Параллельные корпуса

Темы курса

- Различные параллельные корпуса в составе НКРЯ (сложности создания)
- Alignment и как его делать
- Разметка
- Методы анализа, исследования на материале корпусов
- Построение датасета на примере пассива
- Извлечение контекстов из корпуса
- Семантические карты
- Графы
- Case study: Проект по церновно-славяским текстам
- Case study: Проект по текстам Нового Завета

Темы курса

- Различные параллельные корпуса в составе НКРЯ (сложности создания)
- Alignment и как его делать
- Разметка
- Методы анализа, исследования на материале корпусов
- Построение датасета на примере пассива
- Извлечение контекстов из корпуса
- Семантические карты
- Графы
- Case study: Проект по церновно-славяским текстам
- Case study: Проект по текстам Нового Завета
- Ваши исследования!

Базовые понятия

Устройство

Параллельный корпус представляет собой коллекцию оригинальных текстов на языке L1 с их переводами на один или более языков L2...Ln, важным атрибутом такого корпуса является наличие в нем **выравнивания**, то есть наличие установленных соответствий между текстовыми единицами.

Устройство

Параллельный корпус представляет собой коллекцию оригинальных текстов на языке L1 с их переводами на один или более языков L2...Ln, важным атрибутом такого корпуса является наличие в нем **выравнивания**, то есть наличие установленных соответствий между текстовыми единицами.

Как правило тексты в параллельном корпусе выровнены по предложениям (реже по фрагментам текста, абзацам), иногда в корпусе может быть также и пословное выравнивание, когда соответствия устанавливаются между словами.

Переводные единицы

Предложение (или несколько предложений) на языке оригинала с соответствующими ему переводами на представленные в корпусе языки называется **переводной единицей**.

EN: **He was interrupted by a knock on the door**.
RU: В дверь постучали.
DE: **Ein Klopfen an der Tür unterbrach** ihn.
ITA: **Fu interrotto da qualcuno** *che bussava alla porta*.
ES: **Le interrumpieron unos golpes en la puerta**.
CZ: **Přerušilo ho zaklepání na dveře**.
BG: **Прекъсна го почукване на вратата**.
SE: **Han avbröts av en knackning på dörren**.
FR: **Il fut interrompu par des coups frappés à la porte**.

Переводные единицы

Предложение (или несколько предложений) на языке оригинала с соответствующими ему переводами на представленные в корпусе языки называется **переводной единицей**.

EN: **He was interrupted by a knock on the door**. ←*Оригинальное предложение (source)*
RU: В дверь постучали. ←*Перевод (target)*
DE: **Ein Klopfen an der Tür unterbrach** ihn.
ITA: **Fu interrotto da qualcuno** *che bussava alla porta*.
ES: **Le interrumpieron unos golpes en la puerta**.
CZ: **Přerušilo ho zaklepání na dveře**.
BG: **Прекъсна го почукване на вратата**.
SE: **Han avbröts av en knackning på dörren**.
FR: **Il fut interrompu par des coups frappés à la porte**.

Переводные единицы

Предложение (или несколько предложений) на языке оригинала с соответствующими ему переводами на представленные в корпусе языки называется **переводной единицей**.

EN: **He was interrupted by a knock on the door**. ←*Оригинальное предложение (source)*

RU: В дверь постучали. ←*Перевод (target)*

DE: **Ein Klopfen an der Tür unterbrach** ihn.

ITA: **Fu interrotto da qualcuno** *che bussava alla porta*.

ES: **Le interrumpieron unos golpes en la puerta**.

CZ: **Přerušilo ho zaklepání na dveře**.

BG: **Прекъсна го почукване на вратата**.

SE: **Han avbröts av en knackning på dörren**.

FR: **Il fut interrompu par des coups frappés à la porte**.

Переводная единица

Переводные эквиваленты

Языковые выражения, используемые при переводе, которые соответствуют словам и конструкциям из оригинального текста, называются **переводными эквивалентами**.

EN: **He was interrupted by a knock on the door**.
RU: В дверь постучали.
DE: **Ein Klopfen an der Tür unterbrach** ihn.
ITA: **Fu interrotto da qualcuno** *che bussava alla porta*.
ES: **Le interrumpieron unos golpes en la puerta**.
CZ: **Přerušilo ho zaklepání na dveře**.
BG: **Прекъсна го почукване на вратата**.
SE: **Han avbröts av en knackning på dörren**.
FR: **Il fut interrompu par des coups frappés à la porte**.

Наполнение корпуса

Важным атрибутом параллельных корпусов является наличие в них соответствий различных частей корпуса друг другу.

Идеальный параллельный корпус выглядит примерно так:

	RU	**EN**	**FR**	**IT**	**DE**
RU	1.5M	1.5M	1.5M	1.5M	1.5M
EN	1.5M	1.5M	1.5M	1.5M	1.5M
FR	1.5M	1.5M	1.5M	1.5M	1.5M
IT	1.5M	1.5M	1.5M	1.5M	1.5M
DE	1.5M	1.5M	1.5M	1.5M	1.5M

