



ЛТИЯ ИЛИ РАН

24 мая 2022 г.

# NoCaCoDa

## (Nominal Causal Construction Database)

### Структура и предварительные результаты

Е. А. Забелина ([windmill161@gmail.com](mailto:windmill161@gmail.com))

Н. Н. Логвинова ([natalielo009@gmail.com](mailto:natalielo009@gmail.com))

С. С. Сай ([serjozhka@yahoo.com](mailto:serjozhka@yahoo.com))

Исследование поддержано грантом РФФ «Причинные конструкции в языках мира: семантика и типология» (№ 18-18-00472, руководитель – В. С. Храковский)

# План доклада

I) Структура базы (напоминание)

II) Результаты про контексты (= семантику)

III) Результаты про языки



# Часть I

## Структура базы (напоминание)

# Основной принцип

- Собираются переводы конкретного набора предложений-стимулов на язык X

– Цель: сопоставимость данных, ср.:

«Как в Вашем языке выражается непосредственная внутренняя объективная причина?» => навязываются ожидания о релевантных признаках, они могут быть неверны

vs.

«Как на Вашем языке сказать ‘Женщина дрожит от холода’?» => исследование отталкивается от осязаемых сопоставимых фактов

# Структура базы (напоминание)

- На входе анкета из 54 предложений (русская или английская версия)
- Стимулы составлены с учетом семантических параметров, известных из литературы, и заранее размечены по ним

# Заложенные семантические параметры

- **Тип**

- онтологическая: *Платок намок <от> [слез].*
- эпистемическая: *Ночью шел дождь <?> [лужи].*
- иллокутивная: *Заходи в дом <?> [холод].*
- логическая: *Это число не может быть меньше единицы <в силу> [математических законов].*

- **Объективная или субъективная**

- объективная: *Ветка сломалась <от> [тяжести].*
- субъективная: *Женщина зашла внутрь <из> [любопытства].*

- **Непосредственная или опосредованная**  
(direct vs. indirect)
  - непосредственная: *Мужчина кричит <от> [боли].*
  - опосредованная: *Женщина не может заснуть <из-за> [комаров].*
- **Внутренняя или внешняя**  
(internal vs. external)
  - внутренняя: *Ребенок заплакал <от> [страха].*
  - внешняя: *Дети подрались <из-за> [мяча].*
- **Таксисные отношения**  
(simultaneous vs. anterior)
  - одновременность: *Женщина дрожит <от> [холода].*
  - предшествование: *У мужчины <от> [пива] болит голова.*

# Вариативность

- Возможна регистрация более одного перевода
- Ровно **один** перевод для каждого стимула помечается как основной (**main**), число альтернативных (**alternate**) не ограничено
- Критерии включения в базу:
  - причина выражена **именной группой**
  - семантически она **максимально близка к данной в стимуле**
  - синтаксически она **не является субъектом** данного предложения



# Структура записи на выходе

- Номер языка
- Номер стимула
- Статус перевода: main vs. alternate
- Перевод на изучаемый язык
- Глоссы (русские и английские)
- Обратный перевод на русский и английский
- **Показатель именной причины**

# Показатели именной причины

- Показатель именной причины – основная «зависимая переменная», важный результат исследовательской работы
- Все показатели дополнительно размечаются
  - **морфосинтаксический** тип  
адлоги, аффиксы
  - **семантические** характеристики  
специализированный vs. разные типы полисемии
  - **эксплицитность** (пока не учитывалась)

# Типы полисемии показателей именной причины

- В чем идея?
  - Причина — абстрактное значение, очень часто осмысляется через что-то **конкретное**  
ср. в русском: *от холода, из вежливости, из-за дождя...* => начальная точка
  - Модели полисемии отражают **КОГНИТИВНЫЕ СХЕМЫ**, стоящие за осмыслением причины  
[Radden 1985; Dirven 1995; Palancar 2001; Yamaguchi 2004; Narrog 2010]
  - Разметка ведется по широким **макротипам**, это облегчает межъязыковое сравнение

# Типы полисемии показателей именной причины

## **Goal**

Конечная точка + Реципиент, Бенефициент, Цель

## **Instrument**

Инструмент + Средство, Путь, Паттерн, Копартиципант

## **Source**

Начальная точка + Временная точка отсчета, Предшествование

## **Location**

Место

## **Object**

Объект + Предмет обсуждения

## **Other**

например, Посессор, Стандарт сравнения...

# Типы полисемии показателей именной причины

- Семантические макротипы не исключают друг друга  
ср. англ. *at*: макротипы Location + Goal
- Dedicated (специализированные) показатели причины — те, у которых нет существенных непричинных значений  
ср. англ. *because of*, рус. *благодаря* и т.д.



# Текущее наполнение базы данных

<b>language_ru</b>	<b>family (WALS)</b>	<b>contributor_ru</b>
английский	Indo-European	Н. Н. Логвинова
японский	Japanese	А. А. Кузнецов
русский	Indo-European	С. С. Сай
французский	Indo-European	Е. Е. Корди, Е. А. Забелина
каталанский	Indo-European	Т. В. Репнина
ассирийский новоарамейский	Afroasiatic	Н. Н. Логвинова
литовский	Indo-European	Эмма Генюшене
гуро	Mande	О. В. Кузнецова
телугу	Dravidian	Е. А. Забелина
белорусский	Indo-European	Н. М. Заика
чувашский (малокарачкинский диалект)	Altaic	Н. М. Заика
арабский	Afroasiatic	Р. Г. Мамедшахов
кхмерский	Austro-Asiatic	С. Ю. Дмитренко
урду	Indo-European	Е. А. Забелина
удмуртский (южный говор)	Uralic	М. О. Черемисинова, Е. А. Забелина
словенский	Indo-European	С. С. Сай
шведский	Indo-European	В. П. Шувалова
лезгинский	Nakh-Daghestanian	Р. Г. Мамедшахов
латышский	Indo-European	А. Д. Даугавет
немецкий	Indo-European	Д. А. Алфимова
итальянский	Indo-European	Наккарато Мария Кьяра

# Как будет организована база данных

- Цель — создать онлайн-интерфейс
  - Благодарности **Дмитрию Николаеву**,  
университет Штутгарта





sentences	
sentence_no	integer
sentence_en	text
sentence_ru	text
type	text
objective_subjective	text
direct_indirect	text
internal_external	text
taxis	text

languages	
language_no	integer
language_ru	text
language_en	text
contributor_ru	text
contributor_en	text
consultant_ru	text
consultant_en	text
glottocode	text
data_collection_year	integer
macroarea	text
family_WALS	text
genus_WALS	text
latitude	real
longitude	real
contact_language	text
coding_responsibility	text
comments	text
markers_xlsx	text

stimulus\_no:sentence\_no

Arabic	
language_no	integer
language_name	text
stimulus_no	integer
status	text
causal_marker	text
sentence	text
glosses_en	text
back_translation_en	text
glosses_ru	text
back_translation_ru	text
comments	text
comments_informal	real
comments_informal	real
comments_informal	real

language\_no, language\_name:language\_no, null

causal\_marker

ArabicMarkers	
language_no	integer
language_name	text
causal_marker	text
morphosyntactic_type	text
explicitness	text
goal	text
instrument	text
location	text
object	text
source	text
other	text
other_details	real
other_details	real
other_details	real

# Часть II

Результаты про контексты,  
или что мы узнали о семантике причины

# Результаты про контексты

II.1 Когнитивные схемы

II.2 «Расстояния» между контекстами

II.3 Априорные семантические оппозиции

## II.1. Когнитивные схемы

КОГНИТИВНЫЕ СХЕМЫ» ≈ макротипы полисемии:  
способ осмысления причинной ситуации

## II.1. КОГНИТИВНЫЕ СХЕМЫ

Для каждого контекста (из 54) посчитано распределение **всех** полученных переводов по макротипам полисемии причинных показателей:

- Специализированные (Dedicated)
- Начальная точка (Source)
- Конечная точка (Goal)
- Инструмент (Instrument)
- Место (Location)
- Объект (Object)
- другое (Other)

## II.1. КОГНИТИВНЫЕ СХЕМЫ

- Основные и альтернативные варианты перевода учитывались наравне.
- Несколько значений одного показателя учитывались наравне.

## II.1. КОГНИТИВНЫЕ СХЕМЫ

- Пример: распределение для #1  
'Платок намок <от> [слез]. '

stimulus_no	dedicated, %	goal, %	instrument, %	location, %	object, %	source, %	other, %
1	12,5	3,1	31,3	9,4	0,0	40,6	3,1

=> в данном случае преобладают схемы Source и Instrument

# Заложенные семантические параметры

- Объективные или субъективные

	<b>dedicated</b>	<b>goal</b>	<b>instrument</b>	<b>location</b>	<b>object</b>	<b>source</b>	<b>other</b>
objective	0,28	0,11	0,22	0,08	0,04	0,25	0,03
subjective	0,27	0,16	0,19	0,08	0,08	0,20	0,01



# Заложенные семантические параметры

- Непосредственные или опосредованные

	<b>dedicated</b>	<b>goal</b>	<b>instrument</b>	<b>location</b>	<b>object</b>	<b>source</b>	<b>other</b>
direct	0,17	0,10	0,25	0,10	0,03	0,32	0,03
indirect	0,44	0,15	0,16	0,05	0,08	0,11	0,01

# Заложенные семантические параметры

- Внутренние или внешние

	<b>dedicated</b>	<b>goal</b>	<b>instrument</b>	<b>location</b>	<b>object</b>	<b>source</b>	<b>other</b>
internal	0,23	0,08	0,25	0,07	0,03	0,31	0,03
external	0,32	0,15	0,18	0,08	0,06	0,18	0,02

# Заложенные семантические параметры

- Предшествование или одновременность

	<b>dedicated</b>	<b>goal</b>	<b>instrument</b>	<b>location</b>	<b>object</b>	<b>source</b>	<b>other</b>
anterior	<b>0,49</b>	0,09	0,13	0,04	0,04	0,20	0,01
simultaneous	0,27	0,10	<b>0,22</b>	0,07	0,04	0,27	0,03

# Dedicated

## Специализированные показатели

- Самые частотные (около 30% всех вхождений).
- Связаны с **опосредованными** и, чуть менее сильно, с **внешними** причинами в таких контекстах, где есть неназванное промежуточное звено:
- #29 The woman was late for work <**because of**> [her husband]. {He forgot his car keys at home}.
- Практически полностью покрывают эпистемические контексты (~100%, но мало данных); ср. рус. *судя по*, англ. *judging by*.
- Скорее связаны с **предшествованием**, чем с одновременностью.

# Dedicated

## Специализированные показатели

- **Иконичность:** эти значения максимально далеки от чего-то осязаемого физического.
- Связь специализированных маркеров с опосредованными причинами до этого обсуждалась в [Say 2021] на славянском материале, связь опосредованностью и предшествованием – в докладе [Say 2019] на основе обзора грамматик.

# Source

## Начальная точка

- Такие показатели, как нем. *aus*, бел. *ад*
- Второй-третий по частотности тип в данных, около 22% случаев.
- Очень четко ассоциируются с **непосредственными**, среди них скорее с **внутренними**. Например:  
#38 ‘Во время войны люди <с> [голода] ели траву.’
- Связь схемы Source с непосредственными причинами до этого обсуждалась в [Say 2021:194ff] на славянском материале.

