



UCL
Université
catholique
de Louvain



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

Study	Less integrated/compact causative	More integrated/compact causative
Givón (1980)	Lower degree of semantic binding between 2 events	Higher degree of semantic binding between 2 events
Comrie (1981; 1989)	Indirect causation	Direct causation
Haiman (1983)	Greater conceptual distance between Cause and Result	Smaller conceptual distance between Cause and Result
Givón (1990)	Human-Agentive Manipulee	Inanimate Manipulee

Analytic vs. lexical causatives

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



b. She *brought* her cat back.



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

**Related to
Caused event**

Relating to
Causee

Relating to
Causer

Action
(Di)transitive

Control
Unwilling ('make')

Fully affected

Indirect

Accidental

With effort, violence

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

Relating to
VERB

**Related to
Causee**

Relating to
Causer

Action

(Di)transitive

Control

Unwilling ('make')

Fully affected

Indirect

Accidental

With effort, violence

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

Relating to
VERB

Relating to
Causee

Related to
Causer

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?

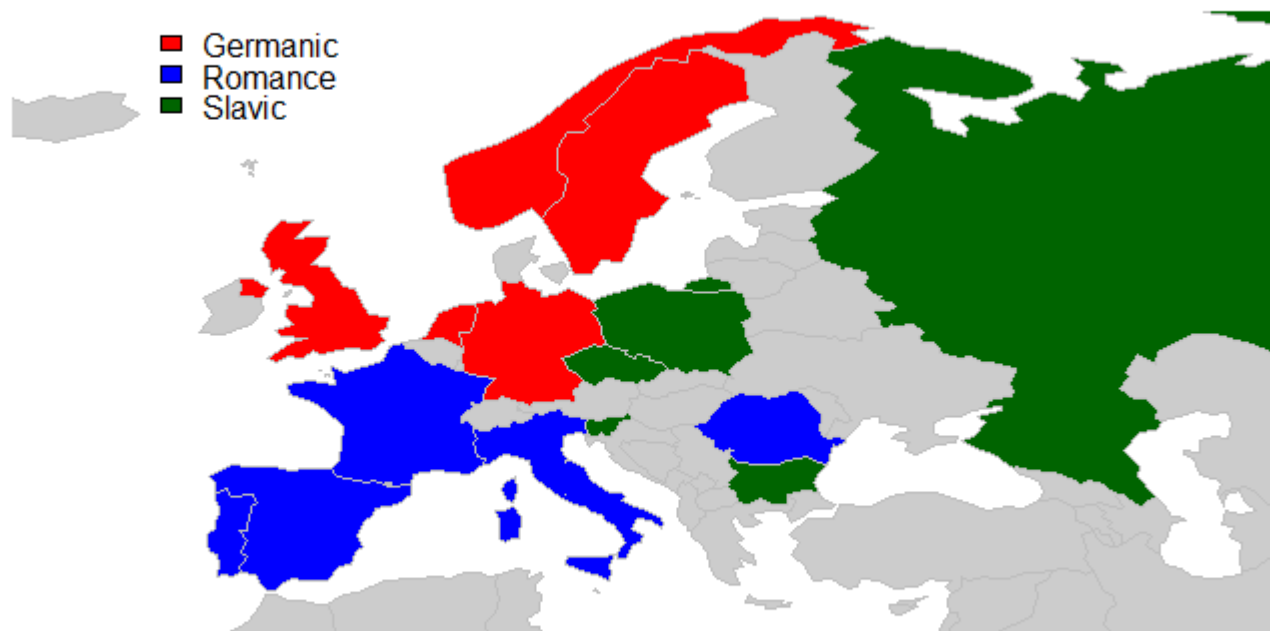
Cross-linguistically

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

Languages



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
- 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

