Факторный анализ

На примере статьи *Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages*

Общая идея

У нас есть некоторое грамматическое явление, которое имеет два вида маркирования. При этом нет четких правил, которые бы объясняли, когда используется тот или иной вариант.

Общая идея

У нас есть некоторое грамматическое явление, которое имеет два вида маркирования. При этом нет четких правил, которые бы объясняли, когда используется тот или иной вариант.

Например, когда употребляется лексический каузатив вроде *feed*, а когда *make eat*, или когда употребляется актив переходного глагола, а когда пассив с выраженным агентивным дополнением.

Как это реализуется: факторы

Можно выделить некоторое множество признаков, которые предположительно влияют на вероятность выбора той или иной конструкции в заданном контексте.

А затем оценить

a) в какой мере каждый из признаков влияет на выбор
b) какие из признаков значимы

Как это реализуется: методы

Логистическая регрессия

- Признакам приписываются коэффициенты, знак +/- говорит нам том, на вероятность появления какой из конструкций он влияет

Как это реализуется: методы

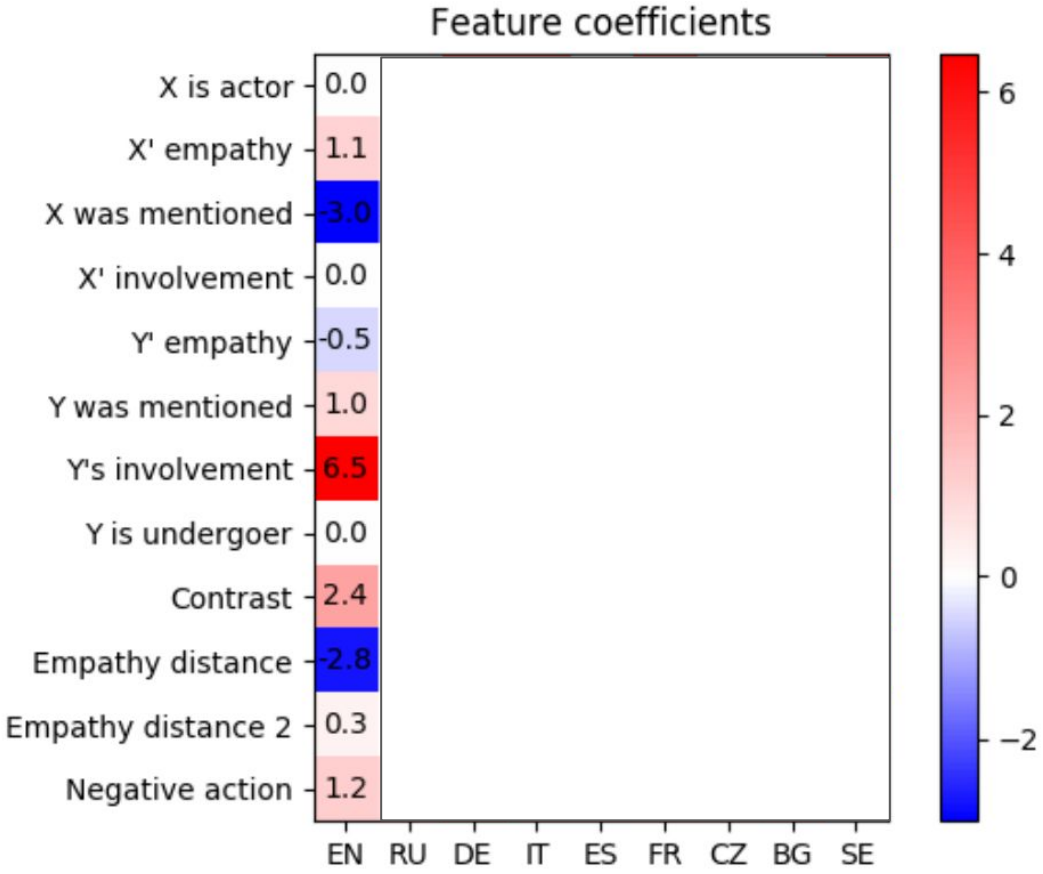
Логистическая регрессия

- Признакам приписываются коэффициенты, знак +/- говорит нам том, на вероятность появления какой из конструкций он влияет
- Показатель p-value указывает, является ли признак значимым

