Quantitative Properties of Languages

Zipf's Law (Zipf, 1936) have attracted much attention. But the known quantitative universals are only about what can be easily calculated given masses of unannotated text: i.e. mostly frequency distributions. But what about quantitative properties of syntax?

Here we present some results from investigations of:

- Word Order Freedom: Languages that allow many word orders in principle might vary in how much freedom they really exhibit. Word order variability is supposed to correlate with the presence of case marking (e.g., Sapir, 1923; Riipsa, 1997; McFadden, 2003). If we want to know if more variability implies more case marking, we need quantitative measures of word order freedom.

- Dependency Length: A large body of research (e.g. Gibson, 1998; Hawkins, 1994, 2004; Gildea & Temple, 2010; Tily, 2012) argues that languages should evolve to minimize the length between heads and their dependents. Average dependency length is a quantitative property of language syntax which should be minimized.

**Data Sources**

Recent interest in multilingual dependency parsing in NLP has resulted in the release of dependency parses in many languages (e.g. the CoNLL 2007 Shared Task (Nivre et al., 2007)).

Differences in annotation have been harmonized by two separate projects: Hamlet (Zeman et al., 2012) and the Google Universal Dependency Treebank (MacDonald et al., 2013). We have combined these corpora and done further harmonization.

The corpora are mostly newspaper text and novels. Exceptions are the Japanese corpus, which is spoken, and Latin and Ancient Greek, which include metered poetry.

**Table 1. Corpora available and their properties.**

**Dependency Formalism**

<table>
<thead>
<tr>
<th>Language</th>
<th>Speakers</th>
<th>Source</th>
<th>Dependency Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>35000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Dutch</td>
<td>13000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>German</td>
<td>65000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>French</td>
<td>60000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Italian</td>
<td>55000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Spanish</td>
<td>40000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Russian</td>
<td>45000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Danish</td>
<td>50000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Swedish</td>
<td>10000000</td>
<td>Tove</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Hebrew</td>
<td>60000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Arabic</td>
<td>35000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Bengali</td>
<td>45000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Finnish</td>
<td>35000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
<tr>
<td>Japanese</td>
<td>10000000</td>
<td>Brown</td>
<td>Universal Dependencies</td>
</tr>
</tbody>
</table>

**Figure 1. An example of the dependency formalism used in the corpora.**

**Figure 2. Branching Direction Entropy and mean normalized dependency length, excluding Latin and Ancient Greek.**

**Figure 3. Order of subjects and objects, colored by case marking system.**

**Figure 4. Log-log plots of frequency distributions of dependency lengths.**

**Figure 5. Average dependency lengths of real, minimal, and random realizations for sentences of varying lengths in the corpora.**

**Figure 6. Branching Direction Entropy and Measuring Branching Direction Entropy**

**Conclusions**

Dependency-parsed corpora make typology of quantitative syntactic possible. We find results that are broadly consistent with previous claims about universal pressures on quantitative syntactic, but with complications. Using dependency corpora we have developed ease of comparison measures of word order freedom and shown that high word order variability of subjects and objects implies case marking, but not vice versa.

We have shown that dependency lengths are minimized across varied languages. But dependency length seems to correlate with word order freedom.

Besides the work presented, we believe that the measures and methods developed here can be used to quantitatively answer long-standing questions about cross-linguistic syntactic phenomena.

**References**


**SO Relation Order Entropy**

Word Order Freedom

**Using Entropy to Quantify Variability**

We use conditional entropy to quantify order conditionality on related type. This is interpretable as the degree of uncertainty about order within relations.

Entropy measures are sensitive to sample size: To make this sure it is not influencing our results, we calculated the measures presented here on small subsets (1000 tokens) of the corpora; we found very little difference in the resulting numbers (r² = 97 between the measures calculated on subsets and measures calculated on the whole corpora).

**Order Entropy of Subj and Obj**

Here we show the entropy of the order of nsubj and dobj relations under verbs where both are present. This is the word order variability for subjects and objects.

Objects are colored for their case marking system. High-variability languages all have case marking, but many case-marked languages have low variability.

- Also for SOV and VSO languages these are case-marked, which fits a noisy-channel communication account of case marking (Gibson et al., 2013).

**Dependency Length and Word Order Freedom**

We find a weak positive correlation (r = 0.45, p < 0.04) between word order freedom and dependency length. This should be unsettling to proponents of dependency length minimization: Are languages using word order freedom to increase their dependency length?