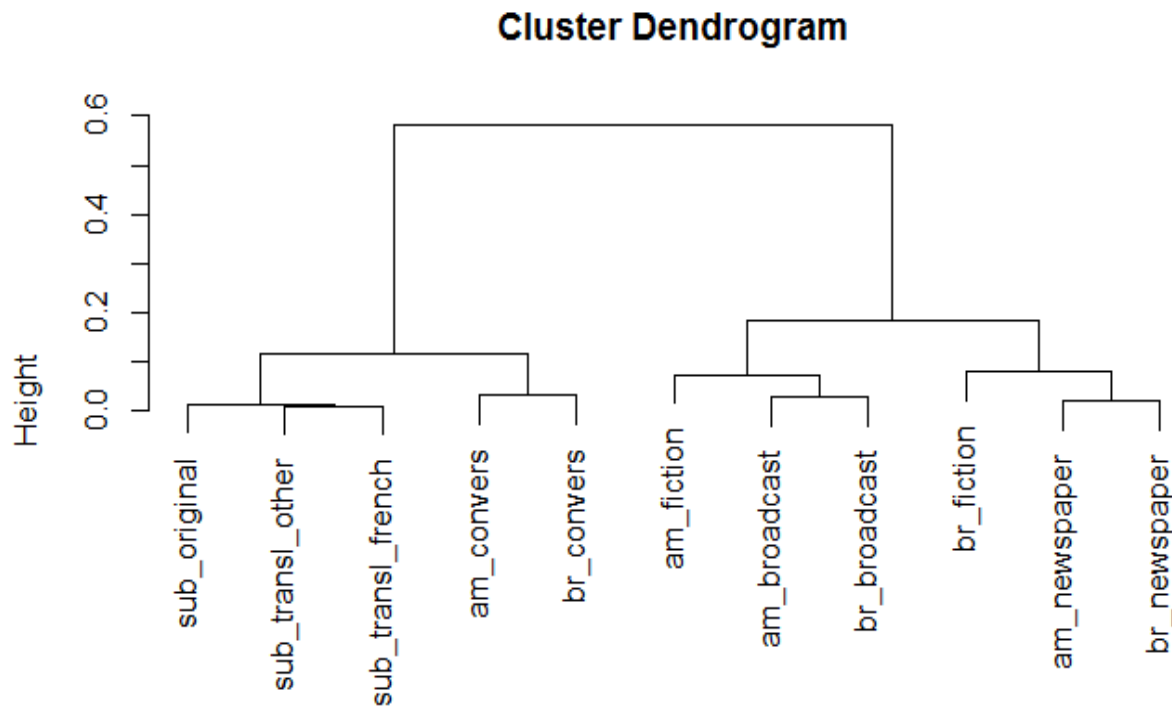
- Yeah... - Yeah?

272

00:33:35,981 --> 00:33:36,814

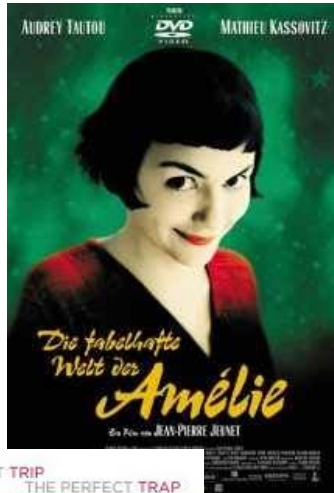
Oh, it's okay.

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

	Variable	Abbreviation	Values	Expectations
1	Aktionsart of the caused event	CausedEvent	'NonAction' 'Action'	lexical analytic
2	Number of main participants	NoPart	'2' '3'	lexical analytic
3	Control of Causee	CeControl	'Yes' 'No'	analytic lexical
4	Making or letting	MakeLet	'Make' 'Let' 'Undef'	lexical analytic -
5	Causer acting directly	CrDirect	'Yes' 'No' 'Undef'	lexical analytic
6	Causer acting intentionally	CrIntent	'Yes' 'No'	lexical analytic
7	Causer acting forcefully	CrForce	'Yes' 'No'	analytic lexical
8	Causer involved in caused event	CrInvolved	'Yes' 'No'	no clear expectations

Dixon's parameters as variables

	Variable	Abbreviation	Values	Expectations
1	Aktionsart of the caused event	CausedEvent	'NonAction' 'Action'	lexical analytic
2	Number of main participants	NoPart	'2' '3'	lexical analytic
3	Control of Causee	CeControl	'Yes' 'No'	analytic lexical
4	Making or letting	MakeLet	'Make' 'Let' 'Undef'	lexical analytic -
5	Causer acting directly	CrDirect	'Yes' 'No' 'Undef'	lexical analytic
6	Causer acting intentionally	CrIntent	'Yes' 'No'	lexical analytic
7	Causer acting forcefully	CrForce	'Yes' 'No'	analytic lexical
8	Causer involved in caused event	CrInvolved	'Yes' 'No'	no clear expectations

**Main
Iconic
factors**

Additional variables

	Variable	Abbreviation	Values	Expectations
9	Semantics of Causer	CrSem	'Anim' 'Inanim'	analytic lexical
10	Semantics of Causee	CeSem	'Anim' 'Inanim'	analytic lexical
11	Coreferentiality of Causer with other main participants	Coref	'Yes' 'No'	no clear expectations
12	Polarity	Polarity	'Pos' 'Neg'	no clear expectations

