Parallel corpora, conditional inference trees and random forests

How to compare constraints cross-linguistically

Natalia Levshina
University of Leipzig
The slides can be downloaded from
www.natalialevshina.com/presentations.html
Aims of the talk

• Demonstrate how one can use data from a parallel corpus to compare semantic and pragmatic constraints on formal variation cross-linguistically

• Show how one can use non-parametric methods based on permutation (conditional inference trees and random forests) for this purpose
Case studies

• T/V distinction in ten European languages
• Lexical, morphological and analytic causatives in ten languages of the world
Why parallel corpora?

• A parallel corpus contains verbalizations of the same or very similar conceptual contents in different languages.

• We can code this conceptual contents for variables and use them as predictors of the formal choices made by translators.

• These choices are an approximation of the choices ‘normal’ speakers make in verbalization of their experience.
Why conditional inference trees and random forests?

• Working with different languages, it is often difficult to obtain a very high number of observations relative to the number of predictors (‘small n large p’).
• Semantic and pragmatic variables, as a rule, are strongly intercorrelated and often participate in complex interactions.
• There can be missing values in the predictors (difficult to code).
• In these circumstances, using traditional logistic regression (*glm* or *lrm* in R) would be problematic.
• Conditional inference trees and random forests are a perfect solution!
Case study 1

T/V forms in European languages
Object of study

- T/V-distinction in addressing the hearer
- The distinction is present in most European languages
  - T forms: informal, familiar, e.g. French *tu*, German *du*,
    Russian *ty* + Verb 2\(^{nd}\) SG
  - V forms: formal, polite, e.g. French *vous*, German *Sie*,
    Russian *vy* + Verb 2\(^{nd}\) PL or 3\(^{rd}\) SG/PL
Cross-linguistic research

• WALS Chapter 45, Helmbrecht 2013

Values

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>No politeness distinction</td>
<td>136</td>
</tr>
<tr>
<td>Binary politeness distinction</td>
<td>49</td>
</tr>
<tr>
<td>Multiple politeness distinctions</td>
<td>15</td>
</tr>
<tr>
<td>Pronouns avoided for politeness</td>
<td>7</td>
</tr>
</tbody>
</table>
Research questions

• What are the cross-linguistic (dis)similarities wrt. the relative frequencies of the forms?
• What are the cross-linguistic (dis)similarities wrt. the preferences of the forms in different communicative situations?
Power and solidarity (Brown and Gilman 1960)

• Power dimension:
  • Based on “older than”, “richer than”, “parent of”, etc.
  • Systematic distinction from the late Middle Ages. Everyone has his/her fixed place in the society.

• Solidarity dimension:
  • Based on “the same age/family/class as”.
  • Emerged with social mobility and egalitarian ideology. Starting from the French revolution (*Citoyen*, *tu*).
  • Currently dominates in major European languages, but there are subtle cross-linguistic differences.
Languages in the sample

• Germanic: Dutch, German and Swedish
• Romance: French and Spanish
• Slavic: Bulgarian, Polish and Russian
• Greek
• Finnish
# T/V forms (standard varieties)

<table>
<thead>
<tr>
<th>Language</th>
<th>Nr of types</th>
<th>T-pronoun</th>
<th>V-pronoun(s), one person</th>
<th>V-verb agreement, one person</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>2</td>
<td>du</td>
<td>Sie</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; person PL</td>
</tr>
<tr>
<td>Dutch</td>
<td>2</td>
<td>jij (je)</td>
<td>u</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; person SG</td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td>du</td>
<td>ni</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
<tr>
<td>French</td>
<td>2</td>
<td>tu</td>
<td>vous</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
<tr>
<td>Spanish</td>
<td>2</td>
<td>tú</td>
<td>usted</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; person SG</td>
</tr>
<tr>
<td>Russian</td>
<td>2</td>
<td>ты [ty]</td>
<td>вы [vy]</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>2</td>
<td>ти [ti]</td>
<td>Вие ['vi.ɛ]</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
<tr>
<td>Polish</td>
<td>2</td>
<td>ty</td>
<td>пан (m)/пани (f)</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; person SG</td>
</tr>
<tr>
<td>Greek</td>
<td>2</td>
<td>еσυ [e'si]</td>
<td>еσείς [e'sis]</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td>sinä</td>
<td>te</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; PL</td>
</tr>
</tbody>
</table>
Data: ParTy corpus

- A Parallel corpus for Typologists
- Online subtitles of films and TED talks in many languages simultaneously
- Mostly Indo-European, but also Chinese, Turkish, Finnish, Indonesian, Vietnamese...
- Partly available from www.natalialevshina.com/corpus.html
Why subtitles?

