Noun-to-verb ratio and word order

Jan Strunk\textsuperscript{1}, Balthasar Bickel\textsuperscript{2}, Swintha Danielsen\textsuperscript{3}, Iren Hartmann\textsuperscript{4}, Brigitte Pakendorf\textsuperscript{5}, Søren Wichmann\textsuperscript{4}, Alena Witzlack-Makarevich\textsuperscript{6}, Taras Zakharko\textsuperscript{2}, Frank Seifart\textsuperscript{1,4}

\textsuperscript{1}U Amsterdam, \textsuperscript{2}U Zürich, \textsuperscript{3}U Leipzig, \textsuperscript{4}MPI for Evolutionary Anthropology Leipzig, \textsuperscript{5}CNRS & U Lyon-Lumiè\textsuperscript{2}, \textsuperscript{6}U Kiel

Also involved in data processing and annotation: Helen Geyer, Olga Khorkhova, Kristina Labs, Marlous Oomen, Robert Schikowski, Lena Sell, Lisa Steinbach, Laura Wägerle, and Evgeniya Zhivotova
Initial observation

Variation in noun vs. verb availability and/or usage across
- the lifespan (Tardif et al. 1997†, Bornstein et al. 2004§, Stoll et al. 2010†)
- brain health status (Bird et al. 2000‡, Thompson et al. 2002º)
- genres, registers, styles (Biber et al. 1998+, Gaenszle et al. 2010%) 
- cultures and languages (Bickel 2003*, Stoll & Bickel 2009#)

or across combinations of these

Similar observations in our project *The relative frequencies of nouns, pronouns, and verbs cross-linguistically (NTVR)*
NTVR project: spoken corpora of 9 languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Speakers</th>
<th>Texts</th>
<th>Annotation</th>
<th>Units</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baure (Arawakan; Danielsen et al. 2009)</td>
<td>15</td>
<td>45</td>
<td>4,925</td>
<td></td>
<td>19,911</td>
</tr>
<tr>
<td>Bora (Boran; Seifart 2009)</td>
<td>46</td>
<td>37</td>
<td>4,037</td>
<td></td>
<td>29,997</td>
</tr>
<tr>
<td>Chintang (Sino-Tibetan; Bickel et al. 2011)</td>
<td>74</td>
<td>40</td>
<td>9,378</td>
<td></td>
<td>37,823</td>
</tr>
<tr>
<td>Dutch (Indo-European; CGN; CGN-Consortium, Language and Speech Nijmegen &amp; ELIS Gent 2003)</td>
<td>42</td>
<td>17</td>
<td>5,822</td>
<td></td>
<td>39,720</td>
</tr>
<tr>
<td>English (NXT-Switchboard Corpus; Godfrey &amp; Holiman 1993; Calhoun et al. 2009)</td>
<td>80</td>
<td>47</td>
<td>6,942</td>
<td></td>
<td>56,143</td>
</tr>
<tr>
<td>Hoocak (Siouan; Hartmann 2013)</td>
<td>30</td>
<td>62</td>
<td>2,961</td>
<td></td>
<td>23,207</td>
</tr>
<tr>
<td>Lamunkhin Even (Tungusic; Pakendorf &amp; Aralova 2010)</td>
<td>32</td>
<td>67</td>
<td>4,755</td>
<td></td>
<td>34,294</td>
</tr>
<tr>
<td>N</td>
<td>uu (!Ui-Taa; Güldemann et al. 2010)</td>
<td>8</td>
<td>33</td>
<td>8,257</td>
<td></td>
</tr>
<tr>
<td>Texistepec Popoluca (Mixe-Zoquean; Wichmann 1996)</td>
<td>1</td>
<td>9</td>
<td>6,453</td>
<td></td>
<td>24,602</td>
</tr>
</tbody>
</table>
A simple example: $\text{NTVR} = \frac{N}{N+V}$

<table>
<thead>
<tr>
<th>clause</th>
<th>N</th>
<th>V</th>
<th>NTVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A <em>man stayed</em> on a <em>farm</em>.</td>
<td>2</td>
<td>1</td>
<td>0.67</td>
</tr>
<tr>
<td>He <em>got</em> hungry.</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>He <em>says</em> to his <em>father-in-law</em>:</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>&quot;<em>Give</em> me some <em>meat!</em>&quot;</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>His <em>father-in-law says</em>:</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>&quot;<em>I have</em> no <em>meat</em>, <em>go</em> to the <em>dune</em>, and <em>hunt!</em>&quot;</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

(English translation of a Nǀuu story)
NTVR variation in our corpora
How to explain differences in noun vs. verb usage

• Earlier research: focused on nouns in argument positions and found explanations in types of agreement systems (Bickel 2003* on referential density)

• NTVR project: focus on noun and verb usage across the board
  - unlikely to be affected by type of agreement system (Bickel et al. 2013#)
  - possible explanation: processing effects resulting from word order
  - for this study, we focus on the simple proportion of nouns rather than nouns vs. verbs
    (relative frequency: nouns / words)
Theory: noun usage dependent on word order?

- Incremental production (for recent review, MacDonald 2013*)
  - alternation of partial utterance planning, execution, and subsequent planning
  - pressure to start and complete plans early

- Good for V-early structures, with early display of plan for proposition (predicate, argument structure, tense, mood, settings, etc.)