Outline

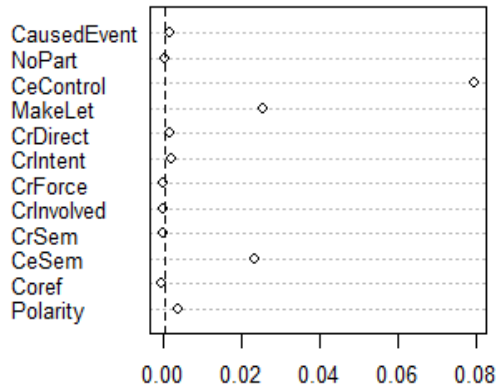
1. Aims of the study
2. Data: ParTy corpus
3. **Statistical analyses: random forests and MDS**
4. Conclusions

Random forests

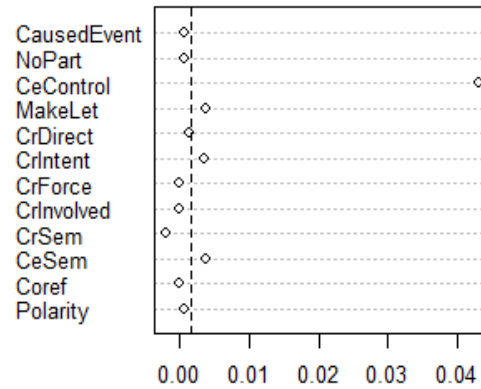
- A non-parametric hypothesis-testing technique based on permutation
- Based on classification trees (conditional inference trees)
- A viable alternative to logistic regression in situations of 'small n , large p ', also with complex interactions
- Return variable importance: which variables help predict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

