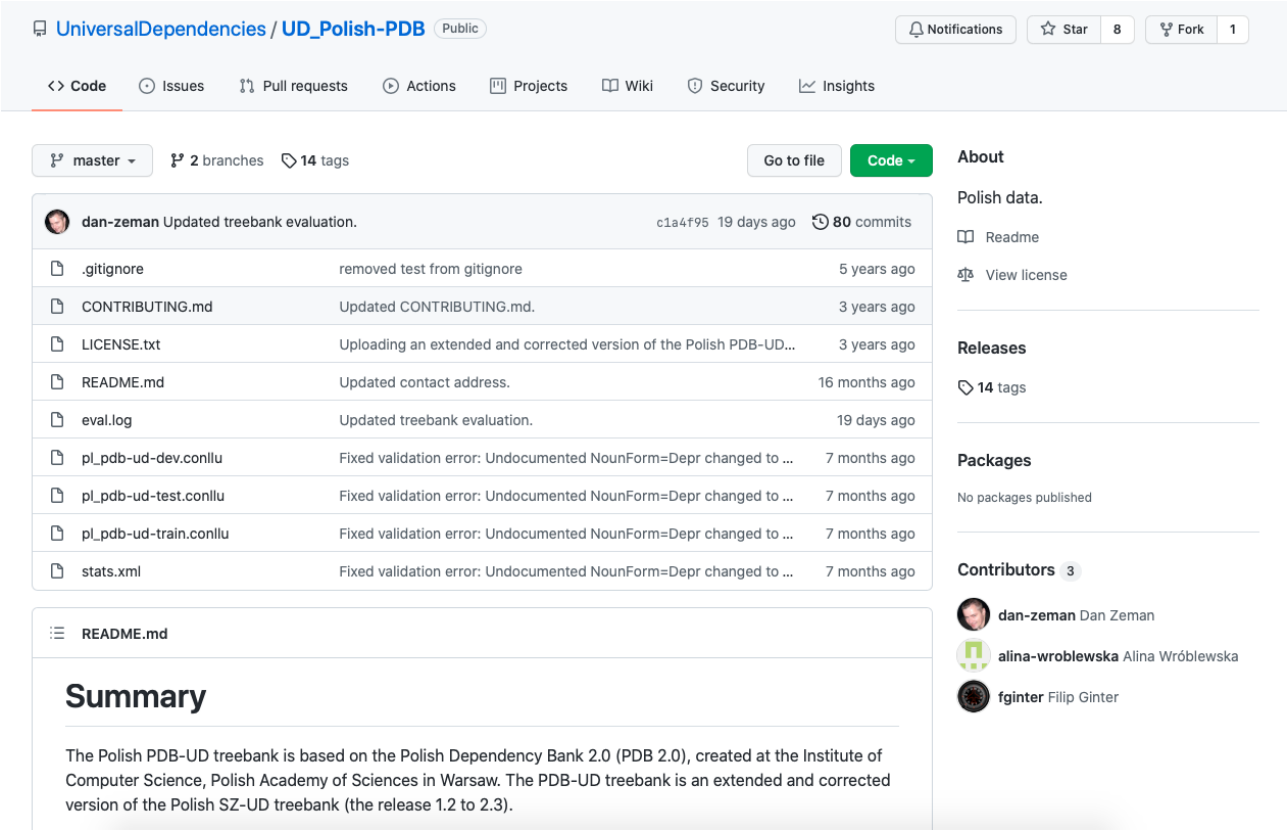
Universal Dependencies to międzynarodowa inicjatywa mająca na celu opracowanie uniwersalnego schematu anotacji drzew zależnościowych i zbudowanie ogromnej wielojęzycznej kolekcji banków drzew zależnościowych anotowanych zgodnie z tym schematem. Największym polskim bankiem drzew w tej kolekcji jest UD_Polish-PDB (Wróblewska, 2018) zawierający 22 tys. grafów zależności i 352 tys. segmentów. Duża część zdań leżących u podstaw banku drzew, tj. 14 tys. zdań, pochodzi z NKJP. Oprócz morfosyntaktycznych anotacji zgodnych ze schematem UD, UD_Polish-PDB zawiera warstwę anotacji ze specyficznymi dla języka polskiego tagami wywodzącymi się z NKJP. Warto również podkreślić, że segmentacja całego banku PDB-UD jest również zbieżna ze schematem segmentacji opracowanym na potrzeby NKJP.



Drugi pod względem wielkości polski bank drzew w repozytorium UD – UD_Polish-LFG (Patejuk i Przepiórkowski, 2018) – zawiera 17 tyś. grafów zależnościowych (130 tys. segmentów), których podstawę stanową wyłącznie zdania pochodzące z NKJP.

Polskie banki drzew zależnościowych z repozytorium UD są wykorzystywane w badaniach naukowych. W badaniach Culbertson i in. (2020) z pogranicza lingwistyki i kognitywistyki opublikowanych w prestiżowym czasopiśmie Language, autorzy sprawdzali, czy struktura frazy rzeczownikowej jest determinowana przez własności obiektów istniejących w świecie rzeczywistym oraz czy na podstawie cech statystycznych tych obiektów można się nauczyć hipotetycznych struktur fraz rzeczownikowych.

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).

W dziedzinach lingwistyki obliczeniowej i przetwarzania języka naturalnego banki drzew są wykorzystywane głównie do trenowania modeli umożliwiających analizę morfosyntaktyczną danych w języku naturalnym dla współczesnych systemów wstępnego przetwarzania języka naturalnego. Przykładowo UD_Polish-PDB został wykorzystany do wytrenowania modeli dla systemów NLP:

- COMBO (Klimaszewski i Wróblewska, 2021): *polish-herbert-large|polish-herbert-base*,
- spaCy (Honnibal i in., 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 i in., 2020): *pl*.

Model COMBO dla polskiego wytrenowany na UD_Polish-PDB został udostępniony publicznie do celów badawczych w aplikacji webowej.