Word order correlations from a quantitative perspective

Matías Guzmán Naranjo and Laura Becker

15.11.2018, Paris
Word order typology

Since Greenberg (1963), crosslinguistic word order correlation and related questions have received a lot of attention in language typology (e.g. Cristofaro, 2018; Dryer, 1992, 2009, 2019; Hawkins, 1994, 2014; Payne, 1992; Siewierska, 1988; Song, 2009).

Some examples of robust crosslinguistic generalizations concerning the verb-object order and the order of other elements in the clause (Dryer, 1991, 1992, 2009):

<table>
<thead>
<tr>
<th>VO</th>
<th>OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>prepositions</td>
<td>postpositions</td>
</tr>
<tr>
<td>postnominal relative clause</td>
<td>prenominal genitive</td>
</tr>
<tr>
<td>prenominal article</td>
<td>postnominal article</td>
</tr>
<tr>
<td>verb - adverb</td>
<td>adverb - verb</td>
</tr>
<tr>
<td>clause-initial complementizer</td>
<td>clause-final complementizer</td>
</tr>
</tbody>
</table>
Word order typology: types of explanations

“Cross-category harmony” (Hawkins, 1983)

a general preference for a head-dependent order within a given language

→ find verb-initial languages with mostly all of the dependents following their heads

→ verb-final languages should mostly have all dependents preceding their heads

“Branching directory theory” (Dryer, 1992, 2009)

Word order correlations reflect a tendency for languages to be consistently left-branching or right-branching.
This traditional approach, based on categorical decisions concerning word order is problematic:

- It is difficult to determine the main word order of a language.
- While some languages show rigid word order, others are much more flexible.
  
  → This approach treats these two types of languages equally.
  
  → For languages with flexible word order, other minor patterns are disregarded.

We can overcome this problem if we take a corpus based approach instead, and model word order tendencies as gradient.
Dataset

Universal Dependencies Treebank 2.2 (Nivre et al., 2016)

- we removed those treebanks without complete annotations
- treebanks for 70 languages of 20 subfamilies (8 are Indo-European)
  - Afro-Asiatic, (4)
  - Altaic (6),
  - Austronesian (2),
  - Basque (1),
  - Defoid (1),
  - Dravidian (2),
  - Indo-European (Armenian (1), Baltic (2), Celtic (2), Germanic (9), Greek (2), Romance (9), Slavic (12))
  - Indo-Iranian (6),
  - Pama-Nyungan (1),
  - Sinitic (2),
  - Uralic (5),
  - Viet-Muong (1)
  - Creole (1), Swedish Sign Language (1)
We are aware some shortcomings of this dataset:

- There is relatively little family variation.
- The corpora for non Indo-European languages are smaller than the datasets for languages like Czech or Russian.
- We entirely depend on the annotation schemes used by the treebank creators.

Typological studies usually take a lot more care in selecting a balanced sample of languages (Bickel, 2008; Dryer, 1989, 2019).

Despite this clear issue, the results we obtain from looking at the Universal Dependency dataset serve as a robust starting point for future work on quantitative word order correlations.
Extracted dependencies

We extracted the dependents from the treebanks for each noun and each verb, and distinguish between their relative order with the head

- head – dependent (following)
- dependent – head (preceding)

We then calculated the proportion of a given dependent following its head (noun or verb).
Extracted dependencies

For verb dependents the following part-of-speech tags were considered:

- NOUN
- VERB
- PROPN (proper noun)
- PRON (pronoun)
- AUX (auxiliary)

For noun dependents we considered all part-of-speech tags.
Verb dependents

We took into account the following types of verb dependents:

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>advcl</td>
<td>adverbial clause modifiers</td>
<td><em>He talked to him in order to secure the account.</em></td>
</tr>
<tr>
<td>advmod</td>
<td>adverbial modifiers (non clausal)</td>
<td><em>genetically modified food</em></td>
</tr>
<tr>
<td>nsubj</td>
<td>nominal subject (noun phrase which acts as subject of the verb), first core argument of the clause</td>
<td><em>There is a ghost in the room.</em></td>
</tr>
<tr>
<td>obj</td>
<td>(direct) object of a verb, second core argument of the clause</td>
<td><em>She gave me a raise.</em></td>
</tr>
<tr>
<td>obl</td>
<td>oblique, or non-core argument of the verb</td>
<td><em>Last night, I swam in the pool.</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>give the toys to the children</em></td>
</tr>
</tbody>
</table>
# Noun dependents