Varimp: Romance languages

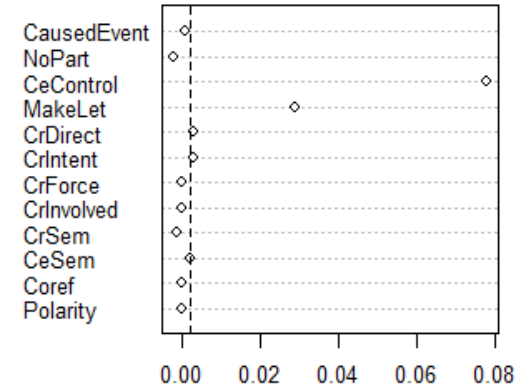
French



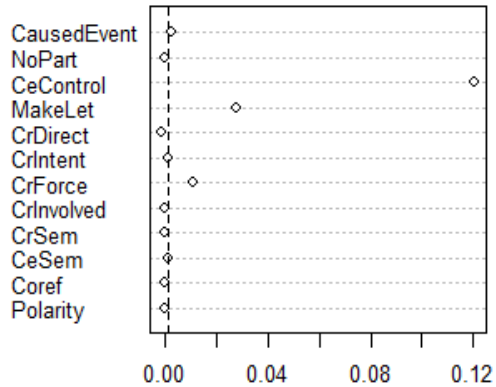
Italian



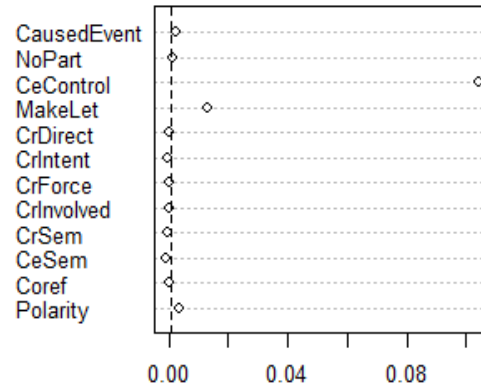
Spanish



Portuguese

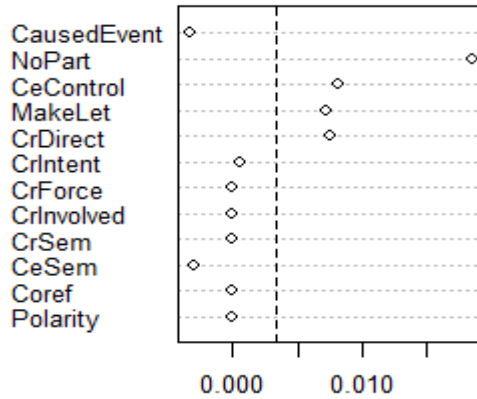


Romanian

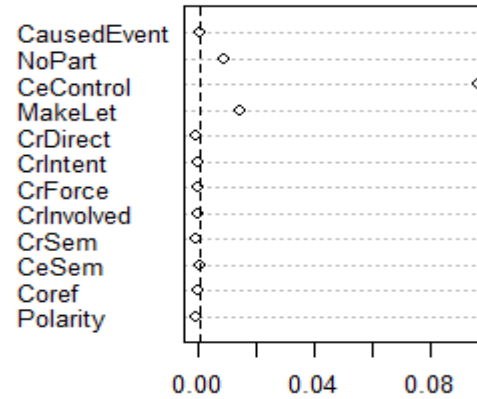


Varimp: Germanic languages

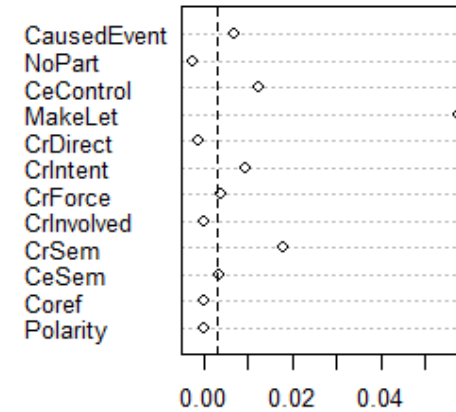
Dutch



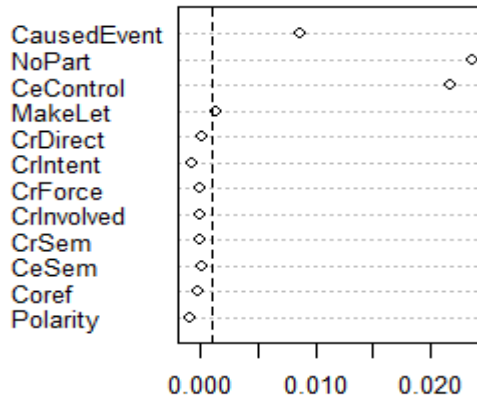
English



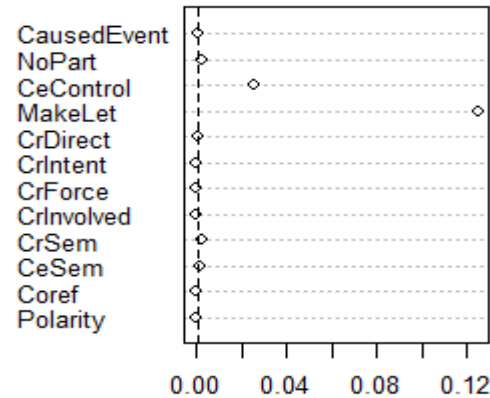
Swedish



German

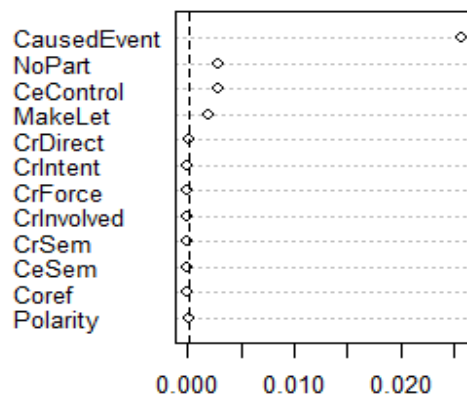


Norwegian