Наполнение корпуса

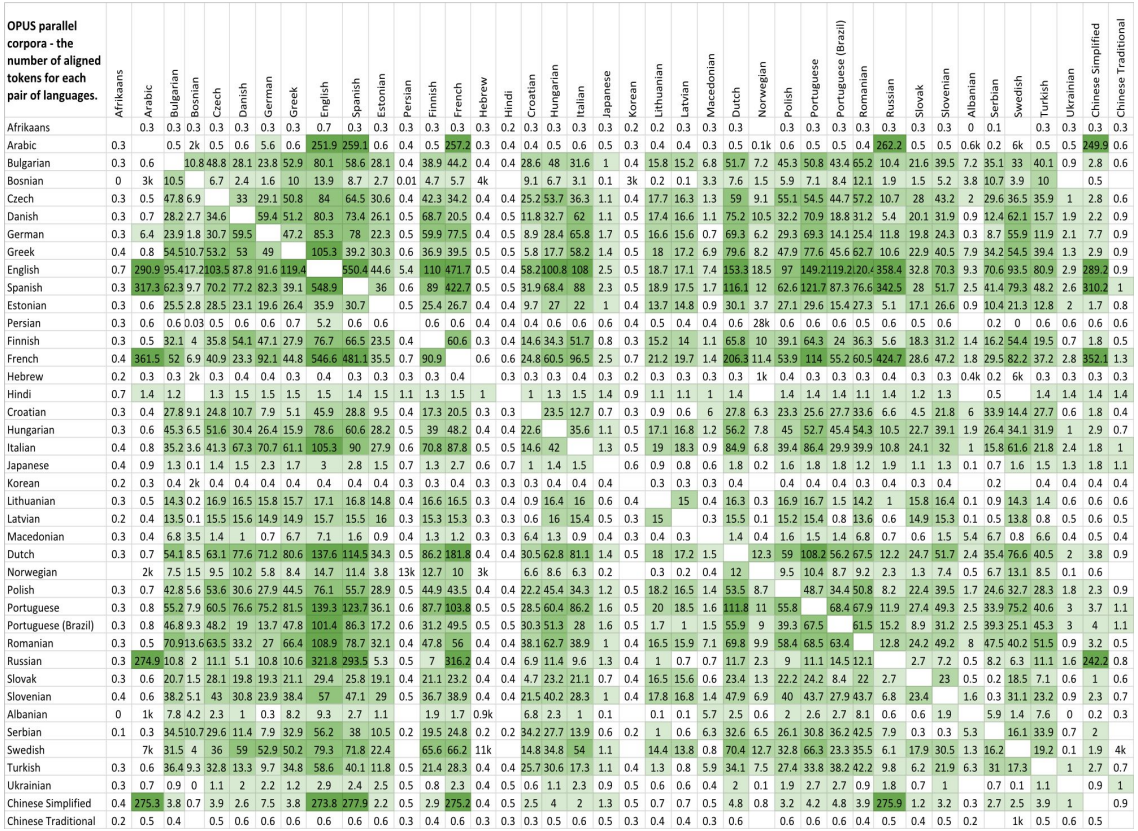
Важным атрибутом параллельных корпусов является наличие в них соответствий различных частей корпуса друг другу.

Идеальный параллельный корпус выглядит примерно так:

	RU	**EN**	**FR**	**IT**	**DE**
RU	1.5M	1.5M	1.5M	1.5M	1.5M
EN	1.5M	1.5M	1.5M	1.5M	1.5M
FR	1.5M	1.5M	1.5M	1.5M	1.5M
IT	1.5M	1.5M	1.5M	1.5M	1.5M
DE	1.5M	1.5M	1.5M	1.5M	1.5M

Например, взяли несколько произведений и для каждого нашли переводы для всех языков, представленных в корпусе

