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.

Correlating ranked treebanks

For a given treebank, different dependencies rankings were obtained using language-specific models. Spearman’s rank 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 < 0.05) between pairs of ranked lists of amods obtained by using different LMs on each treebank.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Target UDT Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Italian</td>
</tr>
<tr>
<td>Italian</td>
<td>1.00</td>
</tr>
<tr>
<td>English</td>
<td>0.84</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.98</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>0.79</td>
</tr>
</tbody>
</table>

The degrees of prototypicality of post-nominal adjectives in English and Bulgarian wrt Italian. The direction of the fluctuation highlights the properties of the used language models. Considering ISDT ranked with the Italian LM as benchmark, we consider how many amods 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 prototypicality of post-nominal adjectives in English and Bulgarian wrt Italian. The direction of the fluctuation highlights the properties of the used language model: depreds going up occur in constructions typical for the used LM, whereas those going down are atypical if not deviant.

The Approach

1) Create Statistical Language Models

![Diagram of creating statistical language models](image1)

2) Assigning quality score to deprels in treebank

![Diagram of assigning quality score to deprels in treebank](image2)

3) Comparing ISDT ranked with different LMs

![Diagram of comparing ISDT ranked with different LMs](image3)

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.

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 with different language-specific models.

**Typological Analysis**

**Correlating ranked treebanks**

For a given treebank, different dependencies rankings were obtained using language-specific models. Spearman’s rank 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 < 0.05) between pairs of ranked lists of amods 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 similarity is asymmetric.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Italian</th>
<th>English</th>
<th>Spanish</th>
<th>Bulgarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>1.00</td>
<td>0.94</td>
<td>0.97</td>
<td>0.79</td>
</tr>
<tr>
<td>English</td>
<td>0.84</td>
<td>1.00</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.98</td>
<td>0.95</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>0.79</td>
<td>0.91</td>
<td>0.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Ranking Variations**

We observe fluctuations in rankings of Italian amod deprels. Considering ISDT ranked with the Italian LM as benchmark, we consider how many amods 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 prototypicality of post-nominal adjectives in English and Bulgarian wrt Italian. The direction of the fluctuation highlights the properties of the used language model: deprels going up occur in constructions typical for the used LM, whereas those going down are atypical if not deviant.