# Instrument

## Инструмент

- Такие показатели, как ит. *per*, англ. *by*
- Второй-третий по частотности тип в данных, около 20% случаев.
- Характерен для единственного “логического” контекста:
  - #54 ‘Это число не может быть меньше единицы <в силу> [математических законов].’

# Instrument

## Инструмент

- Как и Source, связан скорее с **непосредственными внутренними, особенно одновременными.**
- Связь этой модели полисемии с семантикой **одновременности** обсуждалась в докладе [Say 2019] на материале обзора грамматик.



# Goal

## Конечная точка

- Такие показатели, как фр. *pour*, рус. *за*.
- Четвертый по частотности тип (13%).
- Довольно пестрая группа контекстов.
- Много случаев, когда это **МОТИВ** осознанного действия:

#46 ‘Дети подрались <из-за> [мяча].’

- в частности, той или иной реакции агенса:

#45 ‘Мужчина побил соседа <за> [его слова].’

Ср. обсуждение на вторичном материале связи между этим типом полисемии и субъективной причиной в [Say 2019]

# Goal

## Конечная точка

- Довольно много глаголов с ЭМОТИВНЫМ КОМПОНЕНТОМ: ‘сердиться’, ‘хвалить’, ‘рассмеяться’.

Ср. обсуждение двойственной природы стимула эмоций (причина + цель) в [Croft, 1993; Klein & Kutscher, 2002; Verhoeven, 2007: 62] и т. д.

- Семантика отчасти близка к целевой (а ассоциация цели и пространственной конечной точки общеизвестна)

# Location

## Место

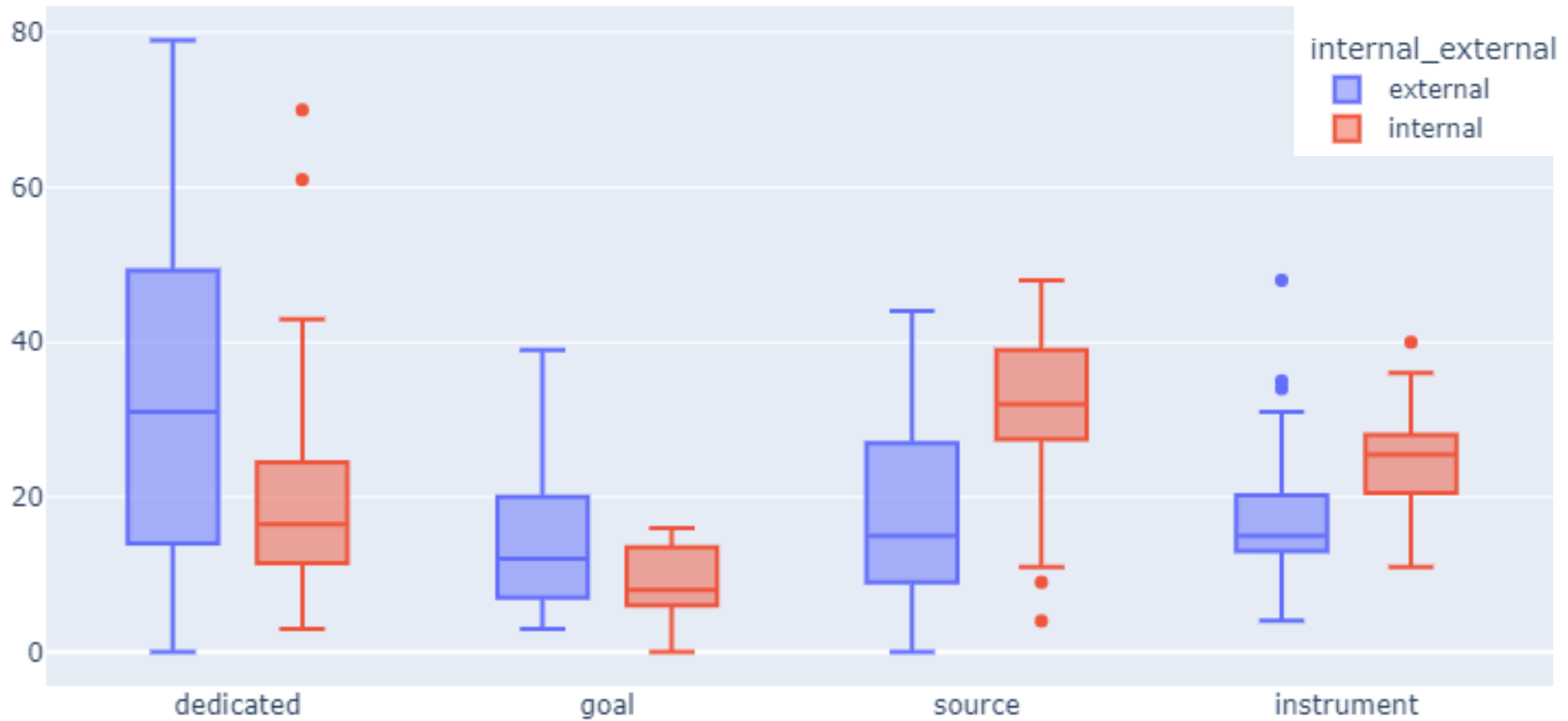
- Такие показатели, как лезг. адессивный падеж, слов. *pod\_INS*
- Общая частотность – 7%
- Но есть 11 контекстов, где этот тип покрывает 13-29%
- Нередко названия **эмоций, одновременных с действием**

#23 ‘Женщина <в> [страхе] схватила брата за руку.’

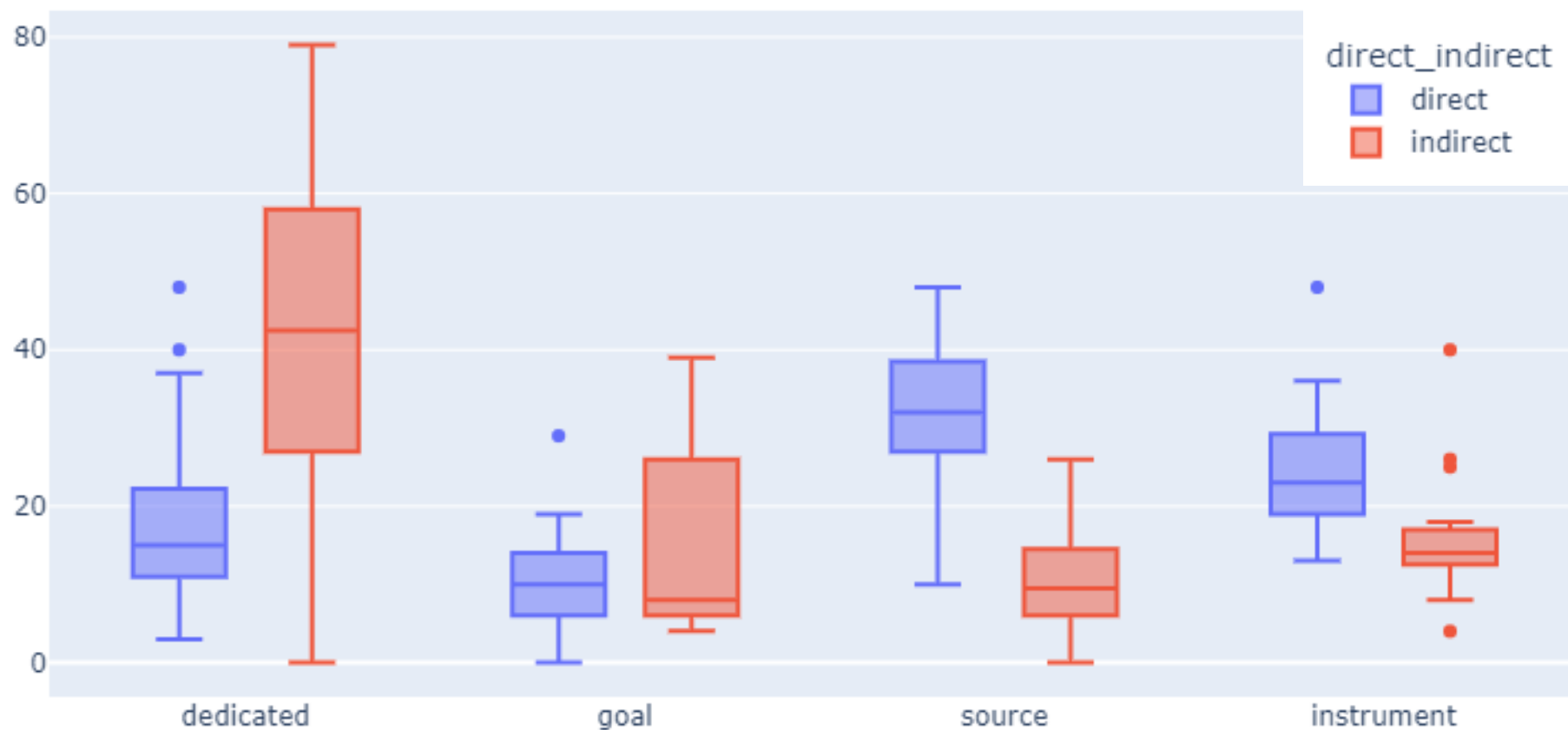
#42 ‘Женщина разбила всю посуду <от> [злости].’

=> пока мало данных для осязаемых выводов.