Cluster Dendrogram

Based on the frequencies of 3-grams (Levshina, Forthcoming)
Films
Data set

• English data: instances of *you/yourselves* used when referring to one person.
• 243 communicative situations with unique participants (in order to ensure maximal diversity)
• Translations into 10 languages coded for T or V
Variables describing relationships between Speaker and Hearer

- **Rel_Age**: is H younger, older or of the same age (approximately) than/as S?
- **Rel_Class**: does H belong to a higher, lower or same social class as S?
- **Rel_Power**: does H have social power over S? E.g. employer > employee, prime-minister > minister, general > soldier
- **Rel_Gender**: M to M, M to F, F to M, F to F
- **Rel_Circle**: family, romance, friends, working together, school or university, “house” (household servants, hotel), prison, acquaintances, strangers
Variables describing Speaker and Hearer

- $S_{Age}$: age of S (child, young person, middle-aged, elderly)
- $H_{Age}$: age of H (child, young person, middle-aged, elderly)
- $S_{Class}$: social class of S (upper, middle, lower, other)
- $H_{Class}$: social class of H (upper, middle, lower, other)
- $S_{Gender}$: gender of S (M or F)
- $H_{Gender}$: gender of H (M or F)
Variables describing communicative settings

• *Others*: are there other people involved in the situation?

• *Office*: does the interaction take place in an office, a government building, prison, school, etc.?

• *Before68*: does the action take place before 1968?

• *Britain*: does the action take place in Great Britain?
Example from *The Grand Budapest Hotel*

*M. Gustave:* What have you done to your fingernails?

*Madame D:* I beg your pardon?

*M. Gustave:* This diabolical varnish. The color is completely wrong.

*Madame D:* Don’t you like it?

*M. Gustave:* It’s not that I don’t like it. I am physically repulsed.
Coding of the situation

• Relational variables:
  • Rel_Age: H is younger
  • Rel_Class: H belongs to a lower social class
  • Rel_Power: H has less power
  • Rel_Gender: F to M
  • Rel_Circle: romance

• Individual variables:
  • S_Age: elderly
  • H_Age: middle
  • S_Class: upper
  • H_Class: lower
  • S_Gender: F
  • H_Gender: M

• Settings:
  • Others: no
  • Office: no
  • Before_68: yes
  • Britain: no
Example

• EN: Mal, what are you doing here? *(Inception)*
  • DE: Mal, was *tust du* hier? (T-pronoun and T-verb form)
  • RU: Мол, что *ты* здесь *делаешь*? (T-pronoun and T-verb form)
  • ES: Mal, qué *haces* aquí? (T-verb form)
  • BG: Какво *правиш* тук? (T-verb form)
Proportions of T/V in 10 languages
Conditional inference trees

- deal with any data, continuous or categorical; can be an alternative to most popular types of generalized linear models
- based on binary recursive partitioning:
  1. The algorithm is in search of the covariate $X$ that is the most strongly associated with the response $Y$ (e.g. has the smallest $p$-value or test statistic).
  2. The algorithm decides on the best way of splitting the data into two subsets with different values of $X$.
     e.g. if $X$ has values a, b and c, one can split it into
     - $[a, b]$ and c
     - a and $[b, c]$
     - $[a, c]$ and b
  3. For each subset, repeat 1 and 2 until certain conditions are met (e.g. there is at least one $p < 0.05$). If not, stop.
Permutation

• Imagine we have a binary predictor X and binary response variable Y.

\[
\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|}
\hline
X & & & & & & & & & & \\
\hline
Y & & & & & & & & & & \\
\hline
9 & 2 & & & & & & & & & \\
\hline
\end{array}
\]

• Compute the test statistic for an independence test (e.g. X-squared).

\[
X \text{-squared} = 6.5455, \ df = 1, \ p\text{-value} = 0.01052
\]
Let’s reshuffle the labels of Y randomly, breaking the dependence, and compute the test statistic again.

\[ \chi^2 = 0.72727, \text{ df} = 1, \text{ p-value} = 0.3938 \]
Permutation

- Repeat the reshuffling again and again and log down the statistics.
Permutation

• After running the procedure many times (e.g. 1000), compute the proportion of observing the actual value (i.e. before permutation) of the test statistic and the more extreme values.

• This proportion is the permutation-based $p$-value.
  • In our example, only 10 permutation runs out of 1000 have the value 6.545 and greater: 10/1000 = 0.01.

