Analytic and lexical causatives in European: A multivariate study based on a parallel corpus of film subtitles

Natalia Levshina
F.R.S. – FNRS, Université catholique de Louvain
Outline

1. Aims of the study
2. Data: ParTy corpus
3. Statistical analyses: random forests and MDS
4. Conclusions
Analytic vs. lexical causatives

a. He *made* his cat *come back*.

b. She *brought* her cat *back*. 
Iconicity-related explanations

<table>
<thead>
<tr>
<th>Study</th>
<th>Less integrated/compact causative</th>
<th>More integrated/compact causative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Givón (1980)</td>
<td>Lower degree of semantic binding between 2 events</td>
<td>Higher degree of semantic binding between 2 events</td>
</tr>
<tr>
<td>Comrie (1981; 1989)</td>
<td>Indirect causation</td>
<td>Direct causation</td>
</tr>
<tr>
<td>Haiman (1983)</td>
<td>Greater conceptual distance between Cause and Result</td>
<td>Smaller conceptual distance between Cause and Result</td>
</tr>
<tr>
<td>Givón (1990)</td>
<td>Human-Agentive Manipulee</td>
<td>Inanimate Manipulee</td>
</tr>
</tbody>
</table>
Analytic vs. lexical causatives

a. He *made* his cat *come* back.

b. She *brought* her cat back.
Dixon’s (2000) parameters

State (or change of state)  Action

Intransitive
- No control
- Willing (‘let’)
- Partially affected
- Direct
- Intentional
- Natural

Related to Caused event
- Relating to Causee
- Relating to Causer

(Di)transitive
- Control
- Unwilling (‘make’)
- Fully affected
- Indirect
- Accidental
- With effort, violence

More compact, e.g. lexical  Less compact, e.g. analytic

More compact, e.g. lexical  Less compact, e.g. analytic
Dixon's (2000) parameters

More compact, e.g. lexical

Less compact, e.g. analytic

State (or change of state)
Intransitive
No control
Willing ('let')
Partially affected
Direct
Intentional
Natural

Action
(Di)transitive
Control
Unwilling ('make')
Fully affected
Indirect
Accidental
With effort, violence

Relating to VERB

Related to Causee

Relating to Causer
Dixon’s (2000) parameters

More compact, e.g. lexical

State (or change of state)
Intransitive
No control
Willing (‘let’)
Partially affected
Direct
Intentional
Natural

Relating to VERB
Relating to Causee
Related to Causer

Less compact, e.g. analytic

Action
(Di)transitive
Control
Unwilling (‘make’)
Fully affected
Indirect
Accidental
With effort, violence
Research questions

• Can the use of analytic and lexical causatives be best explained by
  - the iconicity-related factors?
  - different semantic and syntactic parameters, which do not boil down to iconicity only?

  Cross-linguistically
  Within one language (new!)

Are there cross-linguistic differences in the way these parameters can explain the use of the constructions? If yes, are there any genealogical and/or areal patterns?
Research questions

• Can the use of analytic and lexical causatives be best explained by
  - the iconicity-related factors?
  - different semantic and syntactic parameters, which do not boil down to iconicity only?

  Cross-linguistically
  Within one language (new!)

• Are there cross-linguistic differences in the way these parameters can explain the use of the constructions? If yes, are there any genealogical and/or areal patterns?
Languages

- Germanic
- Romance
- Slavic
Outline

1. Aims of the study
2. Data: ParTy corpus
3. Statistical analyses: random forests and MDS
4. Conclusions
ParTy Corpus

- Subtitles of films and TED talks
- Massively parallel corpus in 15 and more languages
- Publicly available from [www.natalialevshina.com/corpus.html](http://www.natalialevshina.com/corpus.html)
- Aligned with English (from srt > XML > aligned bitexts in txt)
- No special software required
- Constantly updated
- Informal language
An example of .srt format

268
00:33:22,546 --> 00:33:24,109
- Here, hold this. - Yeah, sure.

269
00:33:25,548 --> 00:33:29,219
You must be so excited.

270
00:33:31,080 --> 00:33:32,668
Are you freaking out?

271
00:33:32,703 --> 00:33:33,740
- Yeah... - Yeah?