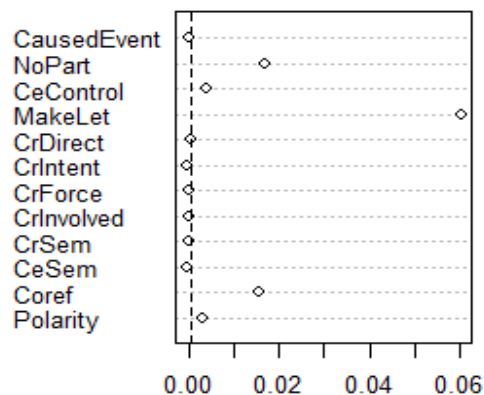


Varimp: Slavic languages

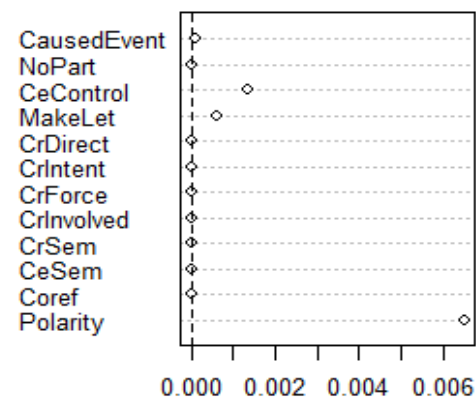
Bulgarian



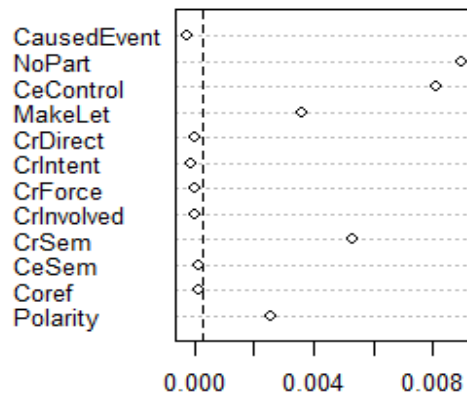
Czech



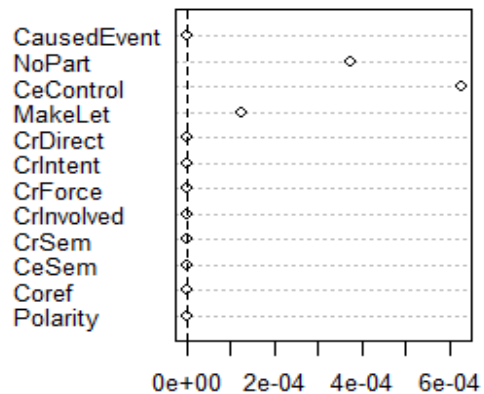
Slovenian



Polish



Russian



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.

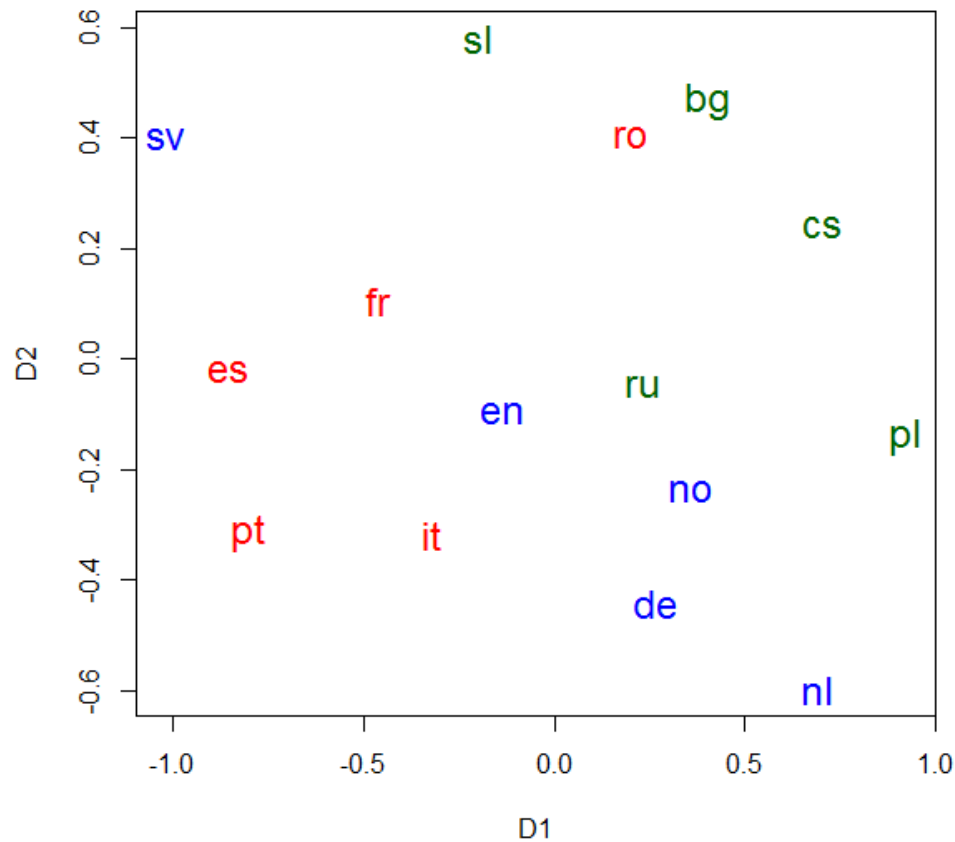
Random forests: conclusions

- 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.

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.
- 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