<table>
<thead>
<tr>
<th>X-squared</th>
<th>0</th>
<th>0.727</th>
<th>2.909</th>
<th>6.545</th>
<th>11.636</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of permutation runs</td>
<td>611</td>
<td>301</td>
<td>78</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>
Random forests

- Aggregate data from many trees (e.g. 500 or 1,000).
- Trees are created from bootstrapped data (a smaller sample is drawn randomly from the original data).
- Improved classification accuracy in comparison with individual trees (adjusts the instability of individual trees).
- Conditional variable importance: how important each predictor is, given the other predictors and possible interactions with them.
  - For each tree: compare the classification accuracy with the given predictor X before and after permutation.
  - The greater the decrease in the classification accuracy, the more important X.
  - Conditional: while controlling for the values of the other predictors (see Strobl et al. 2008)
  - Average this information across the trees.
For R users

• Conditional inference trees and random forests are implemented in R package party (Hothorn et al. 2006); see ctree and cforest functions.

• advantages in comparison with other regression and classification trees (e.g. CART algorithm implemented in rpart) and random forests (e.g. randomForest):
  • No selection bias towards categorical predictors with many categories, where many possible splits are possible.
  • No need to prune (no overfitting) because of the build-in stopping criterion (p-value).
Conditional inference tree: Swedish
Conditional inference tree: German
Random forest: German

Conditional variable importance
Random forests: Germanic

Conditional variable importance
Random forests: Romance

Conditional variable importance
Random forests: Slavic

Conditional variable importance
Random forests: Greek and Finnish

Conditional variable importance
Results

• The most important variable is *Rel\_Circle* (no. 1 in most languages, no. 2 in some):
  • In all countries, strangers and acquaintances (low intimacy) are more frequently addressed by V, and family and friends (high intimacy) as T. There is variation, as far as other levels of intimacy are concerned. E.g. in German and French, the T forms are restricted to friends and family. In Finnish, Greek and Polish, they are the least restricted.

• The second most important variable is relative social class:
  • Usually, when H is higher on the social ladder than S, he/she is addressed by V.
  • Conditional on time (in Swedish, before 1968) and circle (in Bulgarian and Finnish, only in higher-intimacy contexts).
• \textit{Rel\_Age}: older H are more frequently addressed by V in low-intimacy contexts in Polish

• \textit{Rel\_Gender}: men speaking to men in low-intimacy situations prefer V in Finnish

• \textit{S\_Age}: younger S use T in low-intimacy situations in Bulgarian

• \textit{H\_Age}: young Hearers are addressed more frequently by T in lower-intimacy situations in French

• \textit{Office}: more V in the office in German

• \textit{Before68}: more V in French (for younger H in low-intimacy situations) and Swedish

• \textit{Britain}: more V between equals or youngsters in low-intimacy contexts in Bulgarian, Greek, Polish and Spanish
Conclusions

• The solidarity dimension is indeed the most universal and powerful, although there is substantial variation regarding the boundary between high and low intimacy.

• The power dimension is still present, however.

• There are more constraints and cross-linguistic variation in low-intimacy contexts. Higher ‘embarrassment potential’? (cf. Kretzenbacher et al. 2006).
Case study 2

Causative constructions
The causative continuum

Lexical <> Morphological <> Analytic (Periphrastic)

The causative continuum

Lexical <> Morphological <> Analytic (Periphrastic)

e.g. kill, break

The causative continuum

Lexical <> Morphological <> Analytic (Periphrastic)

e.g. *kill*, *break*  
   e.g. Turkish *öl dü r*- “kill”  
       from *öl*- “die”

The causative continuum

Lexical <-> Morphological <-> Analytic (Periphrastic)

e.g. *kill, break*  e.g. Turkish *öldür*- “kill” from *öl-* “die”  e.g. *cause X to die, make X disappear*

The causative continuum

Lexical <> Morphological <> Analytic (Periphrastic)

- e.g. *kill*, *break*
- e.g. Turkish *öldür*—“kill”
  from *öl*—“die”
- e.g. *cause* X to die,
  *make* X disappear

most compact  FORM  least compact

The causative continuum

Lexical <> Morphological <> Analytic (Periphrastic)

e.g. *kill*, *break*  
e.g. Turkish *öldür*- “kill”  
from *öl*- “die”  
e.g. *cause* *X* to *die*,  
*make* *X* disappear

**FORM**  
most compact  
least compact

**MEANING**  
most direct causation  
least direct causation

Example

- LC: I raised the cup to my lips. [Causer acting directly]
- AC: I caused the cup to rise to my lips. [Causer acting indirectly, e.g. using telekinesis]