Наполнение: проблемы



OPUS parallel corpora - the number of aligned tokens for each pair of languages.	Afrikaans	Arabic	Bulgarian	Bosnian	Czech	Danish	German	Greek	English	Spanish	Estonian	Persian	Finnish	French	Hebrew	Hindi	Croatian	Hungarian	Italian	Japanese	Korean	Lithuanian	Latvian	Macedonian	Dutch	Norwegian	Polish	Portuguese	Portuguese (Brazil)	Romanian	Russian	Slovak	Slovenian	Albanian	Serbian	Swedish	Turkish	Ukrainian	Chinese Simplified	Chinese Traditional
Afrikaans		0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.7	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.2	0.2	0.3	0.3	0.3	0.3		0.3	0.3	0.3	0.3	0.3	0.3	0.3	0	0.1		0.3	0.3	0.3	0.3
Arabic	0.3		0.5	2k	0.5	0.6	5.6	0.6	251.9	259.1	0.6	0.4	0.5	257.2	0.3	0.4	0.4	0.5	0.6	0.5	0.3	0.4	0.4	0.3	0.5	0.1k	0.6	0.5	0.5	0.4	262.2	0.5	0.5	0.6k	0.2	6k	0.5	0.5	249.9	0.6
Bulgarian	0.3	0.6		10.8	48.8	28.1	23.8	52.9	80.1	58.6	28.1	0.4	38.9	44.2	0.4	0.4	28.6	48	31.6	1	0.4	15.8	15.2	6.8	51.7	7.2	45.3	50.8	43.4	65.2	10.4	21.6	39.5	7.2	35.1	33	40.1	0.9	2.8	0.6
Bosnian	0	3k	10.5		6.7	2.4	1.6	10	13.9	8.7	2.7	0.01	4.7	5.7	4k		9.1	6.7	3.1	0.1	3k	0.2	0.1	3.3	7.6	1.5	5.9	7.1	8.4	12.1	1.9	1.5	5.2	3.8	10.7	3.9	10		0.5		
Czech	0.3	0.5	47.8	6.9		33	29.1	50.8	84	64.5	30.6	0.4	42.3	34.2	0.4	0.4	25.2	53.7	36.3	1.1	0.4	17.7	16.3	1.3	59	9.1	55.1	54.5	44.7	57.2	10.7	28	43.2	2	29.6	36.5	35.9	1	2.8	0.6
Danish	0.3	0.7	28.2	2.7	34.6		59.4	51.2	80.3	73.4	26.1	0.5	68.7	20.5	0.4	0.5	11.8	32.7	62	1.1	0.5	17.4	16.6	1.1	75.2	10.5	32.2	70.9	18.8	31.2	5.4	20.1	31.9	0.9	12.4	62.1	15.7	1.9	2.2	0.9
German	0.3	6.4	23.9	1.8	30.7	59.5		47.2	85.3	78	22.3	0.5	59.9	77.5	0.4	0.5	8.9	28.4	65.8	1.7	0.5	16.6	15.6	0.7	69.3	6.2	29.3	69.3	14.1	25.4	11.8	19.8	24.3	0.3	8.7	55.9	11.9	2.1	7.7	0.9
Greek	0.4	0.8	54.5	10.7	53.2	53	49		105.3	39.2	30.3	0.6	36.9	39.5	0.5	0.5	5.8	17.7	58.2	1.4	0.5	18	17.2	6.9	79.6	8.2	47.9	77.6	45.6	62.7	10.6	22.9	40.5	7.9	34.2	54.5	39.4	1.3	2.9	0.9
English	0.7	290.9	95.4	17.2	103.5	87.8	91.6	119.4		550.4	44.6	5.4	110	471.7	0.5	0.4	58.2	100.8	108	2.5	0.5	18.7	17.1	7.4	153.3	18.5	97	149.2	119.2	120.4	358.4	32.8	70.3	9.3	70.6	93.5	80.9	2.9	289.2	0.9
Spanish	0.3	317.3	62.3	9.7	70.2	77.2	82.3	39.1	548.9		36	0.6	89	422.7	0.5	0.5	31.9	68.4	88	2.3	0.5	18.9	17.5	1.7	116.1	12	62.6	121.7	87.3	76.6	342.5	28	51.7	2.5	41.4	79.3	48.2	2.6	310.2	1
Estonian	0.3	0.6	25.5	2.8	28.5	23.1	19.6	26.4	35.9	30.7		0.5	25.4	26.7	0.4	0.4	9.7	27	22	1	0.4	13.7	14.8	0.9	30.1	3.7	27.1	29.6	15.4	27.3	5.1	17.1	26.6	0.9	10.4	21.3	12.8	2	1.7	0.8
Persian	0.3	0.6	0.6	0.03	0.5	0.6	0.6	0.7	5.2	0.6	0.6		0.6	0.6	0.4	0.4	0.4	0.6	0.6	0.6	0.4	0.5	0.4	0.4	0.6	28k	0.6	0.6	0.6	0.5	0.6	0.5	0.6		0.2	0	0.6	0.6	0.6	0.6
Finnish	0.3	0.5	32.1	4	35.8	54.1	47.1	27.9	76.7	66.5	23.5	0.4		60.6	0.3	0.4	14.6	34.3	51.7	0.8	0.3	15.2	14	1.1	65.8	10	39.1	64.3	24	36.3	5.6	18.3	31.2	1.4	16.2	54.4	19.5	0.7	1.8	0.5
French	0.4	361.5	52	6.9	40.9	23.3	92.1	44.8	546.6	481.1	35.5	0.7	90.9		0.6	0.6	24.8	60.5	96.5	2.5	0.7	21.2	19.7	1.4	206.3	11.4	53.9	114	55.2	60.5	424.7	28.6	47.2	1.8	29.5	82.2	37.2	2.8	352.1	1.3
Hebrew	0.2	0.3	0.3	2k	0.3	0.4	0.4	0.3	0.4	0.3	0.3	0.3	0.3	0.4		0.3	0.3	0.3	0.4	0.3	0.2	0.3	0.3	0.3	0.3	1k	0.4	0.3	0.3	0.3	0.4	0.3	0.3	0.4k	0.2	6k	0.3	0.3	0.3	0.3
Hindi	0.7	1.4	1.2		1.3	1.5	1.5	1.5	1.5	1.4	1.5	1.1	1.3	1.5	1		1	1.3	1.5	1.4	0.9	1.1	1.1	1	1.4		1.4	1.4	1.4	1.1	1.4	1.2	1.3		0.5		1.4	1.4	1.4	1.4
Croatian	0.3	0.4	27.8	9.1	24.8	10.7	7.9	5.1	45.9	28.8	9.5	0.4	17.3	20.5	0.3	0.3		23.5	12.7	0.7	0.3	0.9	0.6	6	27.8	6.3	23.3	25.6	27.7	33.6	6.6	4.5	21.8	6	33.9	14.4	27.7	0.6	1.8	0.4
Hungarian	0.3	0.6	45.3	6.5	51.6	30.4	26.4	15.9	78.6	60.6	28.2	0.5	39	48.2	0.4	0.4	22.6		35.6	1.1	0.5	17.1	16.8	1.2	56.2	7.8	45	52.7	45.4	54.3	10.5	22.7	39.1	1.9	26.4	34.1	31.9	1	2.9	0.7
