Tweaking UD annotations to investigate the placement of determiners, quantifiers and numerals in the noun phrase

Luigi Talamo

(luigi.talamo@uni-saarland.de)

SIGTYP2022 — JULY,14th

UNIVERSITÄT DES SAARLANDES

Introduction

- Most of the work using Universal Dependencies to study variation across world languages uses **curated collections of annotated texts**, or ‘UD Treebanks’.
- When we turn to **automatically parsed texts**, such as Leipzig corpora in Levshina (2019) or CIEP+ in Talamo and Verkerk (2022), some problems (mostly: **wrong annotations**) arise.
- Can we achieve a **decent quality of analysis** by using automatically parsed texts?

Spoiler: with some tweaks, **we can**.

UD Treebanks

- “Dependency corpora of the HamleDT 2.0 and Universal Dependencies 1.00” (Futrell et al. 2015)
- “the Universal Dependencies Treebank version 2.2” (Naranjo and Becker 2018)
- “a selection of 55 treebanks from Universal Dependencies v2.4” (Yu et al. 2019)
- “Surface-Syntactic Universal Dependencies (SUD) [treebanks]” (Gerdes et al., 2019)
- “Universal Dependencies project, release 2.1” (Futrell et al., 2020).

UD Treebanks are fine for quantitative research, as the quality of linguistic annotation is very high.

However, UD Treebanks **dramatically differ for size and content:** how can we compare, for instance, **Hungarian (42K tokens, 1 treebank, news)** with **French (1,2K tokens, 8 treebanks, 8 different genres)**?

Parallel corpora: CIEP+

- Since late 2019, Annemarie Verkerk and I have been working on **CIEP+**, a parallel **Corpus of Indo-European Prose and More.**
- The corpus has been currently parsed using **Stanford Stanza 1.3 (Qi et al. 2020)** plus **UD Models 2.8 (de Marneffe et al. 2021).**
- This short paper is based on a sample of 10 languages belonging to the **Western branches of the IE family**: Balto-Slavic, Celtic, Germanic, Hellenic and Romance.
- All languages except one (Irish) feature **120K parallel sentences** (1M of tokens), for a total of **18 different texts**.

Order of determiners, quantifiers and numerals in the NP

We are concerned here with the order of **three 'minor' word categories** in the **noun phrase,** which is measured using Shannon's entropy.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Determiners, quantifiers and numerals are often confused in the traditional grammatical analysis and changeling from a cross-linguistic perspective:

- **Determiners** is a **macro-category** containing **articles** (where available) **and demonstratives**.
- Quantifiers are treated in several grammars as **a sub-set** of either **determiners, pronouns** or even numerals.

This is somewhat reflected in the UD annotations:

- at the syntactic level (**UD Relations**), **determiners and quantifiers** are lumped into the det Relation;
- at the word category level (**UPOS**), determiners are pos-tagged as DET and quantifiers as either DET or PRON.