<table>
<thead>
<tr>
<th>Dependency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>advcl</td>
<td>Adverbial clause modifiers</td>
</tr>
<tr>
<td>acl</td>
<td>Clausal modifiers of nouns</td>
</tr>
<tr>
<td>amod</td>
<td>Adjectival modifiers</td>
</tr>
<tr>
<td>case</td>
<td>Used for any case-marking element which is treated as a separate syntactic word (mostly prepositions, but also postpositions, and clitic case markers)</td>
</tr>
<tr>
<td>compound</td>
<td>Relation used to mark noun compounding</td>
</tr>
<tr>
<td>det</td>
<td>Nominal determiners</td>
</tr>
<tr>
<td>nmod</td>
<td>Nominal modifiers of other nouns (not appositional)</td>
</tr>
<tr>
<td>nummod</td>
<td>Numerical modifiers of nouns</td>
</tr>
</tbody>
</table>

Examples:

- *He was the one present when it happened.*
- *There are many online sites offering booking facilities.*
- *the issues as he sees them*
- *Sam eats red meat*
- *the office of the Chair*
- *phone book*
- *which book, the woman*
- *the dog’s bone*
- *Sam ate 3 potatoes*
Results

We explore three questions in this section, exploring the proportions of head-following dependents:

1. the density distributions of head-following dependents
2. the order correlations among noun dependents as well as among verb dependents (intra-categorial correlations)
3. predictability of noun dependent orders from verb dependent orders and vice versa (cross-categorial correlations)
Distributions

We first explore the distribution of all dependents and their position with respect to their heads.

We look at the density of the proportion of follows for each dependent.
Density distribution: verb dependents

![Graph showing density distribution of verb dependents. The graph plots the proportion of followings against the density for different parts of speech: advcl, advmod, nsubj, obj, obl. The legend on the right indicates the color coding for each part of speech.]
Density distribution: noun dependents

![Graph showing density distribution of noun dependents](attachment:image.png)

The graph illustrates the distribution of different types of noun dependents (n_advc, acl, amod, compound) over the proportion of follows. The x-axis represents the proportion of follows, while the y-axis shows the density values. The key for the graph is provided on the right side.
Density distribution: noun dependents
Intra-categorial correlations

We see how verb dependents and noun dependents are correlated among them.

noun dependents

verb dependents

<table>
<thead>
<tr>
<th></th>
<th>acl</th>
<th>advcl</th>
<th>advmod</th>
<th>amod</th>
<th>case</th>
<th>compound</th>
<th>det</th>
<th>nmod</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>0.6</td>
<td>0.4</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>advcl</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>advmod</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>amod</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>case</td>
<td>0.4</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>compound</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>det</td>
<td>0.4</td>
<td>0.6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>nmod</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>n_advcl</th>
<th>nummod</th>
<th>nsubj</th>
<th>obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_advcl</td>
<td>0.3</td>
<td>0.6</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>nummod</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>nsubj</td>
<td>0.4</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>obj</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.4</td>
</tr>
</tbody>
</table>

rho values:

-1.0  -0.5  0.0  0.5  1.0
Cross-categorial correlations

We see how verb noun dependents are correlated with verb dependents.
Models

We fitted beta regression models for each factor (verb or noun dependent) as a dependent variable, and using family as a random effect.

To prevent overfitting we carried out stepwise factor elimination.

For each model, we calculated the marginal and conditional $R^2$ values following the method developed by (Nakagawa, Johnson, and Schielzeth, 2017; Nakagawa and Schielzeth, 2013).
We used:

- **Marginal $R^2$**: Portion of the data explained by the fixed effects (dependents).
- **Conditional $R^2$**: Portion of the data explained by the fixed (dependents) and random (families) effects.