Italian	0.4	0.8	35.2	3.6	41.3	67.3	70.7	61.1	105.3	90	27.9	0.6	70.8	87.8	0.5	0.5	14.6	42		1.3	0.5	19	18.3	0.9	84.9	6.8	39.4	86.4	29.9	39.9	10.8	24.1	32	1	15.8	61.6	21.8	2.4	1.8	1
Japanese	0.4	0.9	1.3	0.1	1.4	1.5	2.3	1.7	3	2.8	1.5	0.7	1.3	2.7	0.6	0.7	1	1.4	1.5		0.6	0.9	0.8	0.6	1.8	0.2	1.6	1.8	1.8	1.2	1.9	1.1	1.3	0.1	0.7	1.6	1.5	1.3	1.8	1.1
Korean	0.2	0.3	0.4	2k	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.4	0.3	0.3	0.3	0.4	0.4	0.4		0.3	0.3	0.3	0.4		0.4	0.4	0.4	0.3	0.4	0.3	0.4		0.2		0.4	0.4	0.4	0.4
Lithuanian	0.3	0.5	14.3	0.2	16.9	16.5	15.8	15.7	17.1	16.8	14.8	0.4	16.6	16.5	0.3	0.4	0.9	16.4	16	0.6	0.4		15	0.4	16.3	0.3	16.9	16.7	1.5	14.2	1	15.8	16.4	0.1	0.9	14.3	1.4	0.6	0.6	0.6
Latvian	0.2	0.4	13.5	0.1	15.5	15.6	14.9	14.9	15.7	15.5	16	0.3	15.3	15.3	0.3	0.3	0.6	16	15.4	0.5	0.3	15		0.3	15.5	0.1	15.2	15.4	0.8	13.6	0.6	14.9	15.3	0.1	0.5	13.8	0.8	0.5	0.6	0.5
Macedonian	0.3	0.4	6.8	3.5	1.4	1	0.7	6.7	7.1	1.6	0.9	0.4	1.3	1.2	0.3	0.3	6.4	1.3	0.9	0.4	0.3	0.4	0.3		1.4	0.4	1.6	1.5	1.4	6.8	0.7	0.6	1.5	5.4	6.7	0.8	6.6	0.4	0.5	0.4
Dutch	0.3	0.7	54.1	8.5	63.1	77.6	71.2	80.6	137.6	114.5	34.3	0.5	86.2	181.8	0.4	0.4	30.5	62.8	81.1	1.4	0.5	18	17.2	1.5		12.3	59	108.2	56.2	67.5	12.2	24.7	51.7	2.4	35.4	76.6	40.5	2	3.8	0.9
Norwegian		2k	7.5	1.5	9.5	10.2	5.8	8.4	14.7	11.4	3.8	13k	12.7	10	3k		6.6	8.6	6.3	0.2		0.3	0.2	0.4	12		9.5	10.4	8.7	9.2	2.3	1.3	7.4	0.5	6.7	13.1	8.5	0.1	0.6	
Polish	0.3	0.7	42.8	5.6	53.6	30.6	27.9	44.5	76.1	55.7	28.9	0.5	44.9	43.5	0.4	0.4	22.2	45.4	34.3	1.2	0.5	18.2	16.5	1.4	53.5	8.7		48.7	34.4	50.8	8.2	22.4	39.5	1.7	24.6	32.7	28.3	1.8	2.3	0.9
Portuguese	0.3	0.8	55.2	7.9	60.5	76.6	75.2	81.5	139.3	123.7	36.1	0.6	87.7	103.8	0.5	0.5	28.5	60.4	86.2	1.6	0.5	20	18.5	1.6	111.8	11	55.8		68.4	67.9	11.9	27.4	49.3	2.5	33.9	75.2	40.6	3	3.7	1.1
Portuguese (Brazil)	0.3	0.8	46.8	9.3	48.2	19	13.7	47.8	101.4	86.3	17.2	0.6	31.2	49.5	0.5	0.5	30.3	51.3	28	1.6	0.5	1.7	1	1.5	55.9	9	39.3	67.5		61.5	15.2	8.9	31.2	2.5	39.3	25.1	45.3	3	4	1.1
Romanian	0.3	0.5	70.9	13.6	63.5	33.2	27	66.4	108.9	78.7	32.1	0.4	47.8	56	0.4	0.4	38.1	62.7	38.9	1	0.4	16.5	15.9	7.1	69.8	9.9	58.4	68.5	63.4		12.8	24.2	49.2	8	47.5	40.2	51.5	0.9	3.2	0.5
Russian	0.3	274.9	10.8	2	11.1	5.1	10.8	10.6	321.8	293.5	5.3	0.5	7	316.2	0.4	0.4	6.9	11.4	9.6	1.3	0.4	1	0.7	0.7	11.7	2.3	9	11.1	14.5	12.1		2.7	7.2	0.5	8.2	6.3	11.1	1.6	242.2	0.8
Slovak	0.3	0.6	20.7	1.5	28.1	19.8	19.3	21.1	29.4	25.8	19.1	0.4	21.1	23.2	0.4	0.4	4.7	23.2	21.1	0.7	0.4	16.5	15.6	0.6	23.4	1.3	22.2	24.2	8.4	22	2.7		23	0.5	0.2	18.5	7.1	0.6	1	0.6
Slovenian	0.4	0.6	38.2	5.1	43	30.8	23.9	38.4	57	47.1	29	0.5	36.7	38.9	0.4	0.4	21.5	40.2	28.3	1	0.4	17.8	16.8	1.4	47.9	6.9	40	43.7	27.9	43.7	6.8	23.4		1.6	0.3	31.1	23.2	0.9	2.3	0.7
Albanian	0	1k	7.8	4.2	2.3	1	0.3	8.2	9.3	2.7	1.1		1.9	1.7	0.9k		6.8	2.3	1	0.1		0.1	0.1	5.7	2.5	0.6	2	2.6	2.7	8.1	0.6	0.6	1.9		5.9	1.4	7.6	0	0.2	0.3
Serbian	0.1	0.3	34.5	10.7	29.6	11.4	7.9	32.9	56.2	38	10.5	0.2	19.5	24.8	0.2	0.2	34.2	27.7	13.9	0.6	0.2	1	0.6	6.3	32.6	6.5	26.1	30.8	36.2	42.5	7.9	0.3	0.3	5.3		16.1	33.9	0.7	2	
Swedish		7k	31.5	4	36	59	52.9	50.2	79.3	71.8	22.4		65.6	66.2	11k		14.8	34.8	54	1.1		14.4	13.8	0.8	70.4	12.7	32.8	66.3	23.3	35.5	6.1	17.9	30.5	1.3	16.2		19.2	0.1	1.9	4k
Turkish	0.3	0.6	36.4	9.3	32.8	13.3	9.7	34.8	58.6	40.1	11.8	0.5	21.4	28.3	0.4	0.4	25.7	30.6	17.3	1.1	0.4	1.3	0.8	5.9	34.1	7.5	27.4	33.8	38.2	42.2	9.8	6.2	21.9	6.3	31	17.3		1	2.7	0.7
Ukrainian	0.3	0.7	0.9	0	1.1	2	2.2	1.2	2.9	2.4	2.5	0.5	0.8	2.3	0.4	0.5	0.6	1.1	2.3	0.9	0.5	0.6	0.6	0.4	2	0.1	1.9	2.7	2.7	0.9	1.8	0.7	1		0.7	0.1	1.1		0.9	1
Chinese Simplified	0.4	275.3	3.8	0.7	3.9	2.6	7.5	3.8	273.8	277.9	2.2	0.5	2.9	275.2	0.4	0.5	2.5	4	2	1.3	0.5	0.7	0.7	0.5	4.8	0.8	3.2	4.2	4.8	3.9	275.9	1.2	3.2	0.3	2.7	2.5	3.9	1		0.9
Chinese Traditional	0.2	0.5	0.4		0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.3	0.4	0.6	0.3	0.3	0.3	0.5	0.6	0.5	0.3	0.4	0.4	0.3	0.6		0.6	0.6	0.6	0.4	0.5	0.4	0.5	0.2		1k	0.5	0.6	0.5	