# Внешняя и внутренняя

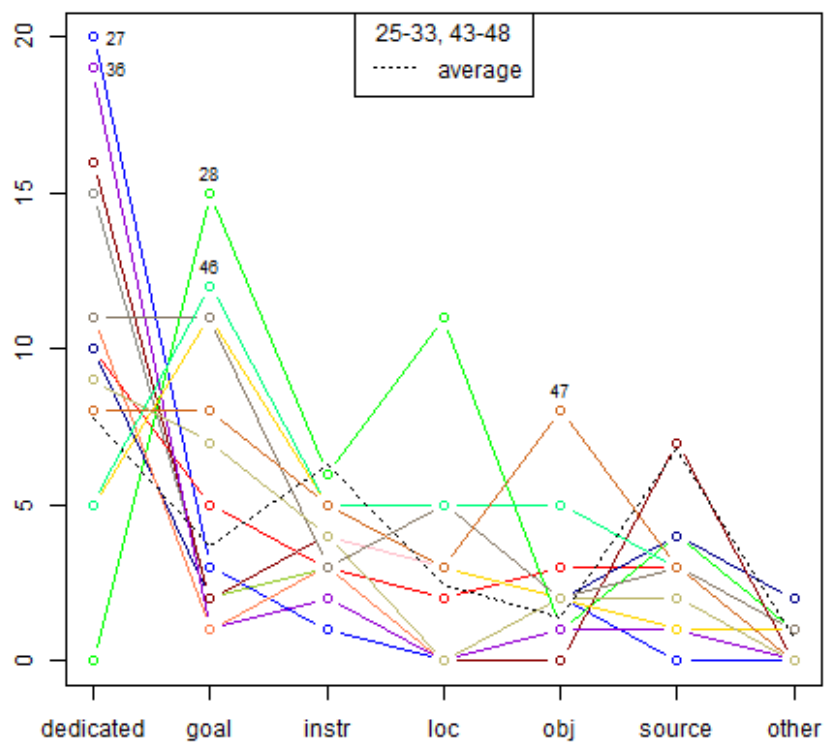


# Непосредственная и опосредованная

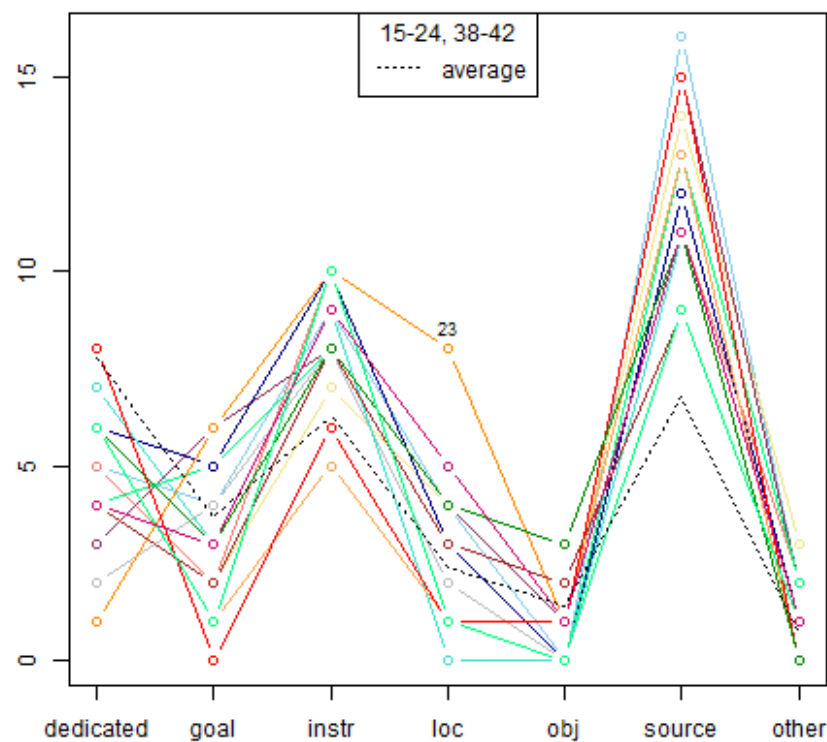


# Два противопоставленных семантических типа

Indirect external



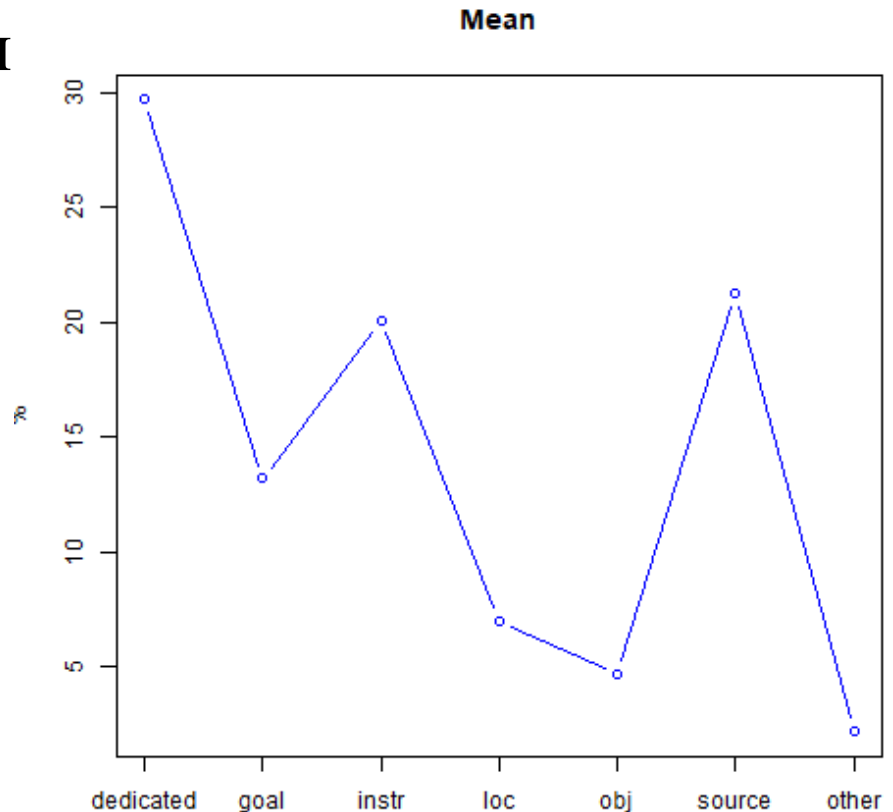
Direct internal



# II.1. КОГНИТИВНЫЕ СХЕМЫ: ВЫВОДЫ

Наиболее частотные типы  
в распределении:

1. Специализированные  
(Dedicated)
2. Начальная точка  
(Source)
3. Инструмент  
(Instrument)



# Распределение

	NoCaCoDa, %	вторичные данные в [Say 2019], %
dedicated	29,7	33,9
source	<b>21,3</b>	14,0
instrument	20,1	23,1
location	7,0	
goal	<b>13,2</b>	21,5
object	4,7	7,4
other	2,2	

- Распределение немного **отличается** от того, что было получено на ранее собранных вторичных данных
- Возможно, из-за несбалансированной выборки в NoCaCoDa: тут много (восточной) Европы



# II.1. КОГНИТИВНЫЕ СХЕМЫ: ВЫВОДЫ

- Из априорных семантических параметров на выбор когнитивной схемы при выражении причины больше всего влияют контрасты:
  - непосредственная vs. опосредованная (direct vs. indirect)
  - внутренняя vs. внешняя (external vs. internal)
- В распределении отчетливо выделяются два паттерна: непосредственные внутренние vs. опосредованные внешние

## II.1. КОГНИТИВНЫЕ СХЕМЫ: ВЫВОДЫ

- Начальная точка (Source) и Инструмент (Instrument) ассоциированы с **непосредственными внутренними** причинными контекстами.
- Специализированные показатели ассоциированы прежде всего с **опосредованными внешними**.
- Конечная точка (Goal), хоть и слабее, ассоциирована с **субъективными внешними** причинными контекстами.
- Все ассоциации **иконичны**.

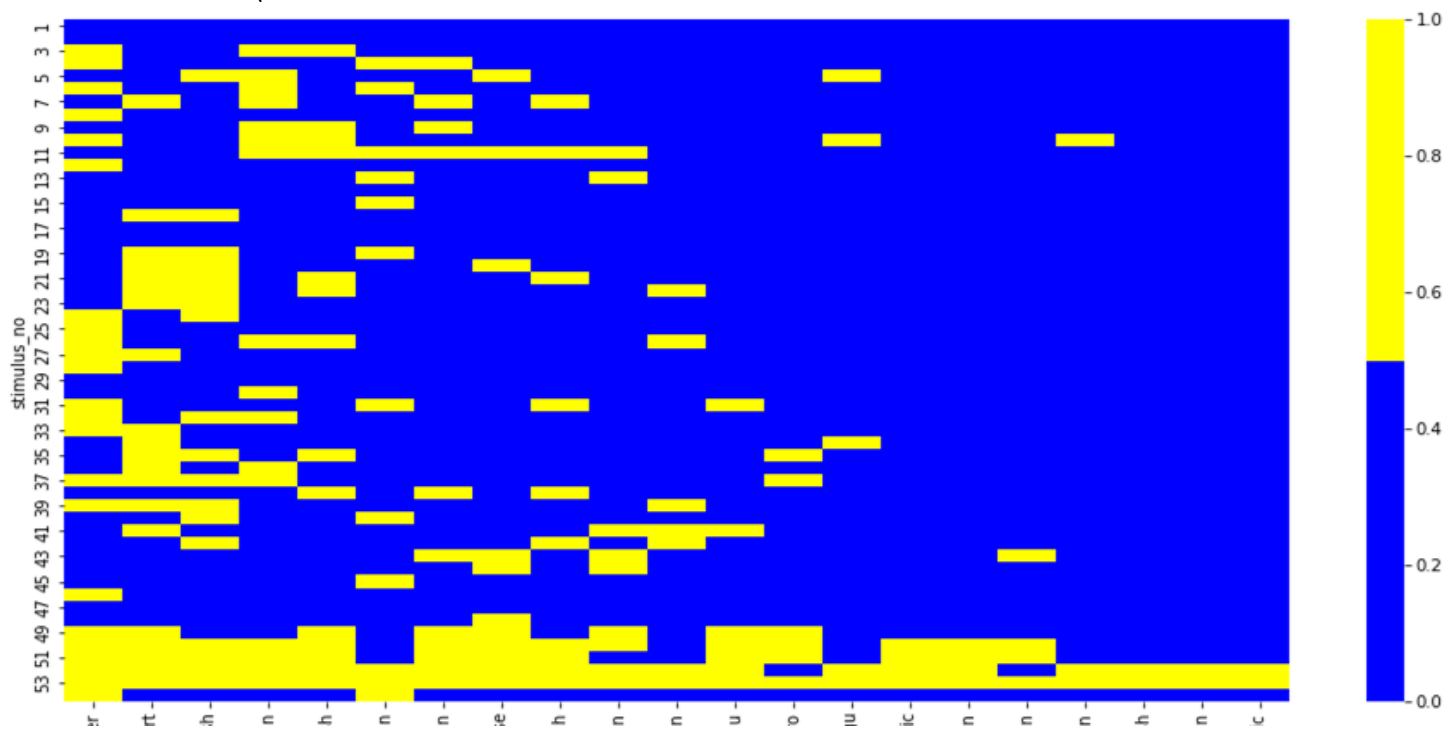
# Результаты про контексты

III.1 Когнитивные схемы

III.2 «Расстояния» между языками

## II.2. «Расстояния» между контекстами

- Оговорка: контексты #11, #49-53 исключены из дальнейшего рассмотрения, т.к. содержат < 70% известных данных, см. желтые сегменты на визуализации
  - строки – контексты
  - столбцы -- языки



## II.2. «Расстояния» между контекстами

- Базовая идея: эмпирически измерить степень (семантического) сходства между стимульными контекстами.
- Мера различия контекстов  $i$  и  $j$  («расстояние») — доля всех случаев, в которых языки использовали **разные** маркеры для кодирования данной пары контекстов.
- В итоге строится матрица расстояний, позволяющая увидеть **попарные** расстояния между 48 контекстами.