Как это реализуется: методы

Логистическая регрессия

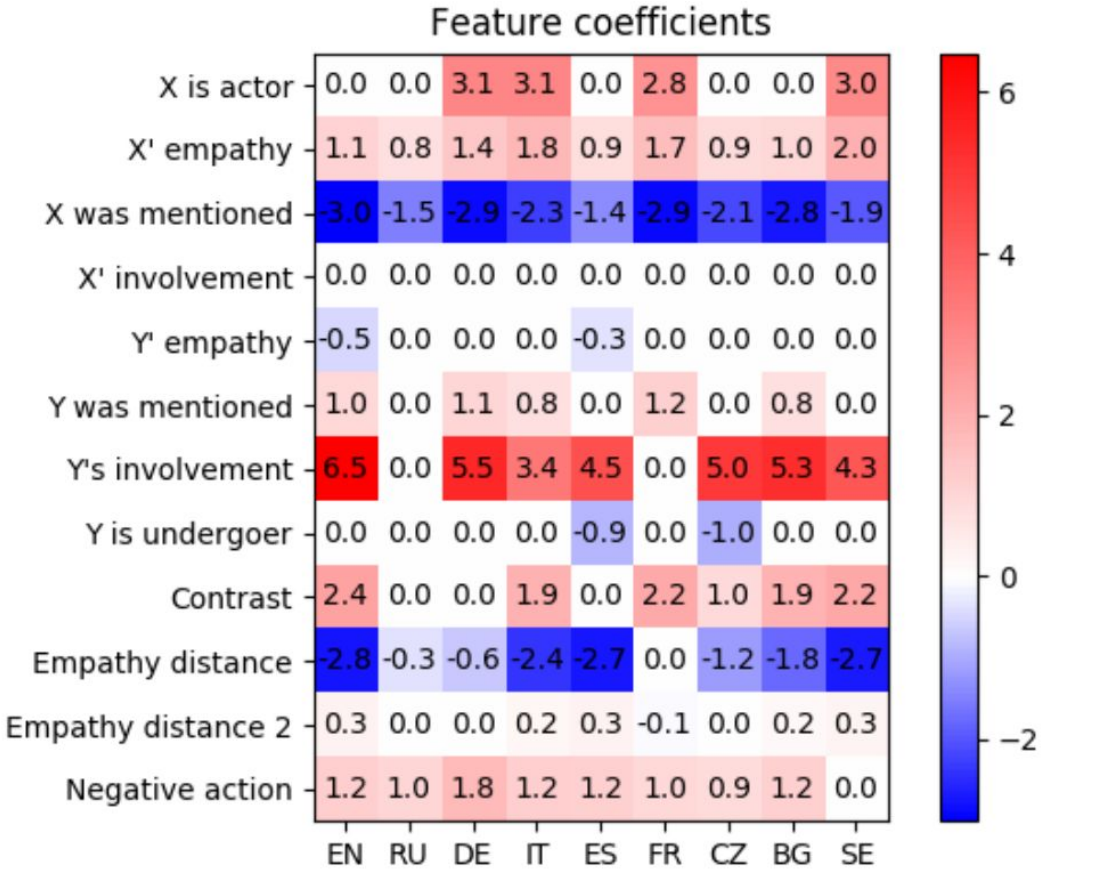
- Признакам приписываются
  вероятность появления как
- Показатель p-value указыва



Как это реализуется: методы

Логистическая регрессия

- Признакам приписываются
  вероятность появления как
- Показатель p-value указыва



Как это реализуется: методы

Логистическая регрессия

- Признакам приписываются коэффициенты, знак +/- говорит нам том, на вероятность появления какой из конструкций он влияет
- Показатель p-value указывает, является ли признак значимым

Random forest

Есть распределение важности признаков и значение, по которому отсекаются значимые и незначимые признаки

Natalia Levshina 2016

Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages

- В этой статье рассматривается вариативность употребления лексических и аналитических каузативов в 15 европейских языках из германской, романской и славянской группы, в качестве материала используется корпус субтитров к фильмам.