This is a reasonable way to evaluate model performance, as well as to know how much of the variation is due to factor correlations, and how much to family biases.
Models predicting noun dependents

<table>
<thead>
<tr>
<th>predicted</th>
<th>intercept</th>
<th>advcl</th>
<th>nsubj</th>
<th>nsubj:obj</th>
<th>obj</th>
<th>obj:obl</th>
<th>obl</th>
<th>R2_m</th>
<th>R2_c</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>0.02</td>
<td>2.02</td>
<td>-1.43</td>
<td></td>
<td>6.39</td>
<td></td>
<td></td>
<td>0.462</td>
<td>0.462</td>
</tr>
<tr>
<td>advcl</td>
<td>-1.29</td>
<td></td>
<td></td>
<td></td>
<td>0.94</td>
<td></td>
<td>3.25</td>
<td>0.428</td>
<td>0.555</td>
</tr>
<tr>
<td>amod</td>
<td>-1.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.56</td>
<td></td>
<td>0.076</td>
<td>0.362</td>
</tr>
<tr>
<td>case</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.48</td>
<td>0.099</td>
<td>0.67</td>
</tr>
<tr>
<td>compound</td>
<td>-1.63</td>
<td>1.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.111</td>
<td>0.285</td>
</tr>
<tr>
<td>det</td>
<td>-2.88</td>
<td></td>
<td>0.74</td>
<td>-9.36</td>
<td>-0.11</td>
<td></td>
<td>2.10</td>
<td>0.170</td>
<td>0.170</td>
</tr>
<tr>
<td>nmod</td>
<td>-0.95</td>
<td>3.71</td>
<td></td>
<td>-5.31</td>
<td>7.20</td>
<td></td>
<td>-1.36</td>
<td>0.246</td>
<td>0.720</td>
</tr>
<tr>
<td>nummod</td>
<td>-2.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.64</td>
<td>0.079</td>
<td>0.409</td>
</tr>
</tbody>
</table>
Models predicting noun dependents

<table>
<thead>
<tr>
<th>predicted</th>
<th>intercept</th>
<th>acl</th>
<th>advcl</th>
<th>case</th>
<th>compound</th>
<th>nmod</th>
<th>R2_m</th>
<th>R2_c</th>
</tr>
</thead>
<tbody>
<tr>
<td>advcl</td>
<td>-0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72</td>
<td>1.57</td>
<td>0.15</td>
</tr>
<tr>
<td>advmod</td>
<td>-2.07</td>
<td></td>
<td>1.65</td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td>0.240</td>
</tr>
<tr>
<td>nsubj</td>
<td>-1.17</td>
<td>-1.54</td>
<td>2.27</td>
<td>-1.26</td>
<td></td>
<td></td>
<td></td>
<td>0.161</td>
</tr>
<tr>
<td>obj</td>
<td>-0.30</td>
<td></td>
<td>2.86</td>
<td>-2.15</td>
<td></td>
<td></td>
<td></td>
<td>0.433</td>
</tr>
<tr>
<td>obl</td>
<td>-1.05</td>
<td></td>
<td>2.92</td>
<td>-1.64</td>
<td></td>
<td></td>
<td></td>
<td>0.445</td>
</tr>
</tbody>
</table>
Fitted models
Fitted models

Motivation
Materials
Methodology
Results
Concluding remarks
References

- Afrikaans
- Amharic
- Ancient Greek
- Arabic
- Armenian
- Basque
- Belarusian
- Breton
- Bulgarian
- Buryat
- Cantonese
- Catalan
- Chinese
- Coptic
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Faroese
- Finnish
- French
- Galician
- German
- Gothic
- Greek
- Hebrew
- Hindi
- Hungarian
- Indonesian
- Irish
- Italian
- Japanese
- Kazakh
- Komi Zyrian
- Korean
- Kurmanji
- Latin
- Marathi
- Naija
- North Sami
- Norwegian
- Old Church Slavonic
- Old French
- Persian
- Polish
- Portuguese
- Romanian
- Russian
- Sanskrit
- Serbian
- Slovak
- Slovenian
- Spanish
- Swedish
- Swedish Sign Language
- Tagalog
- Tamil
- Telugu
- Turkish
- Ukrainian
- Upper Sorbian
- Urdu
- Uyghur
- Vietnamese
- Warlpiri
- Yoruba

nmod observed vs nmod predicted
Fitted models

The diagram illustrates the fitted models across various languages, with the axes representing the observed and predicted values. The languages are plotted according to their linguistic and demographic data, indicating how well the models fit the observed trends.
Fitted models
Concluding remarks

Using treebanks helps to gain new insights on word order typology.

