the investigations of:

Word Order Freedom

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

Entropy measures are sensitive to sample size: To make sure this is not influencing our results, we also calculated the measures presented here on small subsets (1000 tokens) of the corpora; we found very little difference in the resulting numbers (p>0.97 between the measures calculated on subsets and measures calculated on the whole corpora).

Branching Direction Entropy

The conditional entropy of head direction conditioned on relation type. (Bounded between 0 and 1, where 0 is totally deterministic, and 1 means there is total uncertainty about head direction.)

There is great variability in EOE: from 0 for Japanese and Korean to near .75 for Finnish and Estonian. No language exceeds .75.

Some related languages are very similar (e.g., Finnish and Estonian, Telugu and Tamil). Other families are more variable (e.g. Romance).

Order Entropy of Subj and Obj

Here we show the entropy of the order of subj and obj relations under verbs where both are present. This is the word order variability.

Languages are colored for their case marking system. High-variability languages all have case marking.

Also: all SOV and VSO languages here are case-marked, which fits a noisy channel communication account of case marking (Gibson et al., 2012).

Figure 6. Branching Direction Entropy and mean normalized dependency length. Including Latin and Ancient Greek. Normalized dependency length: The length of an observed dependency, scaled between 0 (the length of that dependency in the minimal projective linearization) and 1 (its average length in a random projective linearization).

Conclusions

Dependency-parsed corpora make typology of quantitative syntax possible. We find results that are broadly consistent with previous claims about universal pressures on quantitative syntax, but with complications.

Using dependency corpora we have developed easily interpretable 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


Table 1. Corpora available and their properties.

<table>
<thead>
<tr>
<th>Dependency Formalism</th>
<th>Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>(root)</td>
<td></td>
</tr>
<tr>
<td>boy</td>
<td></td>
</tr>
<tr>
<td>kicked (adj)</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
</tr>
<tr>
<td>red</td>
<td></td>
</tr>
<tr>
<td>ball</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>