Natalia Levshina 2016

Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages

- В этой статье рассматривается вариативность употребления лексических и аналитических каузативов в 15 европейских языках из германской, романской и славянской группы, в качестве материала используется корпус субтитров к фильмам.
- Используя список типологических параметров и статистические методы анализа, автор работы выясняет, какие из параметров релевантны для выбора типа каузативов.

Типы каузативов

Согласно известной классификации, отсылающей нас к работе Комри (например, 1981: Ch. 8), каузативы можно поделить на три большие группы:

analytic – morphological – **lexical**

a. The sheriff caused Bill to die.

b. The sheriff killed Bill.

Аналитические каузативы представляют собой менее семантически интегрированные события.

Подход Dixon

Согласно подходу Диксон, степень компактности коррелирует с семантическими и синтаксическими признаками. Если в языке есть две формы каузативов (более компактная и менее компактная), они будут отличаться по признакам.

Компактность уменьшается в направлении от лексических к преифрастическим:

– Lexical causatives, e. g., breakTR or walkTR;

– Morphological causatives, e. g., tone change, reduplication, or affixation;

– Complex predicates, e. g., serial verbs, French faire ‘make’ + V INF, or causative particles;

– Periphrastic causatives, where the causatives are represented by verbs that belong to separate clauses, e. g., French laisser ‘let’ + NP + VINF or Portuguese fazer ‘make’ + (NP) + VINF.

Подход Dixon: параметры (факторы)

Table 1: Dixon's (2000) semantic and syntactic parameters of variation of causative constructions.

	More compact forms	Less compact forms
1.	Non-causal verb describing a state	Non-causal verb describing an action
2.	Intransitive (or intransitive and simple transitive) non-causal verb	Transitive (ditransitive) non-causal verb
3.	Causee lacking control	Causee having control
4.	Causee willing ("let")	Causee unwilling ("make")
5.	Causee partially affected	Causee fully affected
6.	Direct causation	Indirect causation
7.	Intentional causation	Accidental causation
8.	Causation occurring naturally	Causation occurring with effort

Извлечение контекстов

- Было найдено 325 каузативных ситуаций.
- Из них только 42.7% содержат либо аналитические, либо лексические каузативы.
  То есть большинство случаев было переведено как-то иначе. (звучит, как что-то знакомое)
- В рамках этой статьи такие случаи не рассматриваются. Почему?

Количество аналитических и лексических каузативов

Table 2: The frequencies of analytic and lexical causatives in the data set.

Genus	Language	Analytic	Lexical	Total
Germanic	Dutch	82	53	135
	English	96	72	168
	German	71	71	142
	Norwegian	87	67	171
	Swedish	68	62	130
Romance	French	116	57	173
	Italian	118	43	164
	Portuguese	75	71	179
	Romanian	71	79	150
	Spanish	77	63	140
Slavic	Bulgarian	43	86	129
	Czech	60	81	141
	Polish	25	72	97
	Russian	28	82	110
	Slovenian	28	77	105
Total		*1,045*	*1,036*	*2,081*