(Haiman 1983: 784)
Functional explanation: Iconicity of cohesion

• “Meanings that belong together more closely semantically are expressed by more cohesive forms” (Haspelmath 2008: 2)

• If cause and effect are closely integrated semantically, the elements that express them will be formally integrated, too.
Dixon’s (2000) multifactorial account

<table>
<thead>
<tr>
<th>More compact causative forms</th>
<th>Less compact causative forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. non-causal verb describing a state</td>
<td>non-causal verb describing an action</td>
</tr>
<tr>
<td>2. intransitive (or intransitive and simple transitive) non-causal verb</td>
<td>transitive (ditransitive) non-causal verb</td>
</tr>
<tr>
<td>3. Causee lacking control</td>
<td>Causee having control</td>
</tr>
<tr>
<td>4. Causee willing (‘let’)</td>
<td>Causee unwilling (‘make’)</td>
</tr>
<tr>
<td>5. Causee partially affected</td>
<td>Causee fully affected</td>
</tr>
<tr>
<td>6. Causer acts directly</td>
<td>Causer acts indirectly</td>
</tr>
<tr>
<td>7. Causer achieves the result intentionally</td>
<td>Causer achieves the result accidentally</td>
</tr>
<tr>
<td>8. causation occurring naturally</td>
<td>causation occurring with effort</td>
</tr>
<tr>
<td>Language</td>
<td>Genus</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Chinese</td>
<td>Chinese</td>
</tr>
<tr>
<td>Finnish</td>
<td>Finnic</td>
</tr>
<tr>
<td>French</td>
<td>Romance</td>
</tr>
<tr>
<td>Hebrew</td>
<td>Semitic</td>
</tr>
<tr>
<td>Indonesian</td>
<td>Malayo-Sumbawan</td>
</tr>
<tr>
<td>Japanese</td>
<td>Japanese</td>
</tr>
<tr>
<td>Russian</td>
<td>Slavic</td>
</tr>
<tr>
<td>Thai</td>
<td>Kam-Tai</td>
</tr>
<tr>
<td>Turkish</td>
<td>Turkic</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>Viet-Muong</td>
</tr>
</tbody>
</table>
Subtitles

Films

TED talks

- Ken Robinson: *Do schools kill creativity?*
- Elizabeth Gilbert: *Your elusive creative genius*
- Amy Cuddy: *Your body language shapes who you are*
- Leslie Morgan Steiner: *Why domestic violence victims don’t leave*
- Dan Gilbert: *The psychology of your future self*
- Simon Sinek: *Why good leaders make you feel safe*
Data set

• 344 causative situations found in the English segment of the ParTy corpus

• Coded for various semantic and syntactic variables (mostly based on Dixon’s parameters); interrater agreement scores were computed.

• Translations in the 10 languages are found and coded into 3 types of constructions (Analytic, Morphological or Lexical)
## Variables (1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>Example(s)</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caused Event Semantics</td>
<td>Non-action Action</td>
<td>John killed Bill. He walked the baby upstairs.</td>
<td>More compact form Less compact form</td>
</tr>
<tr>
<td></td>
<td>Action</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of main arguments</td>
<td>2</td>
<td>John killed Bill. John made Bill kill Mary.</td>
<td>More compact form Less compact form</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Causee having control</td>
<td>No</td>
<td>John killed Bill. Bring your friends!</td>
<td>More compact form Less compact form</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Making or Letting</td>
<td>Let Make</td>
<td>She let him go. John killed Bill.</td>
<td>More compact form Less compact form</td>
</tr>
<tr>
<td>Willing Causee</td>
<td>No</td>
<td>John caused Bill to die. The police let him go.</td>
<td>More compact form Less compact form</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Variables (2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>Example(s)</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causer acting directly</td>
<td>Yes, No</td>
<td>He cut his finger. He had his hair cut.</td>
<td>More compact form, Less compact form</td>
</tr>
<tr>
<td>Causer acting intentionally</td>
<td>Yes, No</td>
<td>She wrote a paper. It makes me happy.</td>
<td>More compact form, Less compact form</td>
</tr>
<tr>
<td>Causer acting forcefully</td>
<td>No, Yes</td>
<td>He got him to do it. He forced him to do it.</td>
<td>More compact form, Less compact form</td>
</tr>
<tr>
<td>Causer involved in caused event</td>
<td>No, Yes</td>
<td>John killed Bill. Bring your friends! (and come, too)</td>
<td>None</td>
</tr>
</tbody>
</table>
# Variables (3)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
<th>Example(s)</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreferentiality</td>
<td>Yes</td>
<td>He killed himself.</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>He killed Bill.</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td>Pos</td>
<td>She let him do it.</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Neg</td>
<td>She didn’t let him do it.</td>
<td></td>
</tr>
<tr>
<td>Semantics of Causer</td>
<td>Anim</td>
<td>She made him stay.</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Inanim</td>
<td>The rain made him stay.</td>
<td></td>
</tr>
<tr>
<td>Semantics of Causee</td>
<td>Anim</td>
<td>John let Mary go.</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Inanim</td>
<td>John let it go.</td>
<td></td>
</tr>
</tbody>
</table>
Example from Avatar