## II.2. «Расстояния» между контекстами

- При расчете расстояния между парами контекстов использовано Жаккардово расстояние.

$$\frac{j \setminus i + i \setminus j}{j \cap i + j \setminus i + i \setminus j}$$

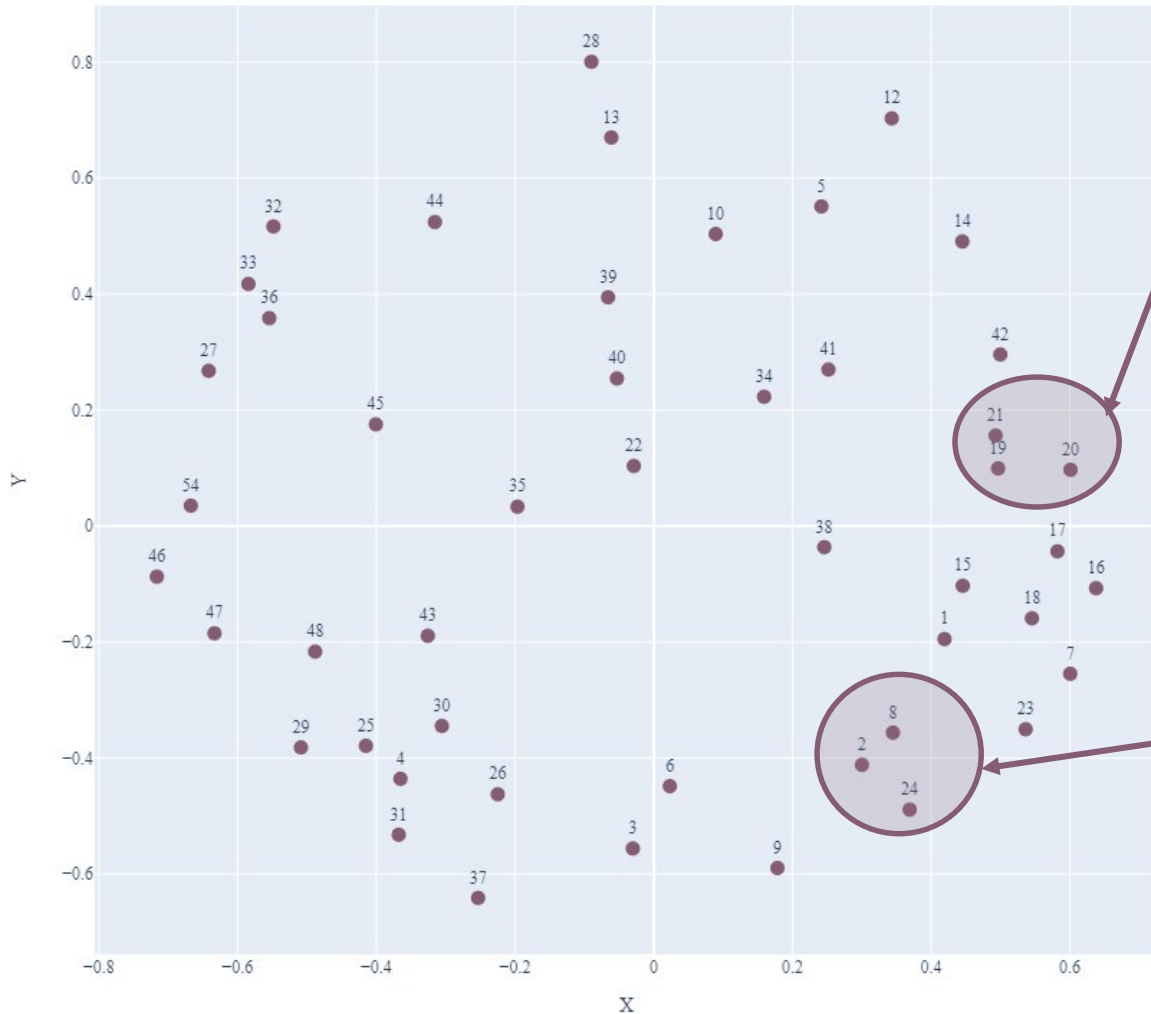
- Получена матрица расстояний между контекстами размером 48x48.

stimulus_no	1	2	3	4	5	...
1	0,00	0,57	0,77	0,91	0,84	
2	0,57	0,00	0,77	0,88	0,84	
3	0,77	0,77	0,00	0,60	0,94	
4	0,91	0,88	0,60	0,00	1,00	
5	0,84	0,84	0,94	1,00	0,00	
...						

## II.2. «Расстояния» между контекстами

- Проблема: такую матрицу невозможно визуализировать графически без искажения расстояний.
- Стандартный алгоритм: многомерное шкалирование (Multi-dimensional scaling, MDS) – сокращение размерности (до двух) с минимальным искажением расстояний

## II.2. «Расстояния» между контекстами



19) Ребенок заплакал от страха.

20) Ребенок рассмеялся от радости.

21) Ребенок от страха спрятался в сарае.

2) Ветка сломалась от тяжести.

8) Женщина проснулась от шума.

24) Женщина устала от работы.

MDS-визуализация расстояний между 48 контекстами



## II.2. «Расстояния» между контекстами

### Отступление

- Альтернатива многомерному шкалированию – кластеризация.
- Но наши контексты на доступных данных плохо поддаются кластеризации.
- Стандартно оптимальное число кластеров определяется как пик на графике силуэта.
- В нашем случае – это всего 2 кластера, заведомо очень грубое деление всех типов.
- **Вывод: отказ от кластеризации.**

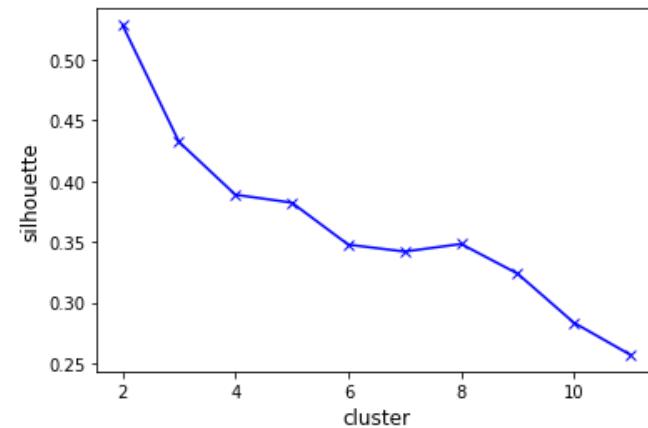
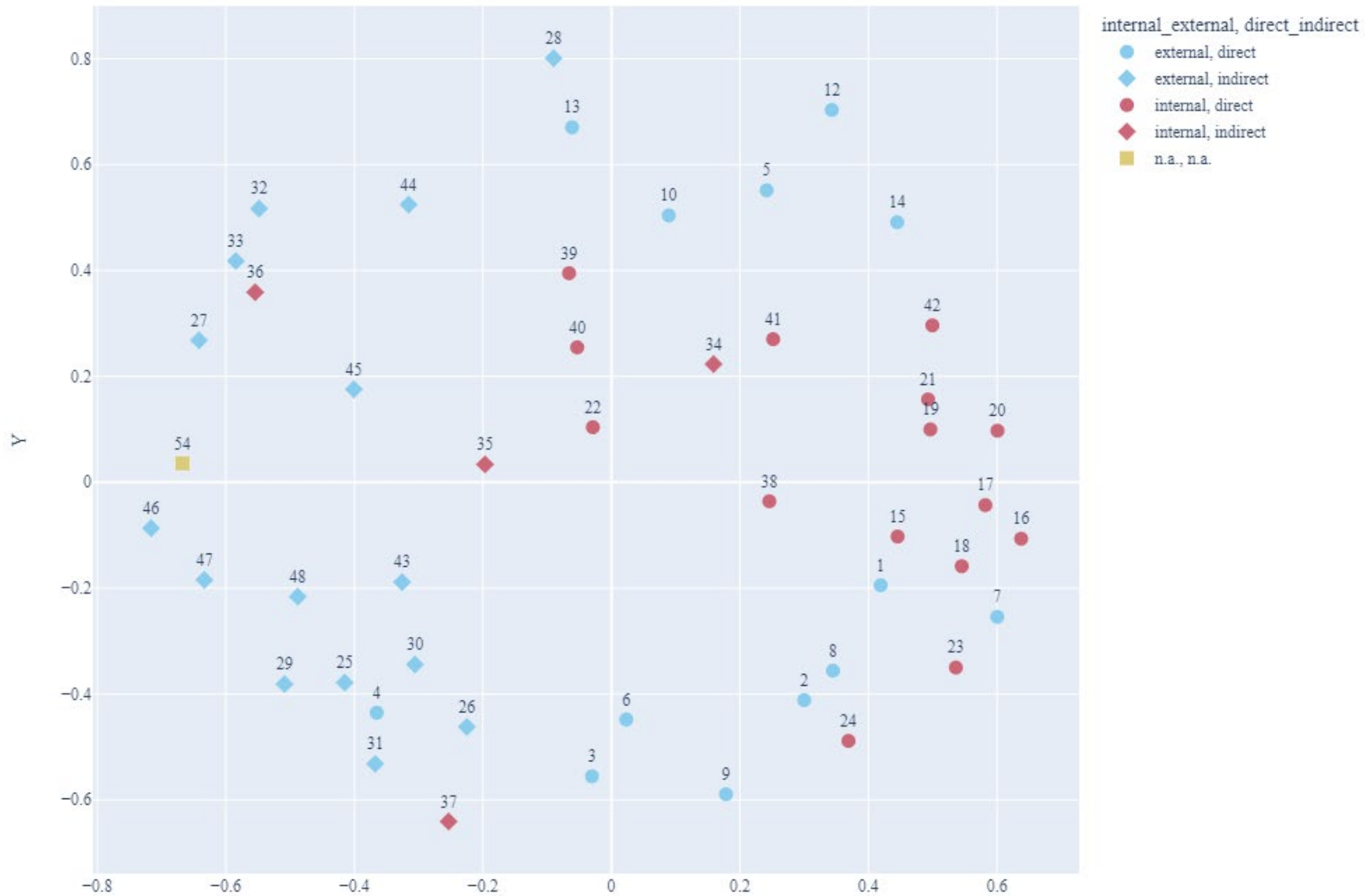
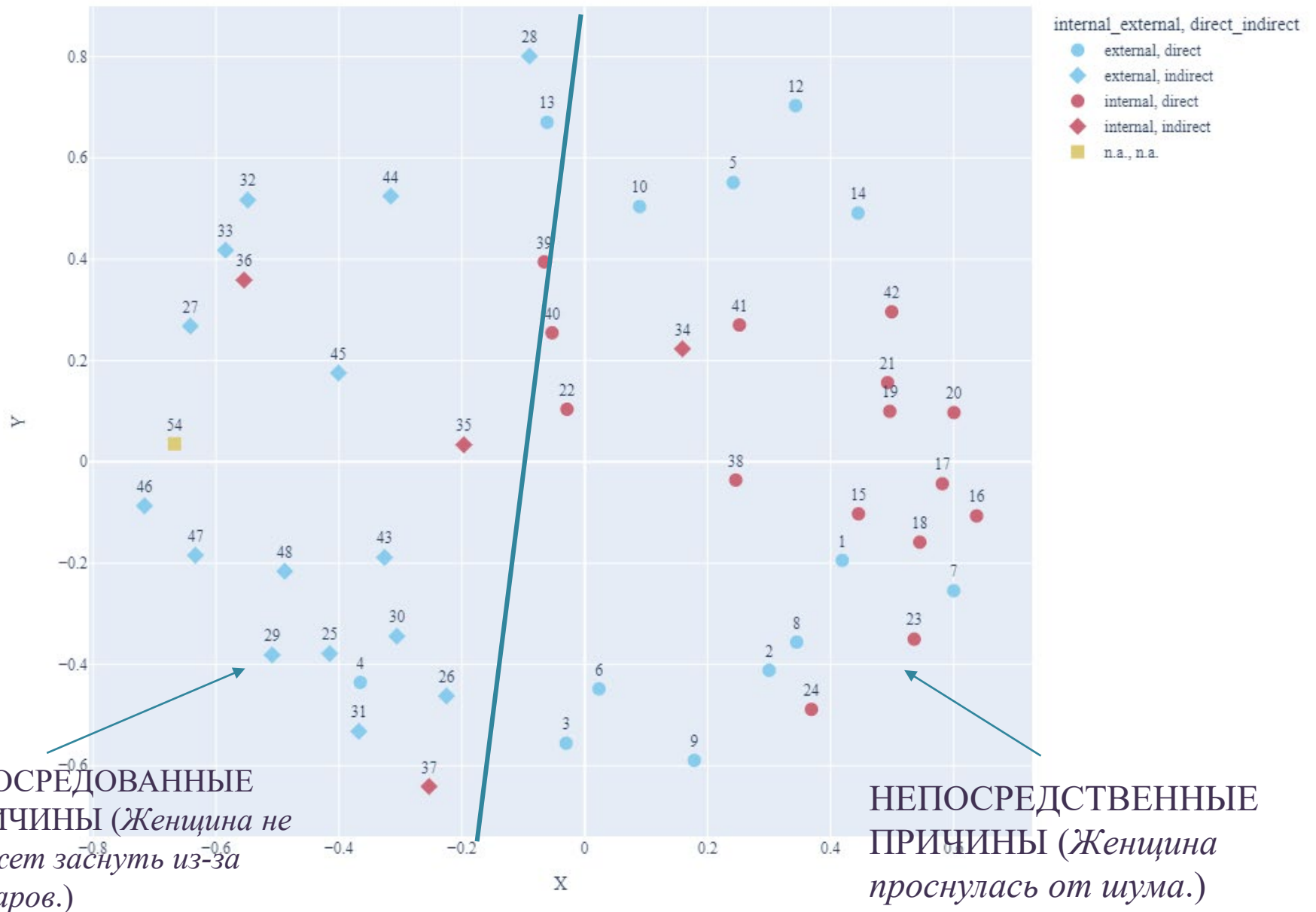


График силуэта для потенциальной кластеризации контекстов.