- Не существует каких-то переводов, либо они недоступны
- Очень специфические тексты становятся массово параллельными

Как обстоят дела с параллельными корпусами?

- С одной стороны ресурсов, предлагающих исследователям в открытом доступе мультиязычные параллельные тексты имеется немало.
- Однако по факту напрямую использовать в исследованиях можно не все из них, это связано с
  - неоднородностью объема соответствий переводных пар в рамках одного корпуса,
  - отсутствием унифицированной разметки,
  - спецификой самих текстов, входящих в корпус
- Двуязычных параллельных корпусов существует гораздо больше. В принципе, чем меньше в корпусе языков, тем больше шансов найти корпус большого объема.

Исследования на материале параллельных корпусов

Вариативность переводных эквивалентов

- Прелесть работы с переводами состоит в том, что изучая некоторую конструкцию или даже просто употребление слова можно найти много нетривиальных соответствий.
- Соответствия не всегда однородные, не всегда прилагательному соответствует прилагательное, а существительному существительное, пассиву соответствует пассив или актив.

Примеры

Иногда это обусловлено наличием в языках нескольких “официально” зафиксированных вариантов выражения, например, для каузативов:

– Lexical causatives, e. g., *break*$_{TR}$ or *walk*$_{TR}$;

Примеры

Иногда это обусловлено наличием в языках нескольких “официально” зафиксированных вариантов выражения, например, для каузативов:

– Lexical causatives, e. g., *break*$_{TR}$ or *walk*$_{TR}$;
– Morphological causatives, e. g., *tone change, reduplication, or affixation*;

Примеры

Иногда это обусловлено наличием в языках нескольких “официально” зафиксированных вариантов выражения, например, для каузативов:

– Lexical causatives, e. g., *break*$_{TR}$ or *walk*$_{TR}$;
– Morphological causatives, e. g., *tone change, reduplication, or affixation*;
– Complex predicates, e. g., serial verbs, French faire ‘make’ + VINF, or causative particles;

Примеры

Иногда это обусловлено наличием в языках нескольких “официально” зафиксированных вариантов выражения, например, для каузативов:

– Lexical causatives, e. g., *break*$_{TR}$ or *walk*$_{TR}$;
– Morphological causatives, e. g., *tone change, reduplication, or affixation*;
– Complex predicates, e. g., serial verbs, French faire ‘make’ + VINF, 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.

Примеры

Пассивные конструкции

- **Соответствие лексической единице**
  - EN: **Her hands shook** slightly

    ITA: **Le sue mani erano scosse da un lieve tremito**
- **Соответствие конструкции вроде *there is/are***
  - EN: **There was a dull murmur** of assent **throughout the class**.

    ITA: **La classe fu percorsa da un cupo mormorio** di assenso.

  - EN: **There were movements** from the watching crowd **in front of the castle**, <...>

    ITA: **La folla davanti al castello fu percorsa da un fremito** <...>

Персоны

Персоны

Bernhard Wälchli, Michael Cysouw

Wälchli B., Cysouw M. Lexical typology through similarity semantics: Toward a semantic map of motion verbs // Linguistics. 2012. No 3 (50). C. 671–710.