272
00:33:35,981 --> 00:33:36,814
Oh, it's okay.
Validation with $n$-grams

From Levshina, Submitted
Film selection
Data set

- 347 causative situations, where at least one language has an analytic causative
- All translations coded for the type of causative: analytic or lexical
- All causative SITUATIONS coded for semantic and syntactic parameters (based on contextual information, including the visual information from the films)
## Dixon’s parameters as variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Values</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Aktionsart of the caused event</td>
<td>CausedEvent</td>
<td>‘NonAction’</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Action’</td>
<td>analytic</td>
</tr>
<tr>
<td>2 Number of main participants</td>
<td>NoPart</td>
<td>‘2’</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘3’</td>
<td>analytic</td>
</tr>
<tr>
<td>3 Control of Causee</td>
<td>CeControl</td>
<td>‘Yes’</td>
<td>analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td>lexical</td>
</tr>
<tr>
<td>4 Making or letting</td>
<td>MakeLet</td>
<td>‘Make’</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Let’</td>
<td>analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Undef’</td>
<td></td>
</tr>
<tr>
<td>5 Causer acting directly</td>
<td>CrDirect</td>
<td>‘Yes’</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td>analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Undef’</td>
<td></td>
</tr>
<tr>
<td>6 Causer acting intentionally</td>
<td>CrIntent</td>
<td>‘Yes’</td>
<td>lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td>analytic</td>
</tr>
<tr>
<td>7 Causer acting forcefully</td>
<td>CrForce</td>
<td>‘Yes’</td>
<td>analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td>lexical</td>
</tr>
<tr>
<td>8 Causer involved in caused event</td>
<td>CrInvolved</td>
<td>‘Yes’</td>
<td>no clear expectations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
</tbody>
</table>
## Dixon’s parameters as variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Values</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aktionsart of the caused event</td>
<td>CausedEvent</td>
<td>‘NonAction’</td>
<td>lexical analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Action’</td>
<td></td>
</tr>
<tr>
<td>Number of main participants</td>
<td>NoPart</td>
<td>‘2’</td>
<td>lexical analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘3’</td>
<td></td>
</tr>
<tr>
<td>Control of Causee</td>
<td>CeControl</td>
<td>‘Yes’</td>
<td>analytic lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
<tr>
<td>Making or letting</td>
<td>MakeLet</td>
<td>‘Make’</td>
<td>lexical analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Let’</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Undef’</td>
<td></td>
</tr>
<tr>
<td>Causer acting directly</td>
<td>CrDirect</td>
<td>‘Yes’</td>
<td>lexical analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Undef’</td>
<td></td>
</tr>
<tr>
<td>Causer acting intentionally</td>
<td>CrIntent</td>
<td>‘Yes’</td>
<td>lexical analytic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Undef’</td>
<td></td>
</tr>
<tr>
<td>Causer acting forcefully</td>
<td>CrForce</td>
<td>‘Yes’</td>
<td>analytic lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
<tr>
<td>Causer involved in caused event</td>
<td>CrlInvolved</td>
<td>‘Yes’</td>
<td>no clear expectations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘No’</td>
<td></td>
</tr>
</tbody>
</table>
Additional variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Values</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>9     Semantics of Causer</td>
<td>CrSem</td>
<td>‘Anim’ ‘Inanim’</td>
<td>analytic lexical</td>
</tr>
<tr>
<td>10    Semantics of Causee</td>
<td>CeSem</td>
<td>‘Anim’ ‘Inanim’</td>
<td>analytic lexical</td>
</tr>
<tr>
<td>11    Coreferentiality of Causer with other main participants</td>
<td>Coref</td>
<td>‘Yes’ ‘No’</td>
<td>no clear expectations</td>
</tr>
<tr>
<td>12    Polarity</td>
<td>Polarity</td>
<td>‘Pos’ ‘Neg’</td>
<td>no clear expectations</td>
</tr>
</tbody>
</table>
Outline

1. Aims of the study
2. Data: ParTy corpus
3. Statistical analyses: random forests and MDS
4. Conclusions
Random forests

• A non-parametric hypothesis-testing technique based on permutation
• Based on classification trees (conditional inference trees)
• A viable alternative to logistic regression in situations of ‘small n, large p’, also with complex interactions
• Return variable importance: which variables help predict the use of lexical or analytic causative
• R package party
Varimp: Romance languages
Varimp: Germanic languages
Varimp: Slavic languages
Random forests: conclusions

• Overall, the iconicity parameters, especially the Causee control, tend to be quite prominent, especially in the Romance languages.

However...
- this does not hold for ALL languages
- in every language, MULTIPLE factors are significant.
Random forests: conclusions

• Overall, the iconicity parameters, especially the Causee control, tend to be quite prominent, especially in the Romance languages.
• However...
  - this does not hold for ALL languages
  - in every language, MULTIPLE factors are significant.
Comparing the languages

• Compare the variable importance ranks between pairs of languages
• If the ranks are similar, small distance; if the ranks are dissimilar, large distance
• Multidimensional scaling...
Multidimensional Scaling
MDS results

• The languages cluster mostly according to their genealogical relationships, with a few exceptions
• Romanian – language contact?
• West-East continuum?
Outline

1. Aims of the study
2. Data: ParTy corpus
3. Statistical analyses: random forests and MDS
4. Conclusions
Conclusions

- Iconicity in general is an important predictor. However, it is not the ONLY factor that explains the use of lexical and analytic causatives: the variation is multifactorial both cross-linguistically and within specific languages. Disentangling multiple parameters is only possible at the level of usage tokens and with the help of multivariate methods.
Conclusions

- Iconicity in general is an important predictor.
- However, it is not the ONLY factor that explains the use of lexical and analytic causatives: the variation is multifactorial both cross-linguistically and within specific languages.
Conclusions

- Iconicity in general is an important predictor.
- However, it is not the ONLY factor that explains the use of lexical and analytic causatives: the variation is multifactorial both cross-linguistically and within specific languages.
- Disentangling multiple parameters is only possible at the level of usage tokens and with the help of multivariate methods.
Thank you!

The slides and corpus are available at

www.natalialevshina.com

Questions or suggestions:

natalevs@gmail.com