- **Gradience** We should rethink the classic word order correlations as being gradient instead of categorical.

- **Order consistency** The consistency of dependent orders vary across different types of dependents across languages: some dependents (e.g. det, acl) show a clear tendency towards preceding or following the head, while others (e.g. nmod, case) do not.

- **Intra-categorical correlations** For both verb and noun dependents, we find some strong intra-categorical order correlations; as well as negative correlations between case and other nominal dependents.

- **Cross-categorical correlations** Dependents that are good predictors are not necessarily well predicted themselves (obl is a better predictor than obj for nominal dependent orders, but both can be predicted equally well). Different types of dependent orders are more (e.g. case) or less (e.g. det) sensitive to family biases.
Future work

We see two potential paths for future work:

- distinguish between different main and subordinate clauses, since subordinate clauses have been shown to be more conservative syntactically (e.g. Bybee, 2002)
- convert the UD format to some other linguistic annotation (HPSG, LFG, TAG, etc.) and see whether the theoretical elements of these theories improve the cross-linguistic patterns
Thank you!


Bibliography II


Languages

Afrikaans, Amharic, Ancient Greek, Arabic, Armenian, Bambara, Basque, Belarusian, Breton, Bulgarian, Buryat, Cantonese, Catalan, Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Galician, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Komi Zyrian, Korean, Kurmanji, Latin, Latvian, Lithuanian, Maltese, Marathi, Naija, North Sami, Norwegian, Old Church Slavonic, Old French, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Swedish Sign Language, Tagalog, Tamil, Telugu, Thai, Turkish, Ukrainian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Yoruba
Verb dependents

Figure: Proportions for verb dependents.
Noun dependents

Figure: Proportions for noun dependents.
Models predicting noun dependents

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Intercept</th>
<th>advcl</th>
<th>nsubj</th>
<th>nsubj:obj</th>
<th>obj</th>
<th>obj²</th>
<th>obj:obl</th>
<th>obl</th>
<th>obl²</th>
<th>R²_m</th>
<th>R²_c</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl</td>
<td>0.02</td>
<td>2.02</td>
<td>-1.43</td>
<td></td>
<td>6.39</td>
<td>-3.81</td>
<td></td>
<td></td>
<td></td>
<td>0.462</td>
<td>0.462</td>
</tr>
<tr>
<td>advcl</td>
<td>-1.29</td>
<td></td>
<td></td>
<td></td>
<td>0.94</td>
<td>-5.45</td>
<td></td>
<td></td>
<td></td>
<td>0.428</td>
<td>0.555</td>
</tr>
<tr>
<td>amod</td>
<td>-1.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.076</td>
<td>0.362</td>
</tr>
<tr>
<td>case</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.099</td>
<td>0.67</td>
</tr>
<tr>
<td>compound</td>
<td>-1.63</td>
<td>1.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.111</td>
<td>0.285</td>
</tr>
<tr>
<td>det</td>
<td>-2.88</td>
<td></td>
<td></td>
<td></td>
<td>0.74</td>
<td>-9.36</td>
<td>-0.11</td>
<td>2.10</td>
<td>3.26</td>
<td>0.170</td>
<td>0.170</td>
</tr>
<tr>
<td>nmod</td>
<td>-0.95</td>
<td>3.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.20</td>
<td>-1.36</td>
<td>0.246</td>
<td>0.720</td>
</tr>
<tr>
<td>nummod</td>
<td>-2.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.079</td>
<td>0.409</td>
</tr>
</tbody>
</table>