Michael Cysouw

Östen Dahl

Dahl Ö. From questionnaires to parallel corpora in typology // Language Typology and Universals. 2007. No 2 (60). C. 172–181

Östen Dahl

Персоны

Bernhard Wälchli, Michael Cysouw

Wälchli B., Cysouw M. Lexical typology through similarity semantics: Toward a semantic map of motion verbs // Linguistics. 2012. No 3 (50). C. 671–710.

Michael Cysouw

Östen Dahl

Dahl Ö. From questionnaires to parallel corpora in typology // Language Typology and Universals. 2007. No 2 (60). C. 172–181

Östen Dahl

Персоны: Bernhard Wälchli, Michael Cysouw, Östen Dahl

Одними из первых указали на преимущества параллельных корпусов

Персоны: Bernhard Wälchli, Michael Cysouw, Östen Dahl

Одними из первых указали на преимущества параллельных корпусов, а именно:

- Охват сразу с нескольких языков
- Возможность применения различных современных методов анализа.
- Работа с ситуациями в одинаковом контекстном окружении. То есть исследователь располагает примерами с одинаковой семантической и прагматической составляющей. Это не разрозненные примеры из разных источников, это сопоставимый материал.

Персоны: Bernhard Wälchli, Mi

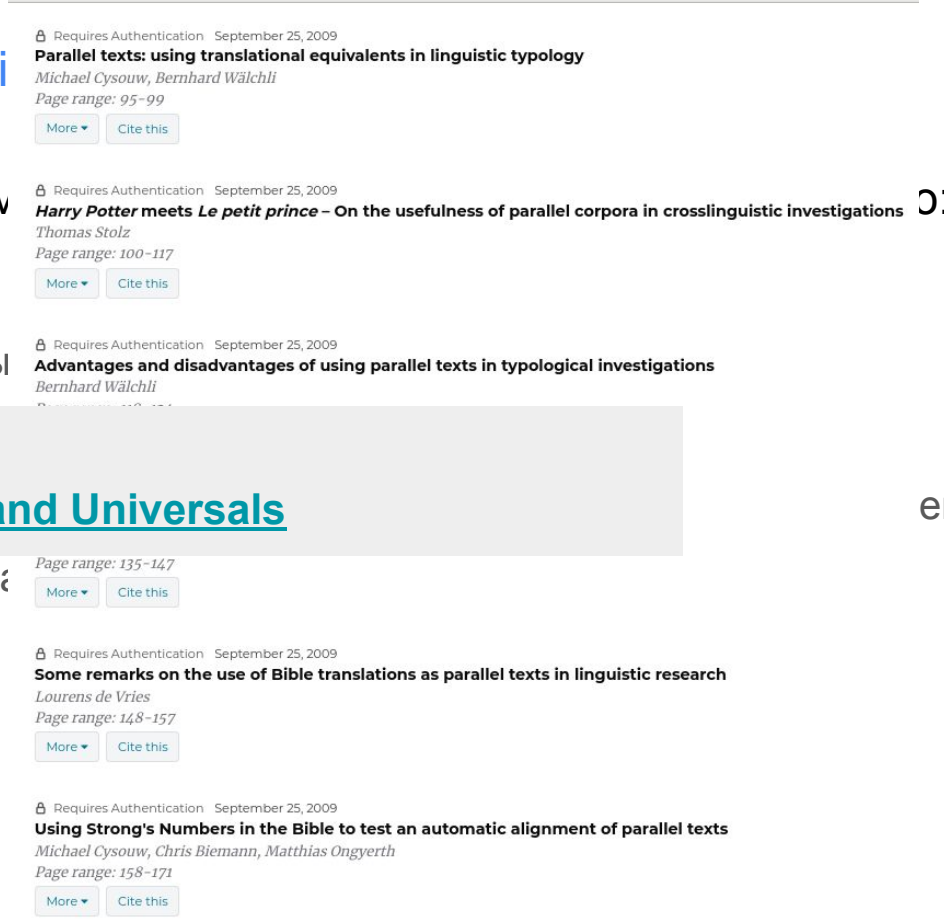
Одними из первых указали на преим о:

- Охват сразу с нескольких языков
- Возможность применения различны

Volume 60 Issue 2,
Issue of STUF - Language Typology and Universals

ей.

Это не разрозненные примеры из ра



Персоны

- **Robert Östling, University of Stockholm**
  - https://www.su.se/english/profiles/robe-1.187515
  - Важно не перепутать с экономистом-тёзкой!
  - Östling R. 6. Studying colexification through massively parallel corpora Berlin, Boston: De Gruyter, 2016
  - Östling R. Word order typology through multilingual word alignment, In: The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Proceedings of the Conference, Volume 2: Short Papers, 2015, p. 205-211
- **Jörg Tiedemann, University of Helsinki**
  - https://researchportal.helsinki.fi/en/persons/j%C3%B6rg-tiedemann
  - Отец OPUS'a
  - Östling R., Tiedemann J. Efficient Word Alignment with Markov Chain Monte Carlo // The Prague Bulletin of Mathematical Linguistics. 2016. No (106). C. 125–146

Персоны

- **Ruprecht von Waldenfels, University of Jena**
  - https://www.gw.uni-jena.de/fakultaet/institut-fuer-slawistik-und-kaukasusstudien/mitarbeiterinnen/von-waldenfels-ruprecht
  - Славист, создатель корпуса ParaSOL
  - Waldenfels R. von Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment // Beiträge der europäischen slavistischen Linguistik (POLYSLAV). 2006. (9). С. 123–138.
- **Дмитрий В. Сичинава, ИРЯ РАН**
  - https://www.ruslang.ru/publica/sichinava
  - Типолог, специалист по славянским языкам, один из создателей параллельных корпусов в составе НКРЯ
  - Sitchinava D. V, others Parallel corpora within the Russian National Corpus // Prace Filologiczne. 2012. No 63. С. 271–278

Персоны

- **Наталия Лёвшина, MPI for Psycholinguistics Nijmegen**
  - Создательница корпуса ParTY, развивает применение квантитативных методов и параллельных корпусов в типологии
  - Levshina N. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages // Folia Linguistica. 2016a. No 2 (50).