MDS with internal\_external and direct\_indirect mapping



MDS с наложением деления по параметрам «внутренняя / внешняя» и «непосредственная / опосредованная причина»



ОПОСРЕДОВАННЫЕ ПРИЧИНЫ (Женщина не может заснуть из-за комаров.)

НЕПОСРЕДСТВЕННЫЕ ПРИЧИНЫ (Женщина проснулась от шума.)

MDS с наложением деления по параметрам «внутренняя / внешняя» и «прямая / опосредованная причина»

## II.3. Априорные семантические оппозиции

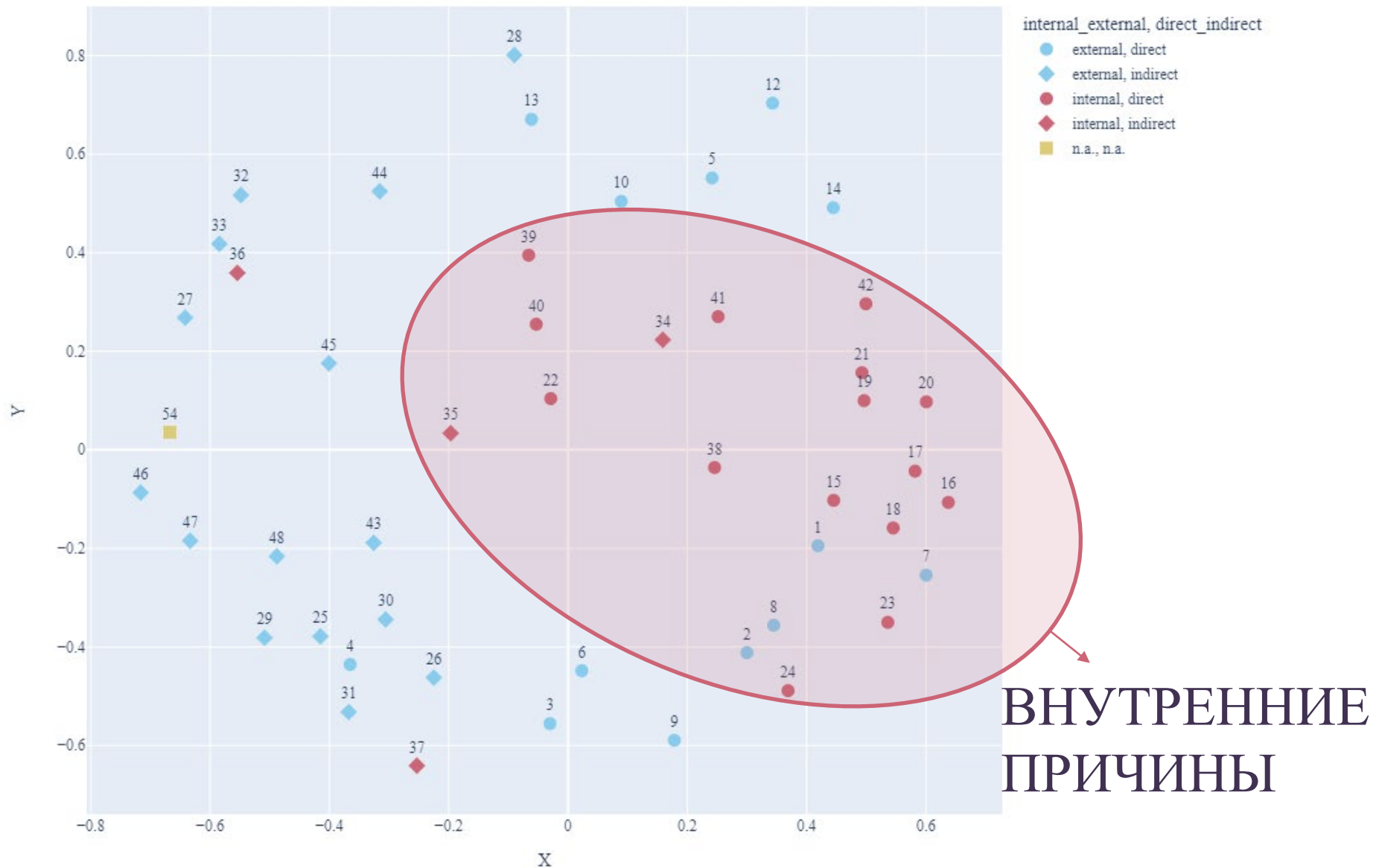
- Известно, что противопоставление непосредственных и опосредованных причин релевантно для славянских языков [Say 2019], в частности для русского [Мельчук, Иорданская 1996; Левонтина 2003].

## II.3. Априорные семантические оппозиции

- Литовский: *iš* и *nuo* последовательно ассоциированы с непосредственными причинами, *dėl* – с опосредованными.
- Маркеры в телугу, имеющие >5 вхождений:

	direct	indirect
DAT	10	0
<i>tō</i>	9	1
<i>valla</i>	5	13

# MDS with internal\_external and direct\_indirect mapping



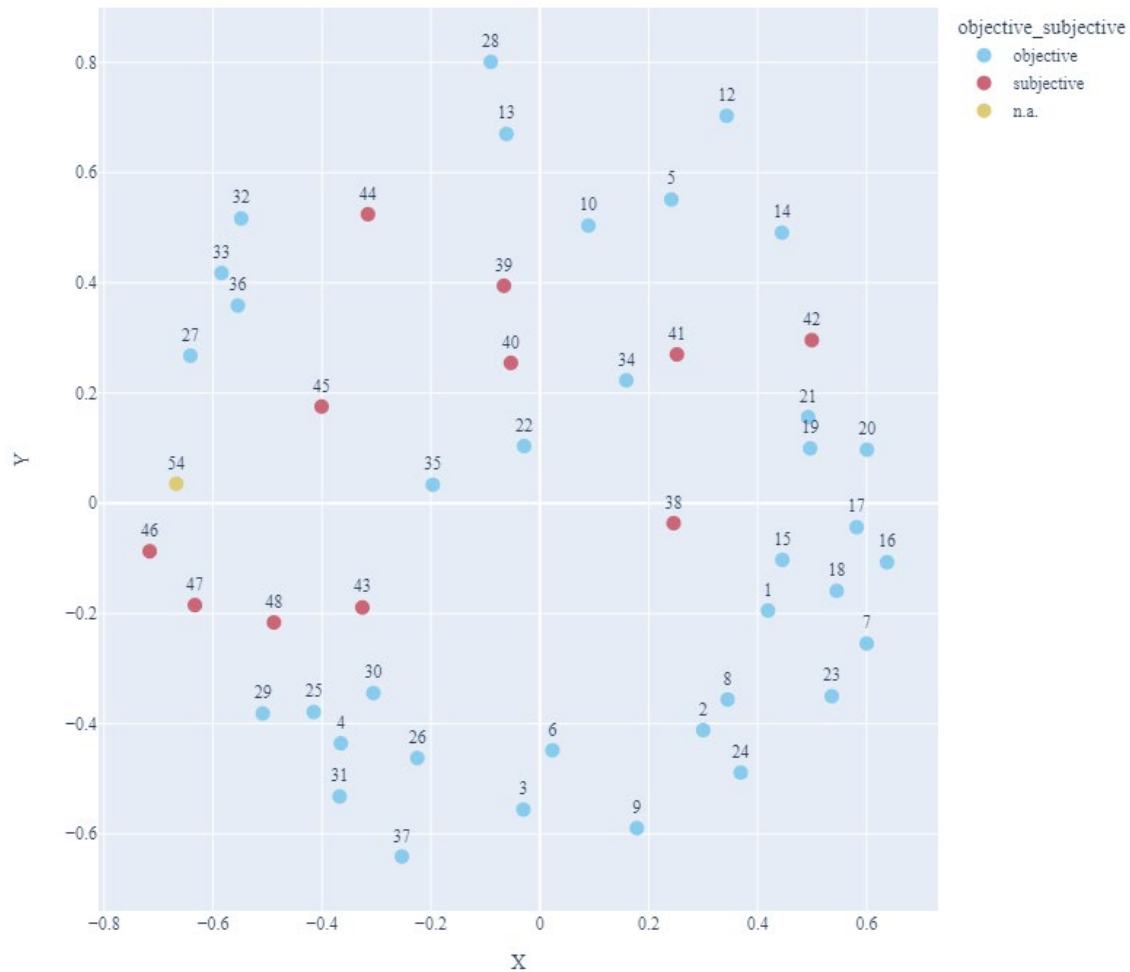
MDS с наложением деления по параметрам «внутренняя / внешняя» и «прямая / опосредованная причина»

## II.3. Априорные семантические оппозиции

- В ряде европейских языков есть показатели, специализирующиеся на внутренних причинах, ср.:
  - немецкое *aus*  
*Das Kind fing aus Angst an zu weinen*  
‘Ребенок заплакал от страха’.
  - фр. *par*  
*La femme par peur a pris son frère par le bras*  
‘В страхе женщина схватила брата за руку’.

