Dissecting Treebanks to Uncover Typological Trends.
A Multilingual Comparative Approach

Chiara Alzetta[•◇], Felice Dell'Orletta[◇], Simonetta Montemagni[◇], Giulia Venturi[◇]

[•]DIBRIS, Università degli Studi di Genova, Italy
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) - ItaliaNLP Lab
chiara.alzetta@edu.unige.it, {name.surname}@ilc.cnr.it

Introduction and Motivation

Typological studies can highly benefit from the synergy between linguistics and computational linguistics that makes possible to acquire quantitative evidence shedding light on how, why and to what extent languages vary with respect to key features covering major areas of language structure. Thus, linguistic typology becomes a way to tackle the bottleneck deriving from the lack of annotated data for many languages for cross- and multi-lingual NLP community.

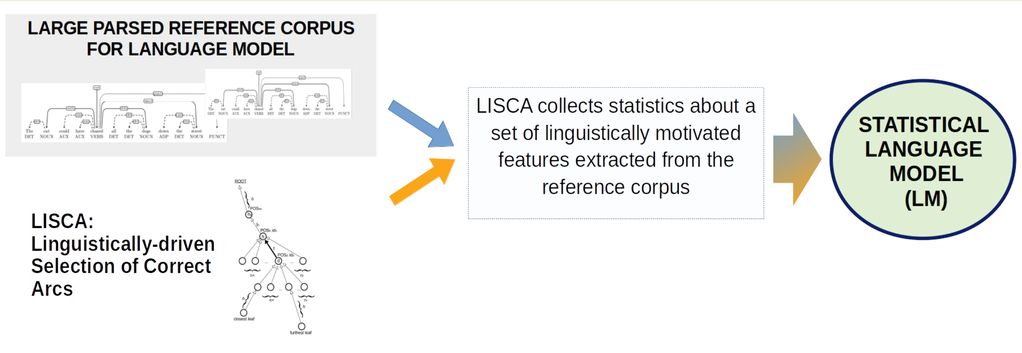
Online Databases. Manually constructed by linguists, they represent the main source to acquire typological information. The *World Atlas of Language Structures* (WALS) is the most commonly-used and broadest database of structural (phonological, grammatical, lexical) properties of languages. Reported data are based on sources such as grammars, dictionaries and scientific papers. Typological evidence inferred from linguistically annotated corpora for different languages can significantly contribute to model linguistic variation within and across languages as well as covering missing information in databases.

UD and Languages Comparability. The *Universal Dependencies project* made available linguistically annotated corpora with a cross-linguistically consistent annotation scheme. These resources are allowing new comparative linguistic studies aimed to identify similarities as well as idiosyncrasies among typologically different languages.

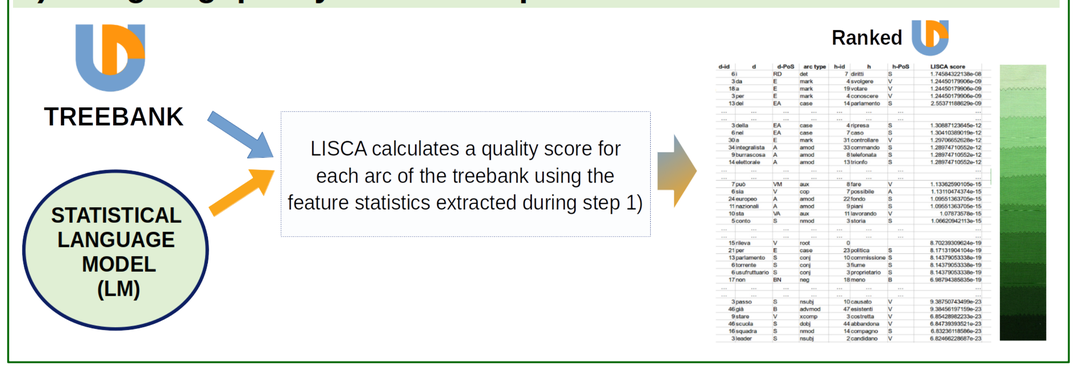
NLP and Typology Corpus-based studies can help to automatically acquire quantitative typological evidence which might be exploited for polyglot NLP. The line of research described here is aimed at acquiring quantitative typological evidence from UD treebanks through a multilingual contrastive approach.