- **Сергей С. Сай, ИЛИ РАН**
  - Типолог, специалист по аргументной структуре
  - Say, Sergey. Nominal causal constructions across Slavic: semantic contrasts in a parallel corpus perspective. Slavia, 2021, 90, 2. P. 182–201.

Методы анализа

- **Построение семантических карт**. Многомерное шкалирование позволяет сгруппировать переводные единицы, с одинаковыми свойствами на основе их сходств в маркировании.
- **Сетевой анализ** дает возможность проанализировать, как взаимодействуют различные способы выражения, без привязки к переводным единицам.
- При помощи **факторного анализа** можно оценивать влияние различных признаков на выбор употребления конструкции.

Перенос разметки

- Наличие выравнивания в параллельном корпусе позволяет осуществлять перенос разметки и использовать это в исследованиях.
  - Например, можно разметить семантические признаки или характеристики ситуации (одушевленность участников, наличие негативного воздействия и др.) и перенести эту информацию с одного языка на другой.

Перенос разметки

- Наличие выравнивания в параллельном корпусе позволяет осуществлять перенос разметки и использовать это в исследованиях.
  - Например, можно разметить семантические признаки или характеристики ситуации (одушевленность участников, наличие негативного воздействия и др.) и перенести эту информацию с одного языка на другой.
- Перенос разметки позволяет упростить процесс аннотирования мультиязычного корпуса: достаточно разметить лишь часть данных.

Перенос разметки

- Наличие выравнивания в параллельном корпусе позволяет осуществлять перенос разметки и использовать это в исследованиях.
  - Например, можно разметить семантические признаки или характеристики ситуации (одушевленность участников, наличие негативного воздействия и др.) и перенести эту информацию с одного языка на другой.
- Перенос разметки позволяет упростить процесс аннотирования мультиязычного корпуса: достаточно разметить лишь часть данных.
- Плохо сработает для грамматических характеристик (иначе нам не нужно было бы использовать кучу парсеров)

S. Say, Nominal causal constructions across Slavic: semantic contrasts in a parallel corpus perspective

С. Сай в своей работе [Say 2021b] рассматривает именные причинные конструкции, то есть такие причинные конструкции, в которых событие-причина выражена именной группой.

(1) Маргарита ... заплакала [**от** бол-**и** в руке и ноге]. (Russian)
‘Margarita ... burst into tears from the pain in her arm and leg.’
(2) Я погибаю [**из-за** любв-**и**]. (Russian)
‘I’m perishing on account of love.

Автор анализирует способы маркирования причинных именных групп в нескольких славянских языках. При помощи семантической карты получилось рассмотреть, как в рамках единого семантического пространства распределены предложные показатели, используемые в языках выборки.

S. Say, Nominal causal constructions across Slavic: semantic contrasts in a parallel corpus perspective

С. Сай в своей работе [Say 2021b] рас , то
есть такие причинные конструкции, в к
группой.

(1) Маргарита ... заплакала [**от** бол-
‘Margarita ... burst into tears from the
(2) Я погибаю [**из-за** любв-**и**]. (Russ
‘I’m perishing on account of love.

Автор анализирует способы маркирова
славянских языках. При помощи семан в
рамках единого семантического простр ии,
используемые в языках выборки.