Признаки

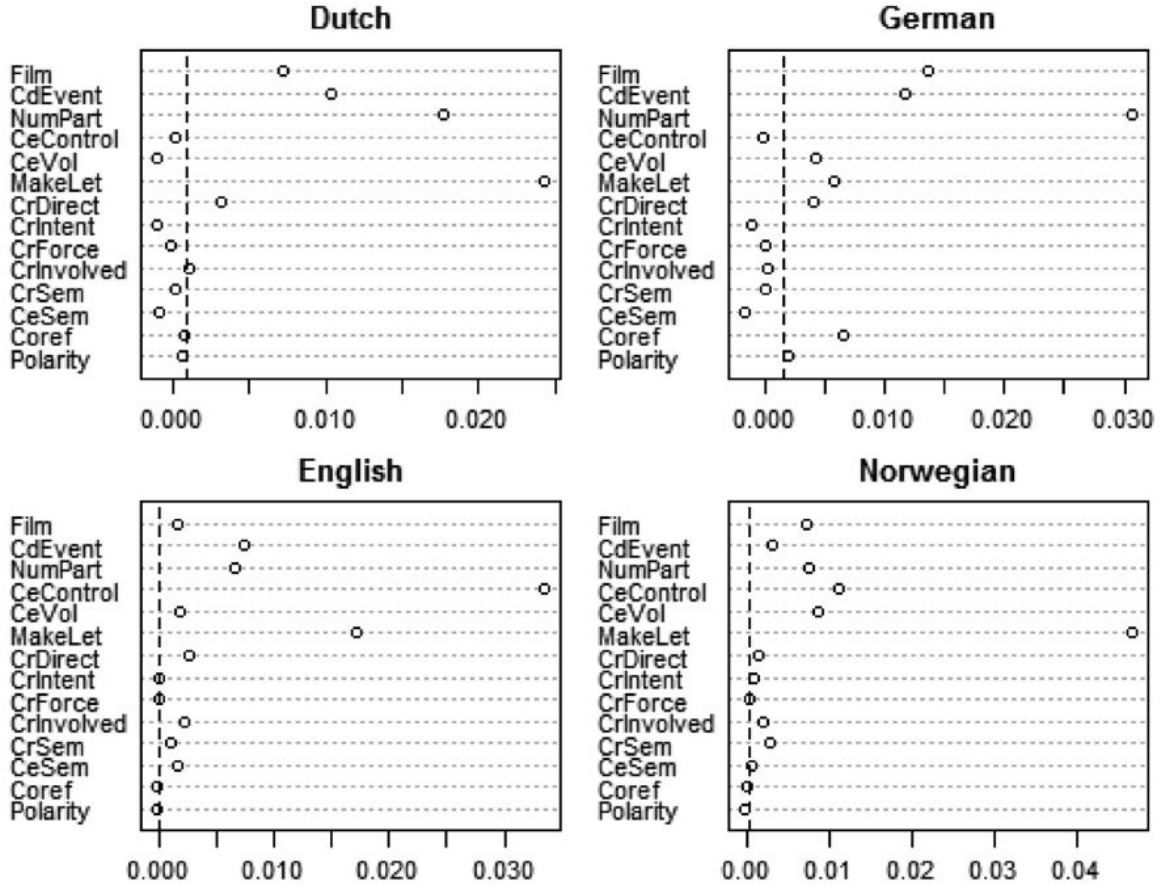
Table 3: Parameters of variation of lexical and analytic causatives operationalized as variables.

	Variable	Abbreviation	Values	Expectations
1	Aktionsart of the caused event	*CdEvent*	“Nonaction”	Lexical
			“Action”	Analytic
2	Number of main participants	*NumPart*	“2”	Lexical
			“3”	Analytic
3	Control of Causee	*CeControl*	“Yes”	Analytic
			“No”	Lexical
4	Causee acting willingly	*CeVol*	“Yes”	Analytic
			“No”	Lexical
			“Undef”	No clear expectations
5	Making or letting	*MakeLet*	“Make”	Lexical
			“Let”	Analytic
6	Causer acting directly	*CrDirect*	“Yes”	Lexical
			“No”	Analytic
7	Causer acting intentionally	*CrIntent*	“Yes”	Lexical
			“No”	Analytic
8	Causer acting forcefully	*CrForce*	“Yes”	Analytic
			“No”	Lexical
9	Causer involved in caused event	*CrInvolved*	“Yes”	No clear expectations
			“No”	
10	Semantics of Causer	*CrSem*	“Anim”	Lexical
			“Inanim”	Analytic
11	Semantics of Causee	*CeSem*	“Anim”	Analytic
			“Inanim”	Lexical
12	Coreferentiality of Causer with other main participants	*Coref*	“Yes”	No clear expectations
			“No”	
13	Polarity	*Polarity*	“Pos”	No clear expectations
			“Neg”	