Original

- ENG: Don't shoot, you'll **piss** him **off**.

Translations

- FRA: Ne tirez pas. Vous **allez l'énerv**er. (Lexical)
- TUR: Ateş **etme**. Ateş **etme**. Onu **kızdıracak**sın. (Morphological, from **kızmek** ‘become angry’).
- VIE: **Đừng bắn**. Cậu sẽ **làm nó nổi điên** đó. (Analytic)
# Examples of constructions

<table>
<thead>
<tr>
<th>Language</th>
<th>Lexical</th>
<th>Morphological</th>
<th>Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>sha “kill”</td>
<td>-</td>
<td>ràng “let, make” + Pred</td>
</tr>
<tr>
<td>Finnish</td>
<td>tappaa “kill”</td>
<td>odotu-tt-aa “make wait”</td>
<td>antaa “give” + V1</td>
</tr>
<tr>
<td>French</td>
<td>tuer “kill”</td>
<td>-</td>
<td>faire + Vinf</td>
</tr>
<tr>
<td>Hebrew</td>
<td>harag “kill” pa‘al</td>
<td>hotsi “take out” hiph’il</td>
<td>natan “give” + le-Vinf</td>
</tr>
<tr>
<td>Indonesian</td>
<td>mem-bunuh “kill”</td>
<td>meng-ingat-kan “remind”</td>
<td>membuat “make” + Pred</td>
</tr>
<tr>
<td>Japanese</td>
<td>korosu “kill”</td>
<td>ikar-ase-ru “make angry”</td>
<td>V_te + morau “get”</td>
</tr>
<tr>
<td>Russian</td>
<td>ubit’ “kill”</td>
<td>-</td>
<td>zastavit’ + Vinf</td>
</tr>
<tr>
<td>Thai</td>
<td>kaa “kill”</td>
<td>-</td>
<td>tham hai “do give” + Pred</td>
</tr>
<tr>
<td>Turkish</td>
<td>açmak &quot;open&quot;</td>
<td>öl-dür- “kill”</td>
<td>V_mA_DAT + izin ver- “allow”</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>giết hai “kill”</td>
<td>-</td>
<td>làm “do” + Pred</td>
</tr>
</tbody>
</table>
Proportions of types of causative constructions in 10 languages

Legend:
- Lex
- Morph
- Ana
Conditional inference trees: analytic vs. lexical in Turkish
## Features that favour AC (vs. LC)

<table>
<thead>
<tr>
<th>Feature</th>
<th>FIN</th>
<th>FRA</th>
<th>HEB</th>
<th>IND</th>
<th>RUS</th>
<th>THA</th>
<th>TUR</th>
<th>VIE</th>
<th>ZHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caused Action (vs. State)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 arguments (vs. 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlling Causee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letting (vs. making)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Willing Causee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animate Causee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-intentional Causer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inanimate Causer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Features that boost MC (vs. LC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letting (vs. making)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-intentional Causer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirectly acting Causer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIN</td>
<td>JPN</td>
<td>TUR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Functional explanations: Economy instead of iconicity

• The iconicity-based account (directness/indirectness) doesn’t explain the multidimensional character of variation.

• A better explanation may be the Principle of Economy: Shorter constructions tend to refer to more frequent causative situations, and longer ones to less frequent causative situations (Haspelmath 2008).

• Supported by corpus-based, experimental and typological evidence (Levshina, In preparation)
General conclusions

• Analysis of semantic and pragmatic constraints on formal variation based on parallel corpora can help us answer fundamental questions about language(s) and make interesting discoveries.

• Conditional inference trees and random forests are a useful tool for modelling linguistic variation based on many interrelated variables in relatively small data sets.
References


References


• Levshina, N. Forthcoming. Subtitles as a Corpus: An n-gram approach. To appear in *Corpora*.