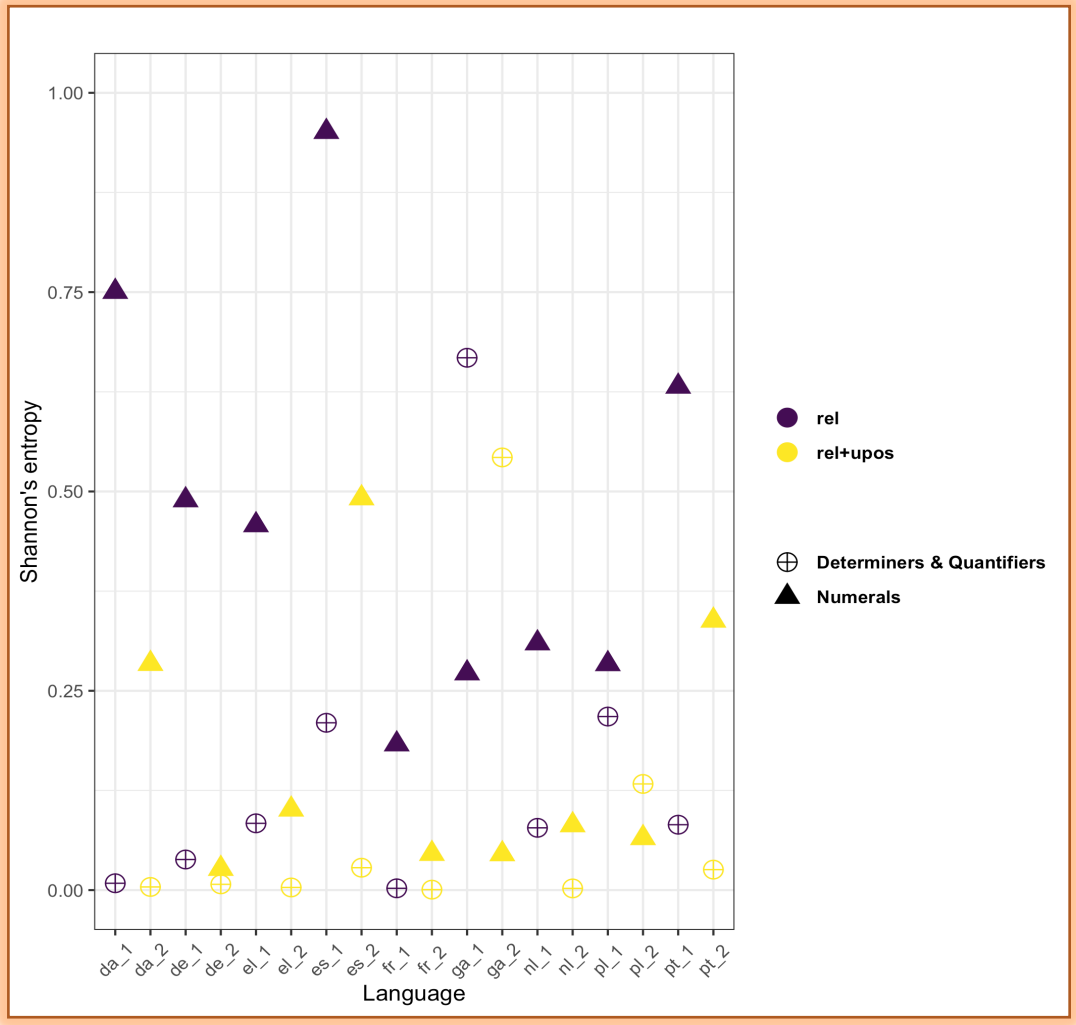
Tweaking the UD annotations

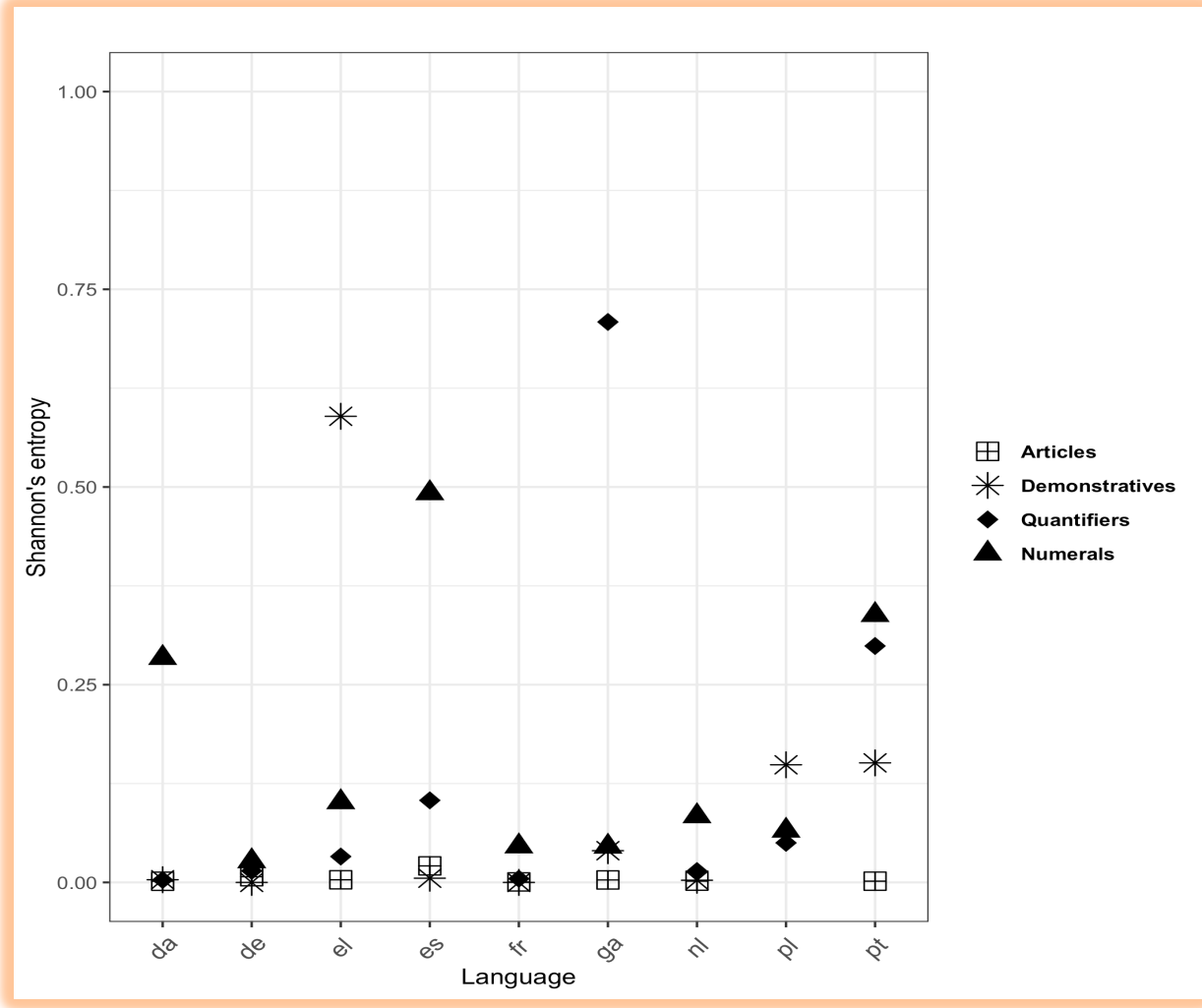
- We start from typologically-adequate **comparative concepts** and we try to **match them against different layers** of UD annotations.

Category	UPOS	UD Relation
nominal head	NOUN, PROPN	-
article	DET	`det`
demonstrative	DET *PRON*	`det`
quantifier	DET *ADJ ADV PRON*	`det det:nummod det:numgov`
numeral	NUM	`nummod nummod:entity nummod:gov nummod:flat`

- **List of Lemmata layer**: **hand-written lists** of articles, demonstratives and quantifiers, as described by grammars
- **Boolean operators**: AND between the layers of annotation; OR between the different values.
- **Do Not Throw Anything Away**: we extract all data from the parsed corpora, then **we apply these simple 'tweaks'** in further steps.

Results


1.00
0.75
0.50
0.25
0.00
Shannon's entropy
da_1
da_2
de_1
de_2
el_1
el_2
es_1
es_2
fr_1
fr_2
ga_1
ga_2
nl_1
nl_2
pl_1
pl_2
pt_1
pt_2
Language
rel
rel+upos
Determiners & Quantifiers
Numerals


1.00
0.75
0.50
0.25
0.00
Shannon's entropy
da
de
el
es
fr
ga
nl
pl
pt
Language
Articles
Demonstratives
Quantifiers
Numerals

Thank you !

Credits: Wiktionary (CC-by-NC)

In a nutshell

- a simple combination of two layers of UD annotation plus language-specific list of lemmata is used to estimate the entropy of determiners, quantifiers and numerals in the NP in a parallel corpus of 10 IE languages;
- the quality of the analysis is improved and the methodology sheds light on previously hidden categories, such as articles, demonstratives and quantifiers;
- high-to-moderate values of entropy in Greek demonstratives, high values of entropy in Irish quantifiers.