Random Forest и особенности применения

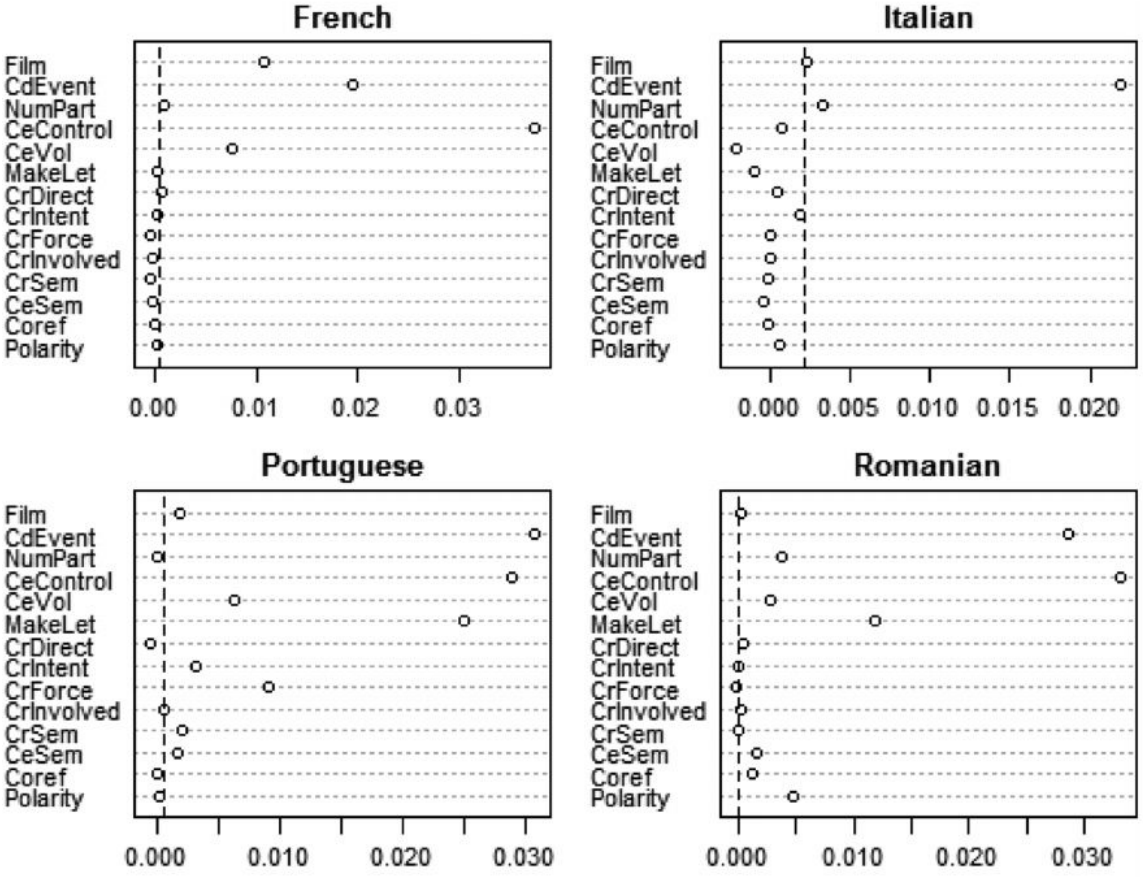
This technique is a perfect solution in our case. First, the data contain too many predictors in comparison with the number of observations (in particularthe frequencies of analytic causatives in some languages are too low for a multiple logistic regression analysis). Second, many of these semantic variables are strongly associated, as was shown in Section 4, e. g., the animacy and control of the Causee, the number of participants and the semantic properties of the caused event. In linguistics, random forests have been successfully applied in variationist studies (Tagliamonte and Baayen 2012).