## II.3. Априорные семантические оппозиции

MDS with OBJECTIVE/SUBJECTIVE mapping

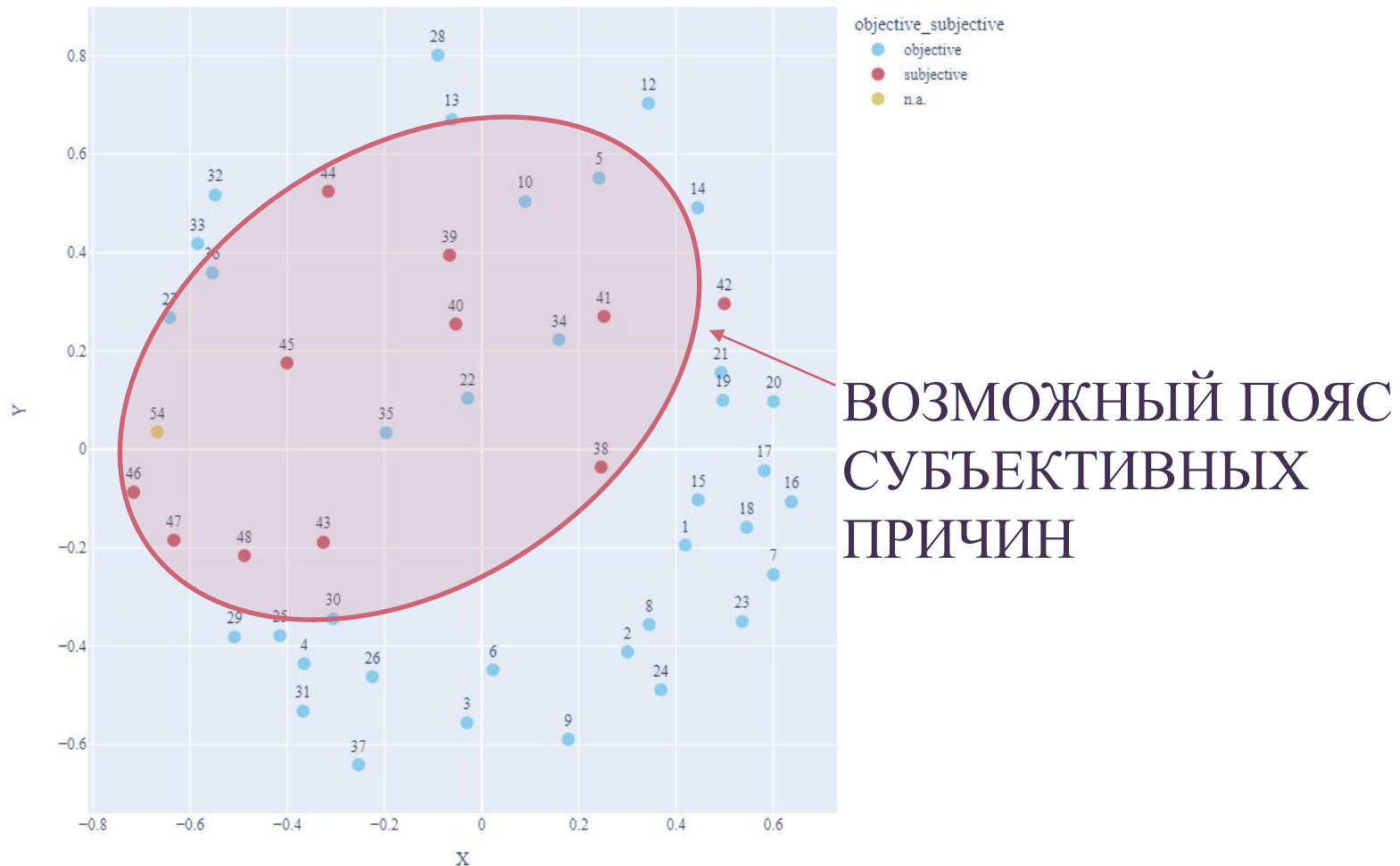


MDS с наложением деления по параметру объективности



## II.3. Априорные семантические оппозиции

MDS with OBJECTIVE/SUBJECTIVE mapping



MDS с наложением деления по параметру объективности

## II.3. Априорные семантические оппозиции

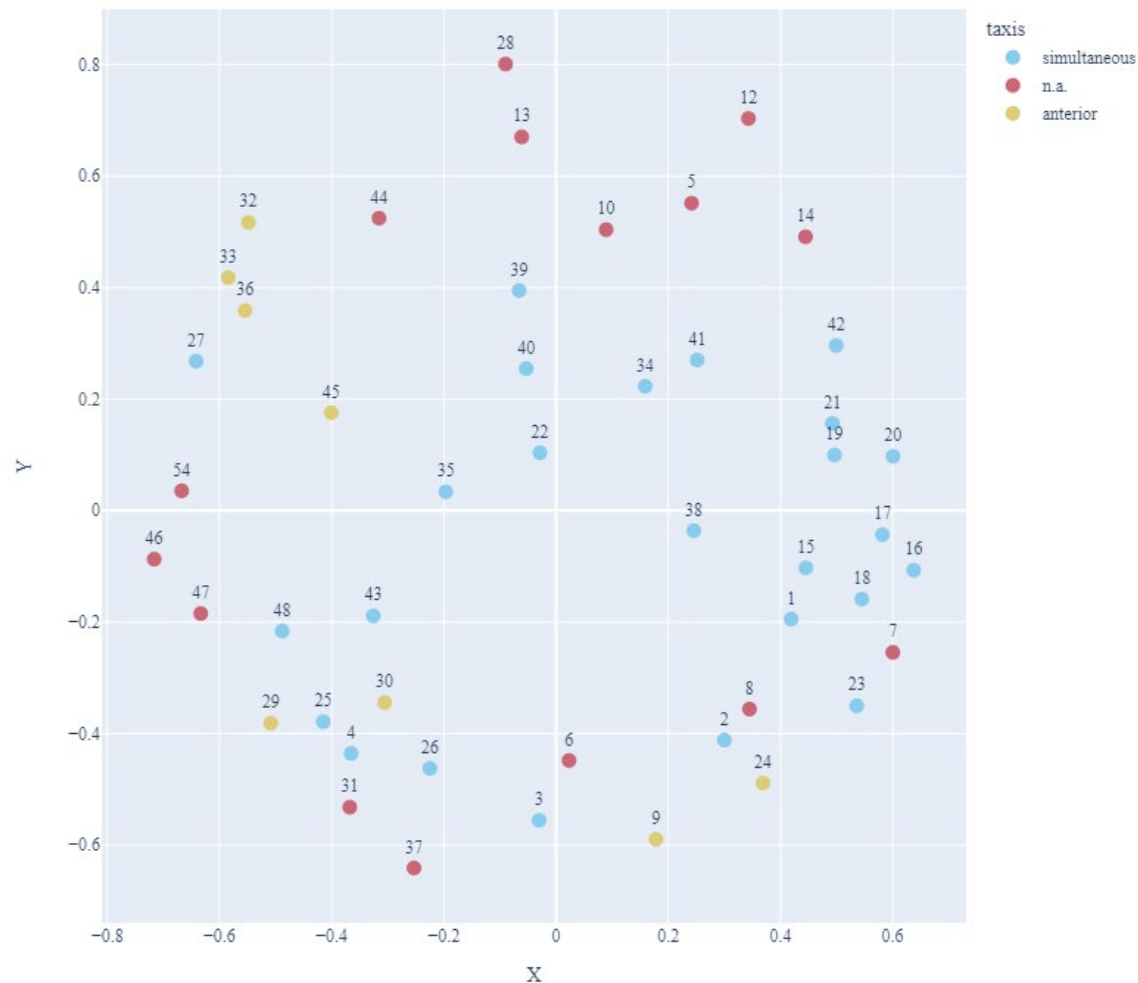
Различие по объективности / субъективности:

- Субъективные причины противопоставлены “нейтральным” объективным причинам, которые представлены по всему полю MDS-визуализации.
- Найти языки, где есть показатели, которые были бы очевидно заточены именно на субъективность, скорее не получается.

## II.3. Априорные семантические оппозиции

- Различие по таксису (одновременность / предшествование).

MDS with TAXIS mapping



MDS с наложением деления по параметру таксиса

## II.3. Априорные семантические оппозиции

- демонстрирует лишь некоторые локальные схождения на визуализации. Вряд ли это типологически очень важный фактор.

## II.3. Априорные семантические оппозиции

### ● Выводы

- Похожесть/непохожесть контекстов измерялась как совпадение/несовпадение средств кодирования в каждой паре контекстов кросс-лингвистически.
- Получившееся пространство расстояний организовано явно не случайно, но кластеризовать его математически пока не получается.

## II.3. Априорные семантические оппозиции

- **Выводы**

- При наложении на получившееся пространство разметки по семантическим типам ярче всего проявляются противопоставления по **непосредственности/опосредованности и внутренней/внешней причине.**
- **NB!** Именно эти параметры также сильно влияют на распределение маркеров по когнитивным схемам.

## II.3. Априорные семантические оппозиции

- **Выводы**

- Кластеризация объективных/субъективных причин работает несколько хуже, хотя область скопления субъективных причин неплохо выделяется.
- Кластеризация таксисных противопоставлений очень фрагментарна.

# Часть III

Результаты про языки,  
или что мы узнали сходствах и различиях  
систем выражения именной причины



# Результаты про языки

III.1 Когнитивные схемы

III.2 «Расстояния» между языками

# III.1. КОГНИТИВНЫЕ СХЕМЫ

- Насколько языки похожи / различаются по использованию когнитивных схем при выражении именной причины?

# III.1. КОГНИТИВНЫЕ СХЕМЫ

- Подсчитывались суммы для отдельных языков по всей анкете
  - альтернативные переводы не учитывались
  - показатели, попадавшие в несколько типов, получали «дробные» баллы

No	Sentence	Goal	Instrument	Location	Object	Source	Other	Dedicated
4	We can't see the shore <b>because of</b> the darkness.	0	0	0	0	0	0	1
5	The man was arrested <b>by</b> mistake.	0	0,5	0,5	0	0	0	0
6	The tree fell <b>because</b> of the wind.	0	0	0	0	0	0	1
7	The man died <b>from</b> electric shock.	0	0	0	0	1	0	0
...	...							

# III.1. КОГНИТИВНЫЕ СХЕМЫ

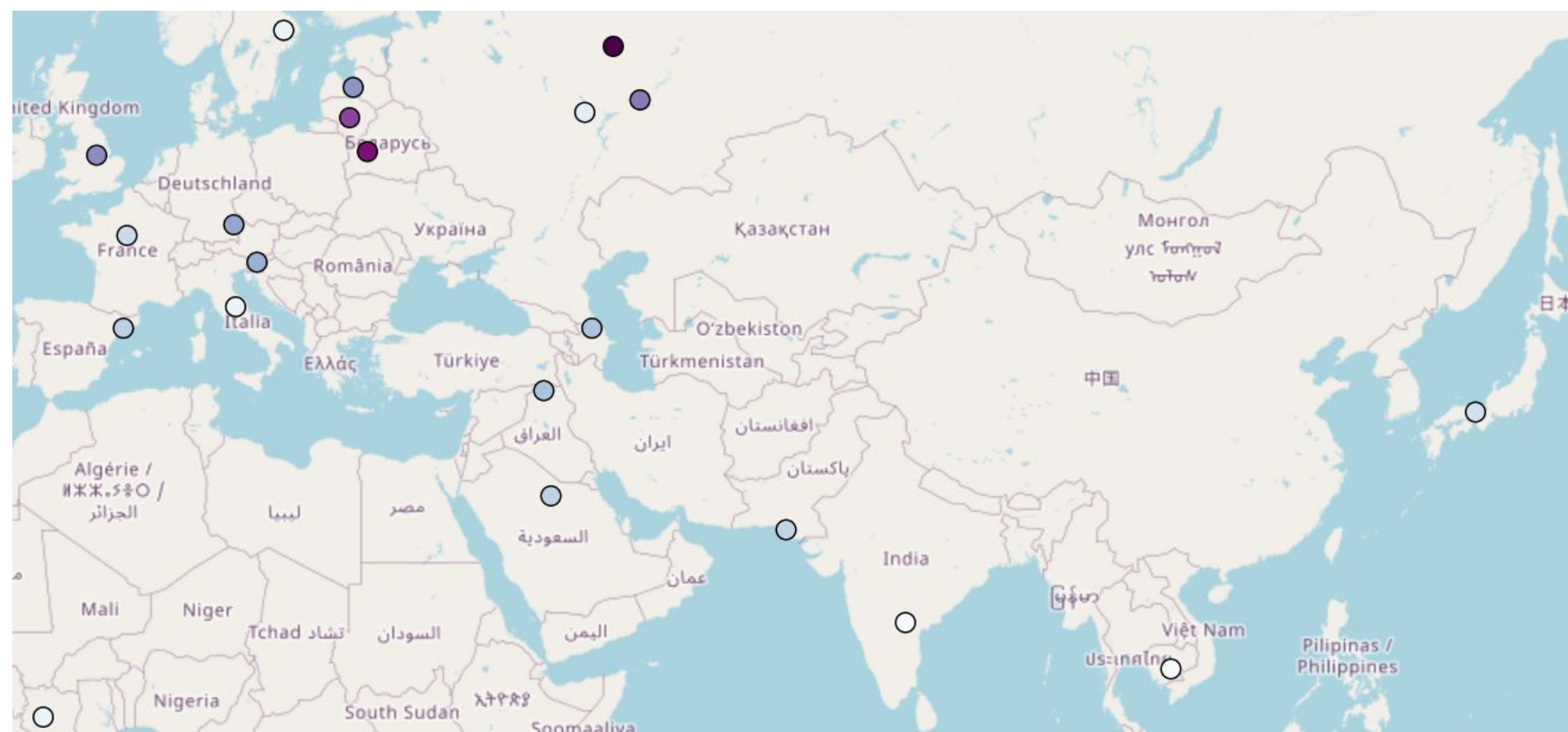
- Многие языки значительно различаются по «долям» разных когнитивных схем, в частности, не демонстрируют «иерархию»:

Dedicated > Source, Instrument > Goal > Location > Other

	Goal	Instrument	Location	Object	Source	Other	Dedicated
Русский	0,06	0,12	0,04	0,00	<b>0,67</b>	0,00	0,12
Итальянский	<b>0,35</b>	<b>0,49</b>	0,00	0,03	0,05	0,00	0,07
Японский	0,07	<b>0,26</b>	<b>0,26</b>	0,05	0,14	0,00	0,21
Кхмерский	0,00	0,00	0,00	0,00	0,00	0,00	<b>1,00</b>

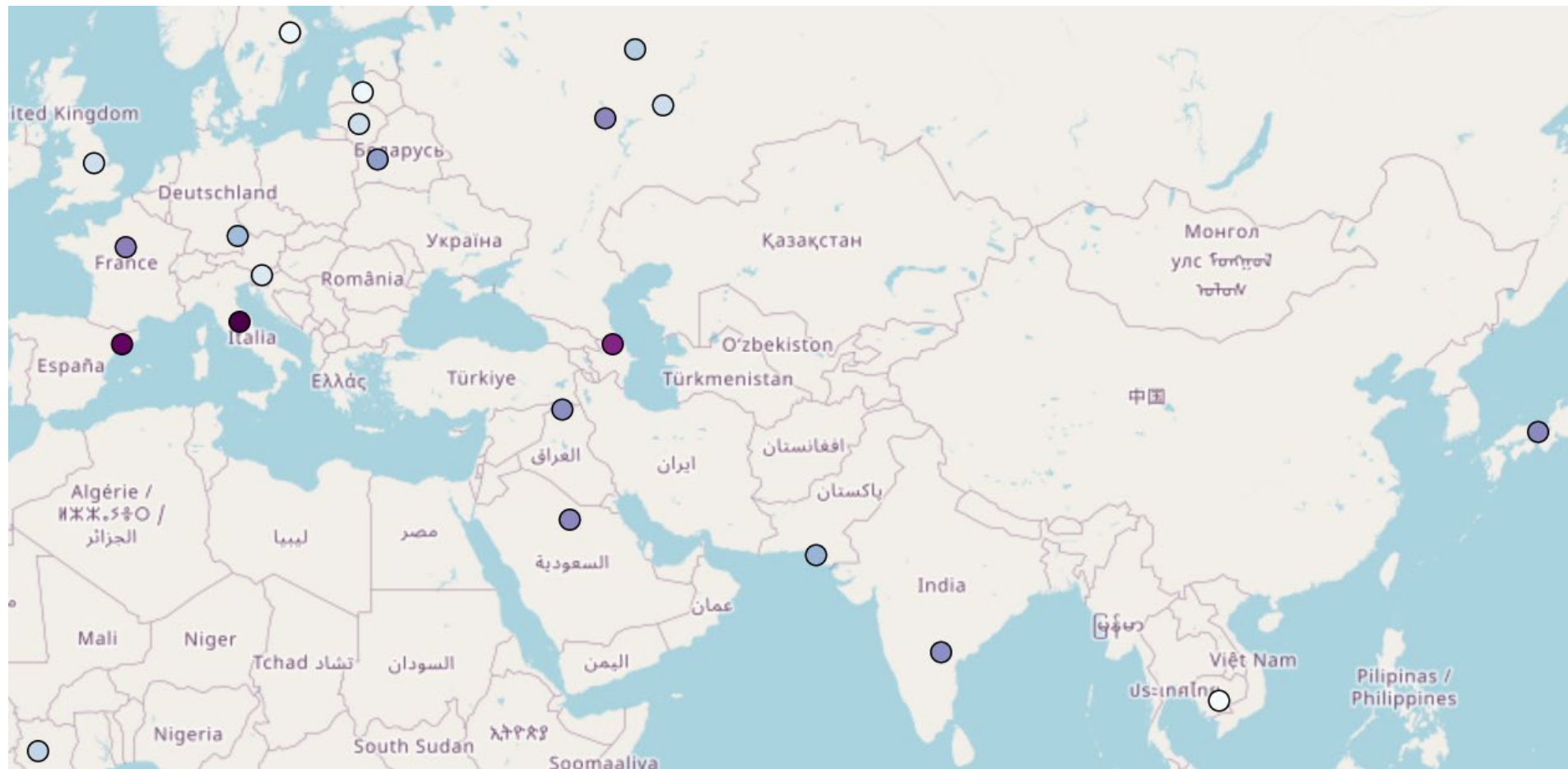
=> Учитывая несбалансированность выборки, наши «усредненные» показатели (см. раздел II.1) могут быть очень искажены

# Карта 1. Представленность типа «Source» по языкам



⇒ Вероятно, ареальная черта с центром в (восточной) Европе – балто-славянские языки?

## Карта 2. Представленность типа «Goal» по языкам



⇒ Широко представлено в романских языках, а также в Азии, нередок синкретизм причины и цели

⇒ Примерно там же сравнительно частотна схема «Instrument», иногда они совмещены (итал. *per*)

# III.1. КОГНИТИВНЫЕ СХЕМЫ

- В целом языков пока мало, чтобы оценить генеалогические и ареальные распределения
- Но уже ясно, что не все языки мира устроены так, как привычные языки Европы
  - Самый яркий пример в выборке – кхмерский, где используются только специализированные показатели, а осмысление причины в терминах пространственных отношений не фиксируется

# III.1. КОГНИТИВНЫЕ СХЕМЫ

- Насколько в разных языках похожи те семантические зоны, для которых типично использование определенных когнитивных схем при осмыслении причины?
- На следующем слайде:
  - контексты идут по строкам и упорядочены по убыванию представленности **специализированных** показателей
  - языки идут по столбцам и упорядочены так же (слева направо)
  - закрашенные ячейки: для искомой комбинации **специализированный** показатель – основной способ











## III.2. «Расстояния» между языками

- Как понять, насколько похожи / различны неродственные языки по принципам организации семантического пространства именной причины?
- Методологическая проблема: невозможна прямая межъязыковая идентификация показателей
- Решение: анализ **распределений** разных контекстов по показателям
- Используется методология, разработанная в рамках проекта VivalTur [Say 2014: 152-156]

## III.2. «Расстояния» между языками

- Отправная точка – энтропия Шеннона, мера «дробности» деления

$$H(x) = - \sum_{i=1}^k p(x_i) \cdot \log(p(x_i))$$

- В кхмерском  $H = 0$ : в качестве основного (main) используется один и тот же маркер
- В итальянском  $H = 1,22$ : распространен предлог *per*
- В немецком  $H = 2,26$ : много разных маркеров, дробное разбиение

## III.2. «Расстояния» между языками

- Чтобы сравнить два языка, нужно построить совместное распределение (только «основные» маркеры -- «Main»)

No	Рус	Англ	Совместное
4	iz-za_GEN	because_of	iz-za_GEN+because_of
5	po_DAT	by	po_DAT+by
6	iz-za_GEN	because_of	iz-za_GEN+because_of
7	ot_GEN	from	ot_GEN+from
8	ot_GEN	from	ot_GEN+from
9	ot_GEN	from	ot_GEN+from
10		at	
11	ot_GEN	from	ot_GEN+from
12	INS	with	INS+with
13	iz-za_GEN	by	iz-za_GEN+by

## III.1. «Расстояния» между языками

- Неформально: чем меньше энтропия совместного распределения, тем больше похожи системы

$$(1) MI (Mutual Information) = H(X) + H(Y) - H(X, Y)$$

- Для подсчета MI в среде R [R Core Team 2021] использовался пакет infotheo [Meyer 2014]



## III.1. «Расстояния» между языками

$$(1) MI (Mutual Information) = H (X) + H (Y) - H (X, Y)$$

- Преобразование MI в меру расстояния

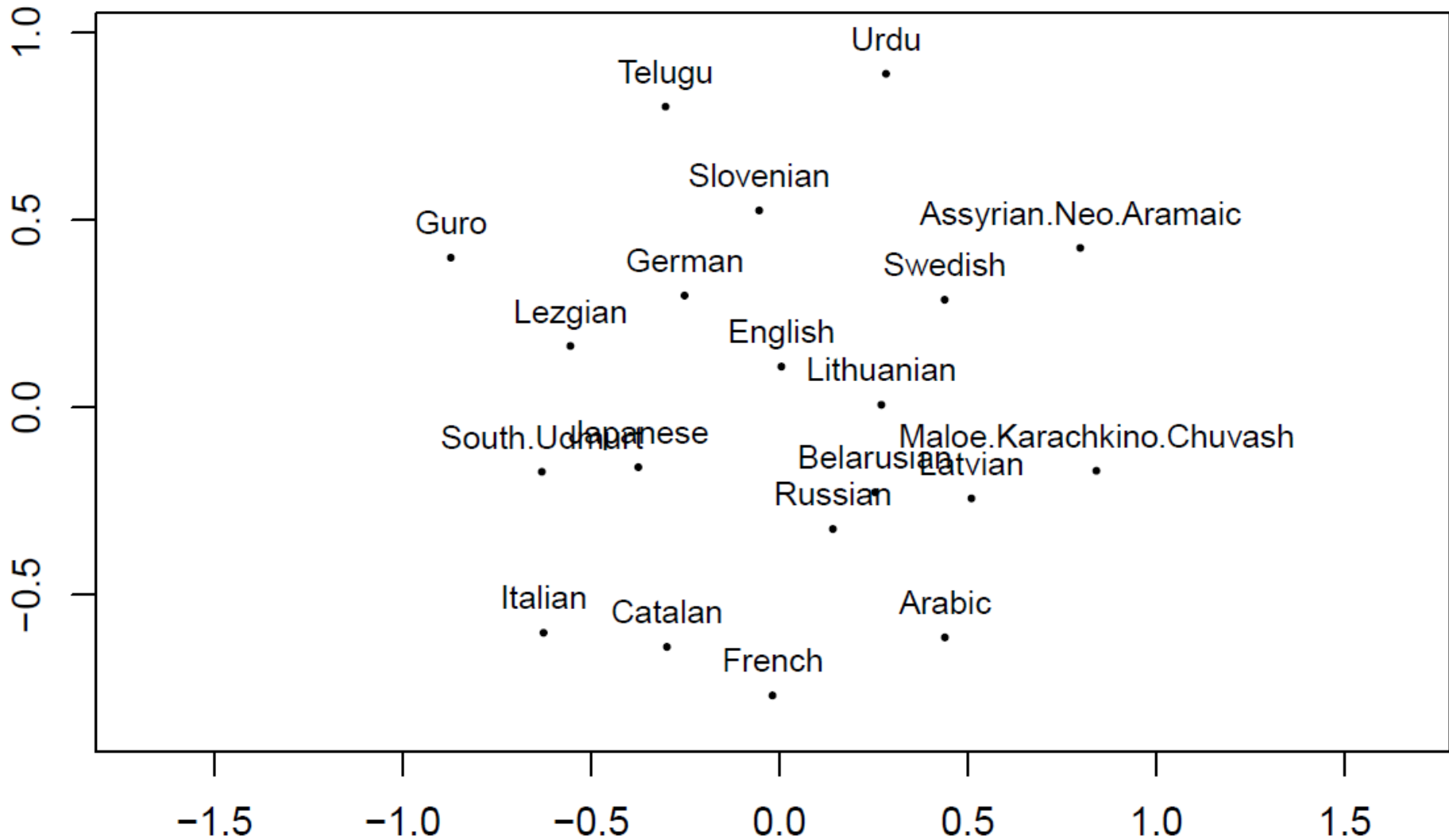
$$(2) D (L1, L2) = 1 - \frac{\frac{MI (L1,L2)}{H (L1)} + \frac{MI(L1,L2)}{H(L2)}}{2}$$