The Approach

1) Create Statistical Language Models



2) Assigning quality score to deprels in treebank



Experimental Data

Reference Corpora:

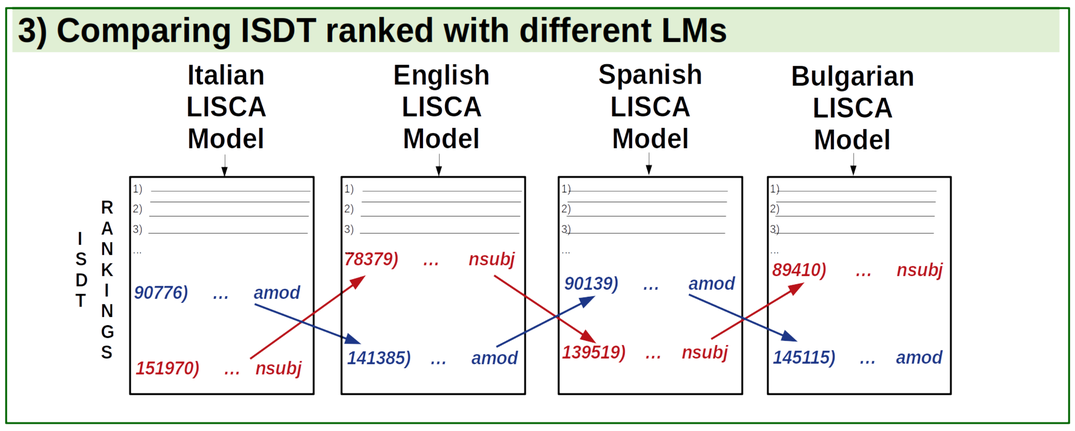
Four monolingual corpora from Wikipedia and newspapers of around 40 million tokens each parsed by the UDPipe pipeline trained on the UD treebanks.

UD treebanks vers 2.2:

- Italian ISDT (288,352 tokens)
- English EWT (231,787 tokens)
- Spanish Ancora (544,040 tokens)
- Bulgarian BTB (141,860 tokens)

The approach is exemplified on a single dependency relation, namely adjectival modifier (*amod*). For each language, we compare *amod* distributions in the ranked treebank obtained wrt different language-specific models.

3) Comparing ISDT ranked with different LMs



Typological Analysis

Correlating ranked treebanks

For a given treebank, different dependencies rankings were obtained using language-specific models. Spearman's correlation coefficients were computed between pairs of rankings of the same treebank: each pair is represented by the ranking obtained using the language models of the target UD treebank (*Target UDT models*) and the ranking obtained using LMs of other languages.

Hypothesis: typologically similar languages used as LISCA language models will produce more similar rankings, both in general and for single deprels.

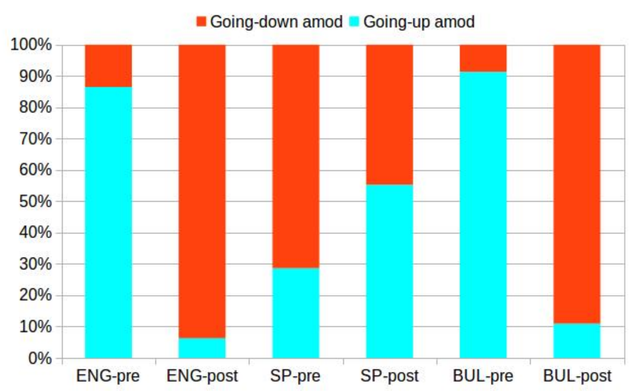
Results: the table below reports Spearman's rank correlation coefficients (p <0.00) between pairs of ranked lists of *amod*s obtained by using different LMs on each treebank.
Higher correlation values are obtained e.g. when the Italian treebank is ranked using the Spanish language model or viceversa. Lower values are reported between e.g. Italian and Bulgarian. Interestingly, correlation values are not always symmetric (see e.g. the contrast between EN vs IT and IT vs EN whose correlation is 0.94 and 0.84 respectively), suggesting that typological symilarity is asymmetric.

Language Model	Target UDT Models			
	Italian	English	Spanish	Bulgarian
Italian	**1.00**	0.94	0.97	0.79
English	0.84	**1.00**	0.87	0.90
Spanish	0.98	0.95	**1.00**	0.93
Bulgarian	0.79	0.91	0.83	**1.00**

Ranking Variations

We observe fluctuations in rankings of Italian *amod* deprels. Considering ISDT ranked with the Italian LM as benchmark, we consider how many *amod*s are placed in higher or lower positions in rankings obtained with LMs of other languages.
Hypothesis: the higher the number of ranking fluctuations, the more typologically distant the languages are.
Results: they reflect the lower degree of prototipicality of post-nominal adjectives in English and Bulgarian wrt Italian. The direction of the fluctuation highlights the properties of the used LM: deprels going up occur in constructions typical for the used LM, whereas those going down are atypical if not deviant.

Language Models	Pre-nominal		Post-nominal	
	Up	Down	Up	Down
English	21,691.18	2,566.57	290.13	40,934.52
Spanish	1,541.31	9,163.83	5,462.83	5,322.66
Bulgarian	30,597.76	967.66	885.15	43,003.64



Typology for Polyglot NLP (TyP-NLP 2019), 1 August 2019, Florence, Italy.