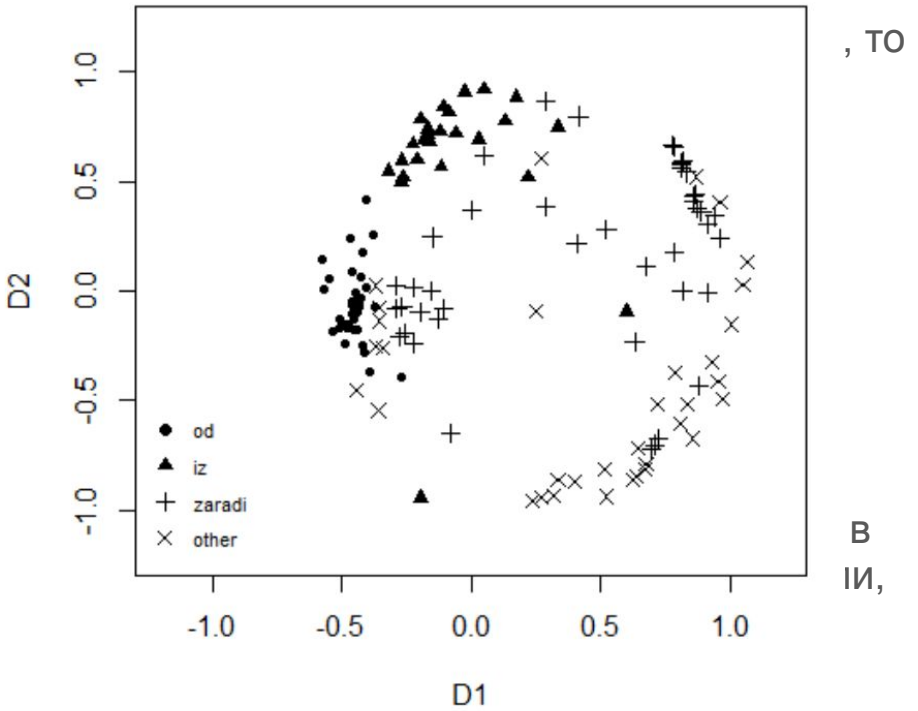


Fig. 7. Bulgarian nominal causal markers

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

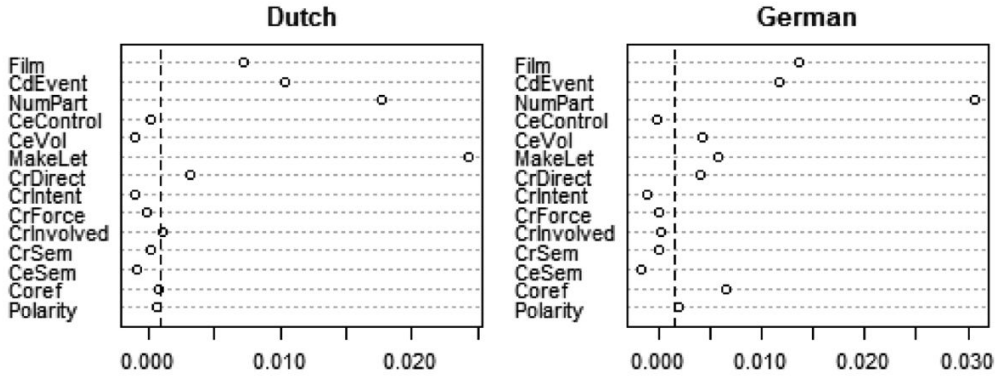
Table 3: Parameters of variation of lexical and analytic causatives operationalized as variables.

	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"	

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

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

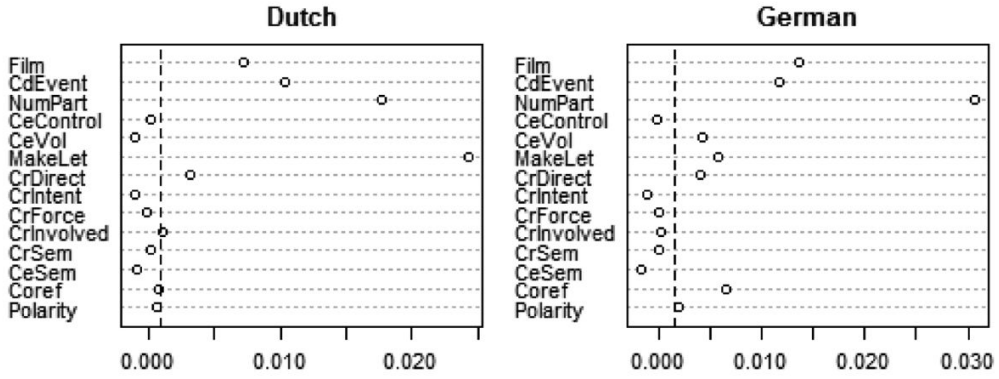
	Variable	Abbreviation	Values	Expectations
1	Aktionsart of the caused event	*CdEvent*	"Nonaction"	Lexical
			"Action"	Analytic
2	Number of main participants	*NumPart*	"2"	Lexical
			"3"	Analytic
3	Control of Causee	*CeControl*	"Yes"	Analytic
			"No"	Lexical
4	Causee acting willingly	*CeVol*	"Yes"	Analytic
			"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"	



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

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"	



Подход:

1) Размечаем признаки на материале одного языка
2) Переносим это на остальные языки
3) Строим модели для языков выборки на основе одинаковых признаков, но разных способов маркирования ситуаций

Создание корпусов

Применение

Преимущества параллельных корпусов

- Охват сразу с нескольких языков
- Возможность применения различных современных методов анализа.
- Работа с ситуациями в одинаковом контекстном окружении. То есть исследователь располагает примерами с одинаковой семантической и прагматической составляющей. Это не разрозненные примеры из разных источников, это сопоставимый материал.

Применение

Преимущества параллельных корпусов

- Охват сразу с нескольких языков
- Возможность применения различных современных методов анализа.
- Работа с ситуациями в одинаковом контекстном окружении. То есть исследователь располагает примерами с одинаковой семантической и прагматической составляющей. Это не разрозненные примеры из разных источников, это сопоставимый материал.

Применение:

➢ Сравнительные исследования, типология
➢ Машинный перевод

Чего мы хотим от параллельных корпусов?

=Что придется реализовать технически

Чего мы хотим от параллельных корпусов?

=Что придется реализовать технически

- Много разных языков
- Много текстов и соответствий
- Несколько этапов обработки

Pipeline создания параллельного корпуса

Сбор текстов (проверка, что все со всем параллельно)

Предобработка текстов (разбиение на предложения, токенизация)

Выравнивание по предложениям

Выравнивание по словам

Разметка

Разметка и выравнивание

Какие проблемы могут возникнуть?

Как можно решать эти проблемы?

Разметка и выравнивание

Какие проблемы могут возникнуть?

Как можно решать эти проблемы?

- Нет универсальных инструментов, а нам нужно обработать очень разнообразные данные

Важные понятия

Выравнивание (alignment)

Переводная единица (translational unit)

Переводные эквиваленты (translational equivalents)

Язык оригинала (source language)

Язык перевода (target language)