- Построение матрицы расстояний для всех\* пар языков

\* кхмерский исключен, так как его энтропия = 0

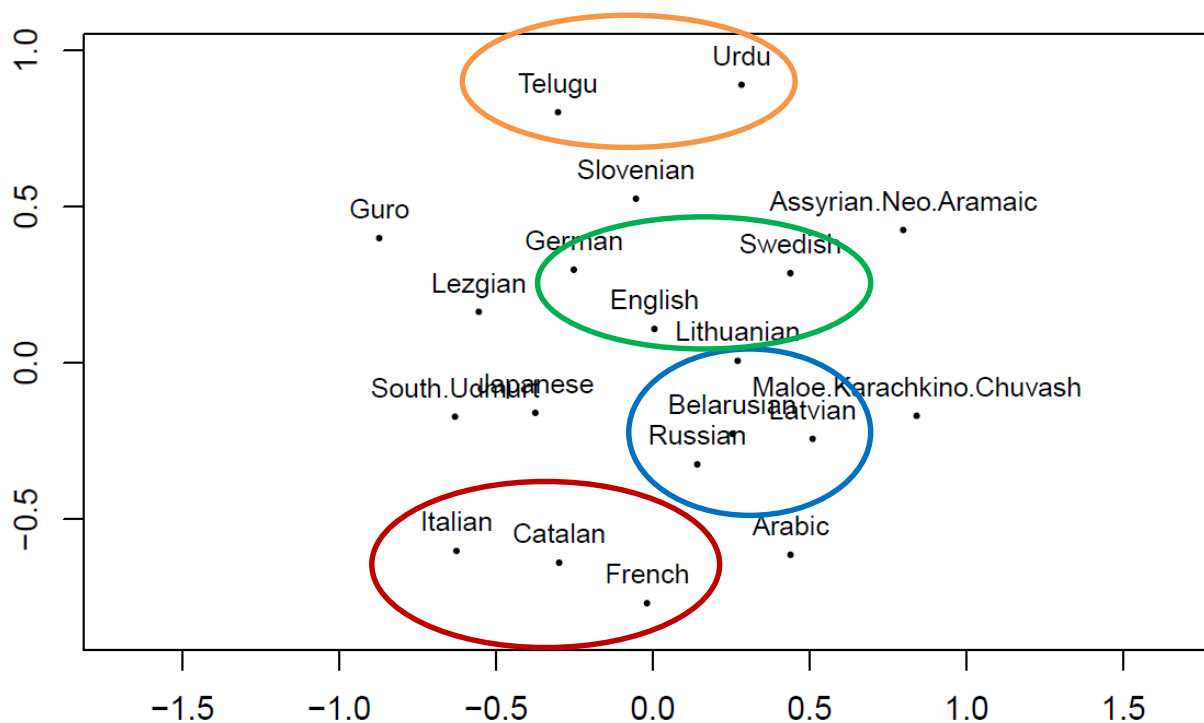
## III.1. «Расстояния» между языками

- Стандартные алгоритмы понижения мерности / визуализации матрицы расстояний:
  - иерархическая кластеризация
  - NeighborNet
  - MDS — многомерное шкалирование (использовался пакет `smacof` в R [de Leeuw, Mair 2009]), см. след. слайд



# III.1. «Расстояния» между языками

NB: Языков очень мало, выводы пока самые предварительные

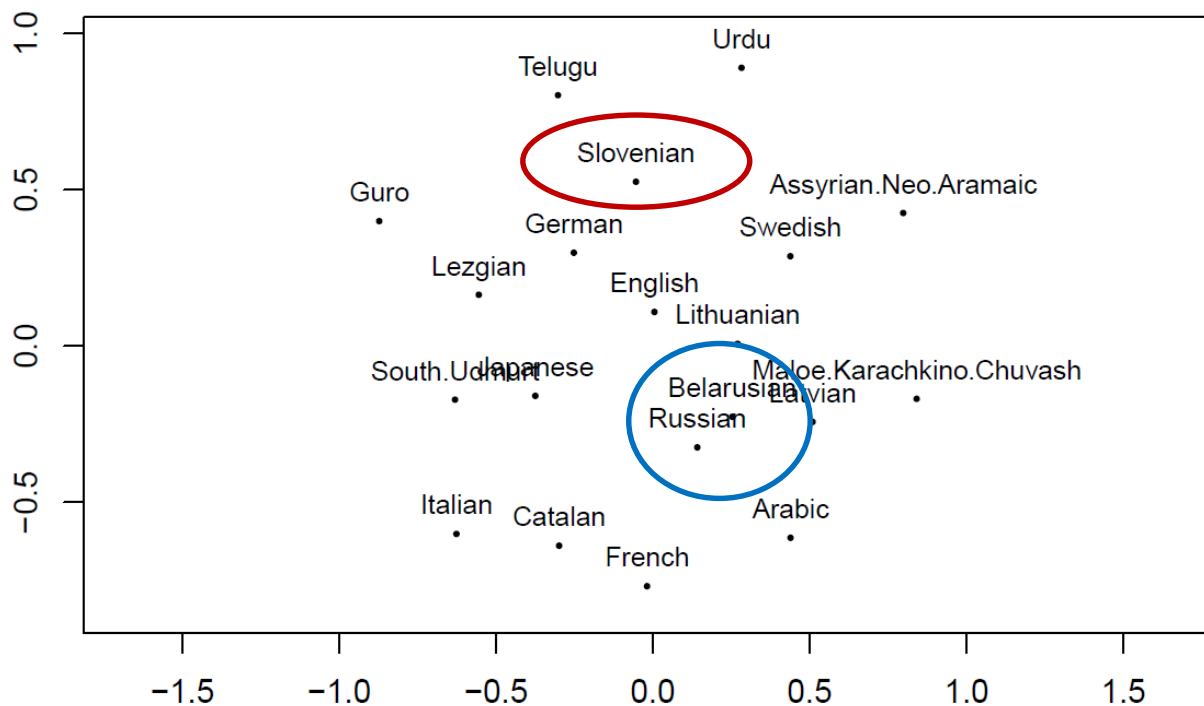


Схождения в группах родственных и/или контактирующих языков:

- романские
- балто-восточно-славянские
- германские
- индийского субконтинента?

# III.1. «Расстояния» между языками

- NB: Ясно, что схождения и расхождения нужно интерпретировать в содержательных терминах
  - Например, словенский выделяется на фоне славянских ослаблением контраста непосредственных и опосредованных причин (*zaradi*)



## III.1. «Расстояния» между языками

- Но системное изучение ареального и генеалогического распределения семантических контрастов в зоне именной причины – дело будущего

Спасибо!

# Литература

## Раздел I

- Croft, W. (1993). Case marking and the semantics of mental verbs. In J. Pustejovsky (Ed.), *Semantics and the lexicon* (pp. 55–72). Kluwer.
- Dirven 1995 — R. Dirven. The construal of cause: The case of cause prepositions // J. R. Taylor, R. MacLaury (eds.). *Language and the Cognitive Construal of the World*. Berlin, New York: Mouton de Gruyter, 1995. P. 95–118.
- Klein, K., Kutscher, S. (2002). Psych-verbs and lexical economy. In *Theorie des Lexikons, 122* (Arbeiten des Sonderforschungsbereichs, 282). 1–41.
- Narrog 2010 — H. Narrog. A diachronic dimension in maps of case functions // *Linguistic Discovery* 8, 1, 2010. P. 233–54.
- Palancar 2001 — E. Palancar. Emergent markers: The relation of instruments, causes, and agents in grammatical constructions // *Sprachtypologie und Universalienforschung*, 54, 4, 2001. P. 365–384.
- Radden 1985 — G. Radden. Spatial metaphors underlying prepositions of causality // W. Paprotté, R. Dirven (eds.). *The Ubiquity of Metaphor*. Amsterdam, Philadelphia: John Benjamins, 1985. P. 177–207.
- Say 2019 – S. Say. 2019. Nominal causal constructions: causal chains and syncretism (a typological study). Talk given at the 15<sup>th</sup> International Cognitive Linguistics Conference. Theme session: “Causation in discourse and cognition: Crosslinguistic perspectives”. August 6-11, Nishinomiya, Japan.
- Say 2021 – S. Say. Nominal causal constructions across Slavic: semantic contrasts in a parallel corpus perspective. *Slavia*, 2021, 90, 2. P. 182–201.
- Verhoeven, E. (2007). *Experiential constructions in Yucatec Maya. A typologically based analysis of a functional domain in a Mayan language*. John Benjamins.
- Yamaguchi 2004 — K. Yamaguchi. A Typological, historical and functional study of adpositions in the languages of the world. PhD dissertation, University of New Mexico, Albuquerque, 2004.



# Литература

## Разделы II-III

Patrick E. Meyer (2014). infotheo: Information-Theoretic Measures. R package version 1.2.0. <https://CRAN.R-project.org/package=infotheo>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Say, Sergey. [Bivalent Verb Classes in the Languages of Europe: A Quantitative Typological Study](#). In: Language dynamics and change, 4 (1), 2014, 116–166. [Pre-final proof of the paper, as of May 26, 2014](#)

Jan de Leeuw, Patrick Mair (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1-30. URL <https://www.jstatsoft.org/v31/i03/>.

И. Б. Левонтина. ИЗ - ЗА 4, ИЗ 8.1, ОТ 6, разг. ПО 1.7, разг. С 1.10, устар. или канц. ЗА II .8, БЛАГОДАРЯ, книжн. ПО ПРИЧИНЕ, книжн. ВСЛЕДСТВИЕ, необходн. В РЕЗУЛЬТАТЕ, офиц. ВВИДУ, книжн. В СИЛУ ‘по той причине, что’ // Ю. Д. Апресян и др. (ред.). Новый объяснительный словарь русского языка. Москва: Школа «Языки славянской культуры», 2003. С. 430–437.

Л. Н. Иорданская, И. А. Мельчук. К семантике русских причинных предлогов // Московский лингвистический журнал 2, 1996 . С. 162–211.