Результаты

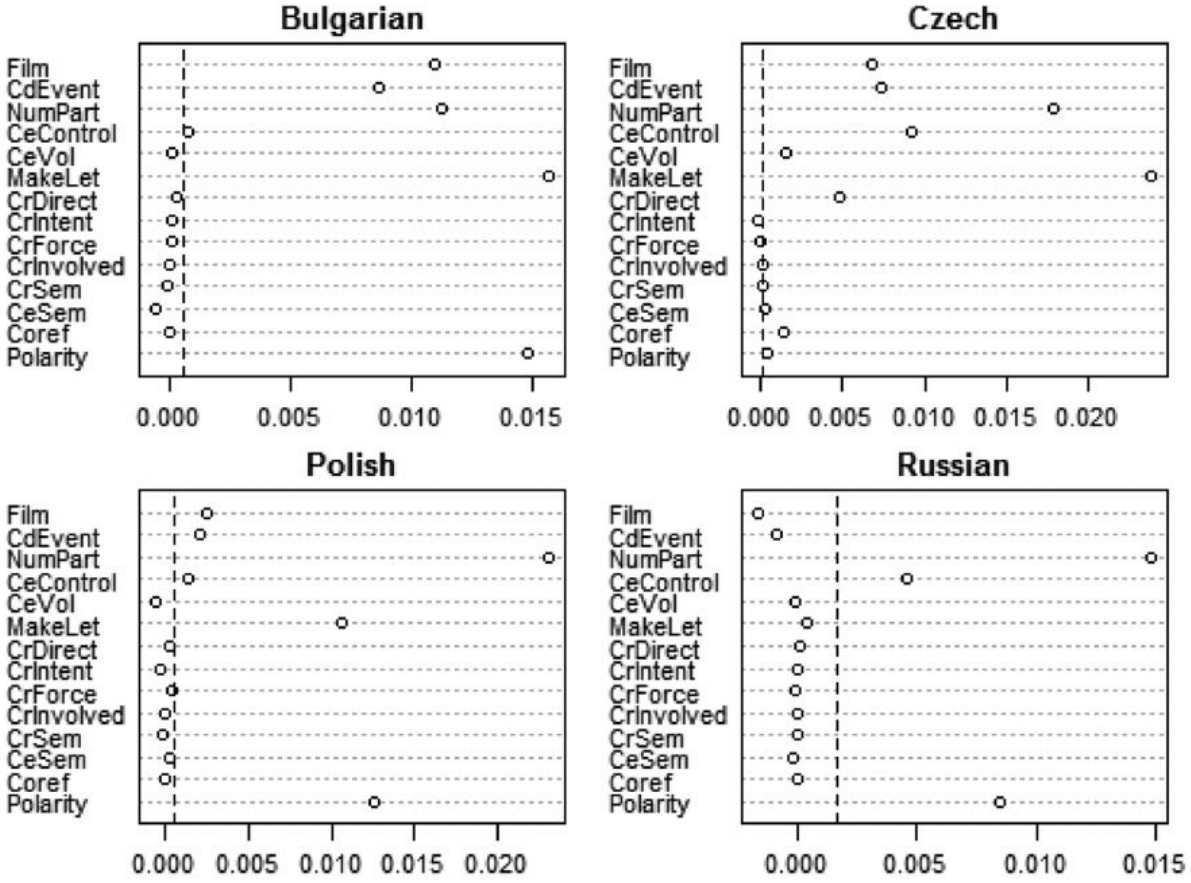


Можно смотреть языки в отдельности, а можно группировать близко родственные

Результаты


French
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.00
0.01
0.02
0.03
Italian
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.005
0.010
0.015
0.020
Portuguese
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.010
0.020
0.030
Romanian
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.010
0.020
0.030