- Predictions from this for V-final structures ...
Possible predictions for V-final structures:

- Increased usage of **non-verb tokens**, especially **nouns** as content words, in order to compensate for the delay in getting to the core information about the proposition
  - *perhaps* also more noun **type** variation (as observed in a correlational study of dictionaries by Polinsky 2012+), for more information load
- but this may be counterbalanced by increased access cost that comes with lexical variation

Possible counter-hypothesis

- Noun usage is costly/harder to process in pre-verbal argument position (Ueno & Polinsky 2009*):
  - increased pro-drop
  - increased use of intransitives

- Other options:
  - production costs can also be avoided by right-dislocation (Pastor & Laka 2013#)
  - production costs can be compensated for by optimizing lexicon shapes/the way semantic space is divided between verbs and nouns (Sauppe et al. 2013%, in prep.)
  - speakers may just live with a slight speed loss (Seifart et al. 2014, in prep: higher N-to-V ratios result in lower production speeds)
Corpus Study

• Test the research hypothesis:
  - Verb final languages exhibit increased noun usage
    (in comparison to verb non-final languages),
  - expect weak signals for tokens
  - and perhaps also for lexical types
Data

- Mapping of language-specific PoS-tags to tags of \{N, V, PRO, OTHER\} per *lexical root*

  
  ![Diagram showing PoS tagging examples]

- Why roots?
- Our hypothesis concerns units with propositionally relevant content; in our corpus, PoS derivation like nominalization usually doesn’t add information (e.g. nominalization for embedding)
- In more than 90% of cases, root and word category are identical
Methods: Linear mixed-effect models

- Linear mixed-effects models* predicting the proportion of
  1. noun tokens per annotation unit (utterance or sentence)
  2. noun types per recording session / text

- An extension of ordinary linear regression models that can account for random idiosyncrasies of natural groups in the data (e.g., texts of the same speaker, register, or language)

- \[ P(\text{nouns}) \sim \text{word order} + \text{plannedness} + (1|\text{session}) \]

- Reads as: The proportion of nouns is predicted on the basis of the two predictors word order and plannedness (fixed effects) while accounting for random variation between recording sessions (random factor).

*lme4::lmer (Bates et al. 2014, CRAN)
• Fixed factors (predictors):
  - basic word order:
    verb final vs. verb non-final (vs. mixed)
  
  - speech setting:
    monologue vs. dialogue vs. multi-party conversation, estimated on the basis of the number of speakers in a recording session
  
  - plannedness:
    - planned: (almost) memorized traditional narratives
    - semi-spontaneous: personal narratives, life stories, procedurals, etc.
    - spontaneous: open conversation
Methods

• **Random factors (for intercepts):**
  
  • **recording session**, capturing genre, topic choice, style, register, speakers and their social relations and interactions
  
  • **language**, capturing other aspects of grammar that might influence noun and verb usage
Results: proportion of nouns depending on word order
Results: proportion of nouns depending on word order

Even (Lamunkhin dialect; Tungusic; Pakendorf and Aralova 2010)
bi: ere-w inen-i-w hun-du njimkar-al-bu ukchen-dzi-m
1SG PROX-ACC day-ACC 2PL.OBL-DAT tale.PL-PL-ACC tell-FUT-1SG
PRO OTHER N PRO N V
‘Today I will tell you folk tales.’ [KKK_Emcheni_007]

Baure (Arawakan; Danielsen et al. 2009)
ni=koyepin-sha-ri-ro jeni
1SG=narrate-IRR=2SG=3SG.M yes
V OTHER
‘I will tell it to you, yes.’ [DC-N121216S.001]
kwe’=ji tech ja jir
exist=QUOT DEM2M HESIT man
V OTHER OTHER N
‘There was that man.’ [DC-N121216S.002]
Results: statistical model (proportion of nouns)

Best-fitting model: \( P(\text{nouns}) \sim \text{word order} \times \text{plannedness} + \text{speech setting} + (1|\text{session}) + (1|\text{language}) \)

interaction: \( p = .009 \), word order: \( p < .001 \), plannedness: \( p = .002 \), speech setting: \( p = .41 \), session: \( p < .001 \), language: \( p < .001 \)
Results: lexical types (proportion of noun root types)

- Results for lexical types are much less clear
- Still a detectable overall word order effect

Volkswagen Foundation DoBeS grant


Discussion

- Heavier noun usage (tokens) in annotation units (sentences) of verb-final languages than in annotation units of verb-non-final languages.

- Effect of word order detectable across categories of plannedness (planned, semi-spontaneous vs. spontaneous) and speech setting (monologue, dialogue vs. multi-party conversation).

- Word order effects mostly play out for the proportion of noun tokens, word order effects on the proportion of noun types (cf. Polinsky’s 2012 dictionary-based approach) are still unclear.
Conclusions

A small relativity effect:

The word order rules you follow also regulate the amount of noun roots you produce.

There is a higher average proportion of nouns in sentences of verb-final languages than in sentence of verb-non-final languages.

This is in line with relativity effects from other aspects of grammar (agreement systems) on noun vs. verb usage (Bickel 2003*, Stoll & Bickel 2009#).

BUT the exact relationship between these effects still needs to be explored.

*Language, #Crossling Appr Psych
Thank you very much for your attention!