Результаты


Bulgarian
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.005
0.010
0.015
Czech
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.005
0.010
0.015
0.020
Polish
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.005
0.010
0.015
0.020
Russian
Film
CdEvent
NumPart
CeControl
CeVol
MakeLet
CrDirect
CrIntent
CrForce
CrInvolved
CrSem
CeSem
Coref
Polarity
0.000
0.005
0.010
0.015

Результаты

Романские языки

- Наиболее однородны
- *Во всех* языках самые важные — Actions и Nonactions (CdEvent), а также контроль над Causee (CeControl).
- В *большинстве* языков: MakeLet.
- *Португальский и испанский:* intentionality of the Causer (CrIntent)
- *Португальский*: forcefulness of causation (CrForce)
- *Румынский*: Polarity

Результаты

Германские языки

- Довольно неоднородны
- Наиболее выдающиеся признаки: making and letting, as well as the number of participants, directness of causation, control and volitionality of the Causee
- Английский: control of the Causee
- Немецкий: number of participants
- Шведский: making and letting (MakeLet)
- Норвежский и датский: больше всего расхождений

Результаты

Славянские языки

- Общая картина: control of the Causee (CeControl), making vs. letting (MakeLet), number of participants (NumPart), и Actions vs. Nonactions (CdEvent).
- Много вариативности
  В болгарском и чешском: Making vs. letting (MakeLet) самый важный фактор
- В русском и польском: the number of participants (NumPart).
- Факторы, связанные с характеристиками каузатора не важны во всех языках

Что следует учесть

Мы работаем с мультиязычными параллельными данными. На что это влияет?

С одной стороны, мы можем воспользоваться свойством параллельности и облегчить процесс разметки.

Что следует учесть

Мы работаем с мультиязычными параллельными данными, на что это влияет.

С одной стороны, мы можем воспользоваться свойством параллельности и облегчить процесс разметки.

С другой стороны, в такого рода экспериментах мы не используем все-все примеры, которые нам попались. На что это влияет? На размер выборки по каждому языку, а также ее сбалансированность.

Что следует учесть

Мы работаем с м
влияет.

С одной стороны
облегчить проце

С другой сторон
примеры, которы
каждому языку.

Table 2: The frequencies of analytic and lexical causatives in the data set.

Genus	Language	Analytic	Lexical	Total
Germanic	Dutch	82	53	135
	English	96	72	168
	German	71	71	142
	Norwegian	87	67	171
	Swedish	68	62	130
Romance	French	116	57	173
	Italian	118	43	164
	Portuguese	75	71	179
	Romanian	71	79	150
	Spanish	77	63	140
Slavic	Bulgarian	43	86	129
	Czech	60	81	141
	Polish	25	72	97
	Russian	28	82	110
	Slovenian	28	77	105
Total		*1,045*	*1,036*	*2,081*

Что следует учесть

Мы работаем с мультиязычными параллельными данными, на что это влияет.

С одной стороны, мы можем воспользоваться свойством параллельности и облегчить процесс разметки.

С другой стороны, в такого рода экспериментах мы не используем все-все примеры, которые нам попались. На что это влияет? На размер выборки по каждому языку.

Размер и сбалансированность выборки могут сильно влиять на результаты.

Это может быть просто необходимость подобрать определенный метод или настроить процесс построения модели, а может быть модель просто не построится или результаты будут не интерпретируемые.

Что следует учесть

Выбирая признаки, нужно иметь в виду, что мы не делаем морфологический анализатор.

Признаки не должны явно указывать на конструкцию
Признаки (должны) быть проецируемыми на все языки, то есть, например, признак “дательный падеж у дополнения” не подходит.