Parallel Corpora

LiRI corpus workshop — Johannes Graën

Wednesday 24th November, 2021

Overview

1. What are parallel corpora? (definition)
2. Where do they come from? (methods of corpus compilation)
3. What are they good for? (use cases)

$\Rightarrow$ Representation and querying of parallel corpora

Parallel corpora

- corpora = collections of language samles
- (text) corpora = collections of texts
- parallel (text) corpora = collections of translated texts

$\Rightarrow$ correspondence on various levels (key word: alignment)

Alignment

Definition (term is ambiguous)

Alignment refers to

- a correspondence relation between different parts of a parallel corpus at a particular level, e.g.
  - a book and its translation to another language
  - a sentence and its translation
  - a word or multiword expression (”potencia visual” ↔ ”sight”)
- a set of those relations
- the process of identifying those sets of relations
- the level of correspondence (word alignment, sentence alignment, ...)

What can be aligned?

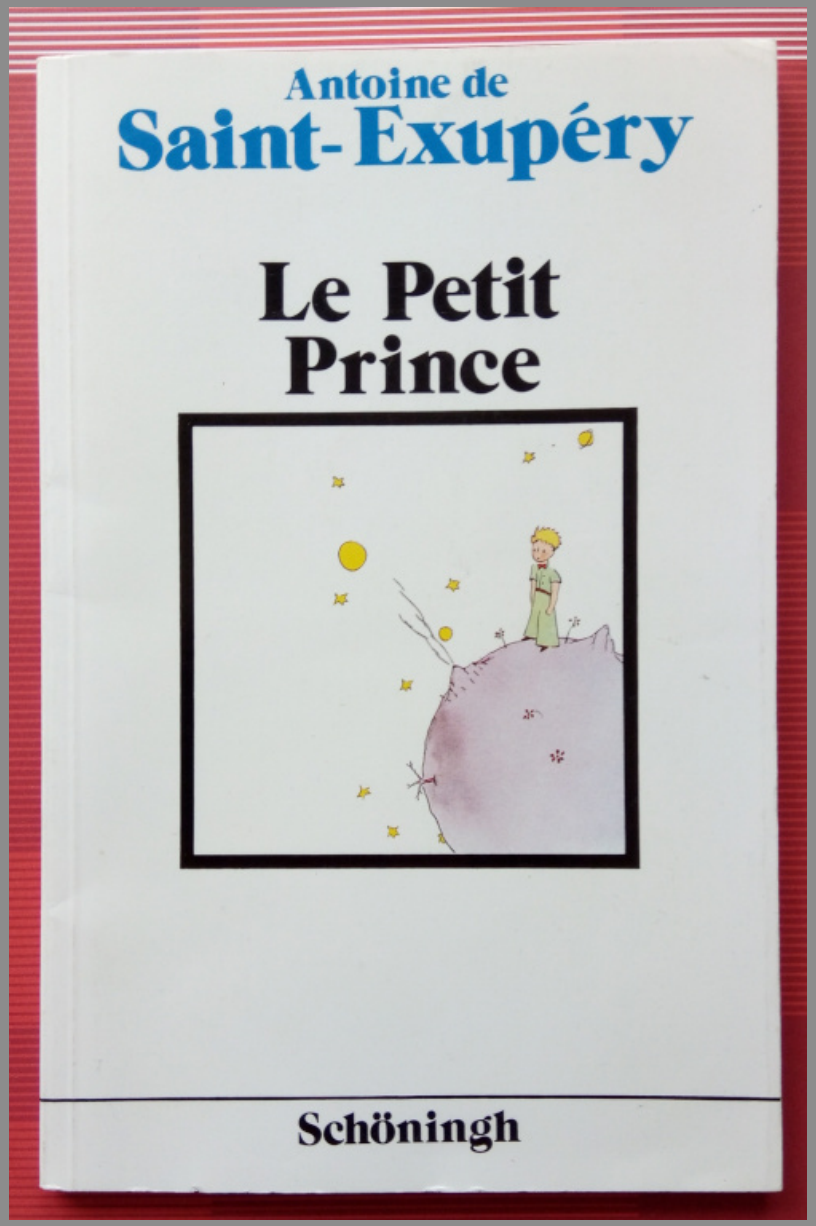
- documents (books, protocols, leaflets, construction manuals, ...)
- any kind of subordinated structural text units (chapters, agenda items, ...)
- paragraphs (?)
- sentences/segments
- sub-sentential units (chunks, constituents, ...)
- words/tokens
- morphems (?)

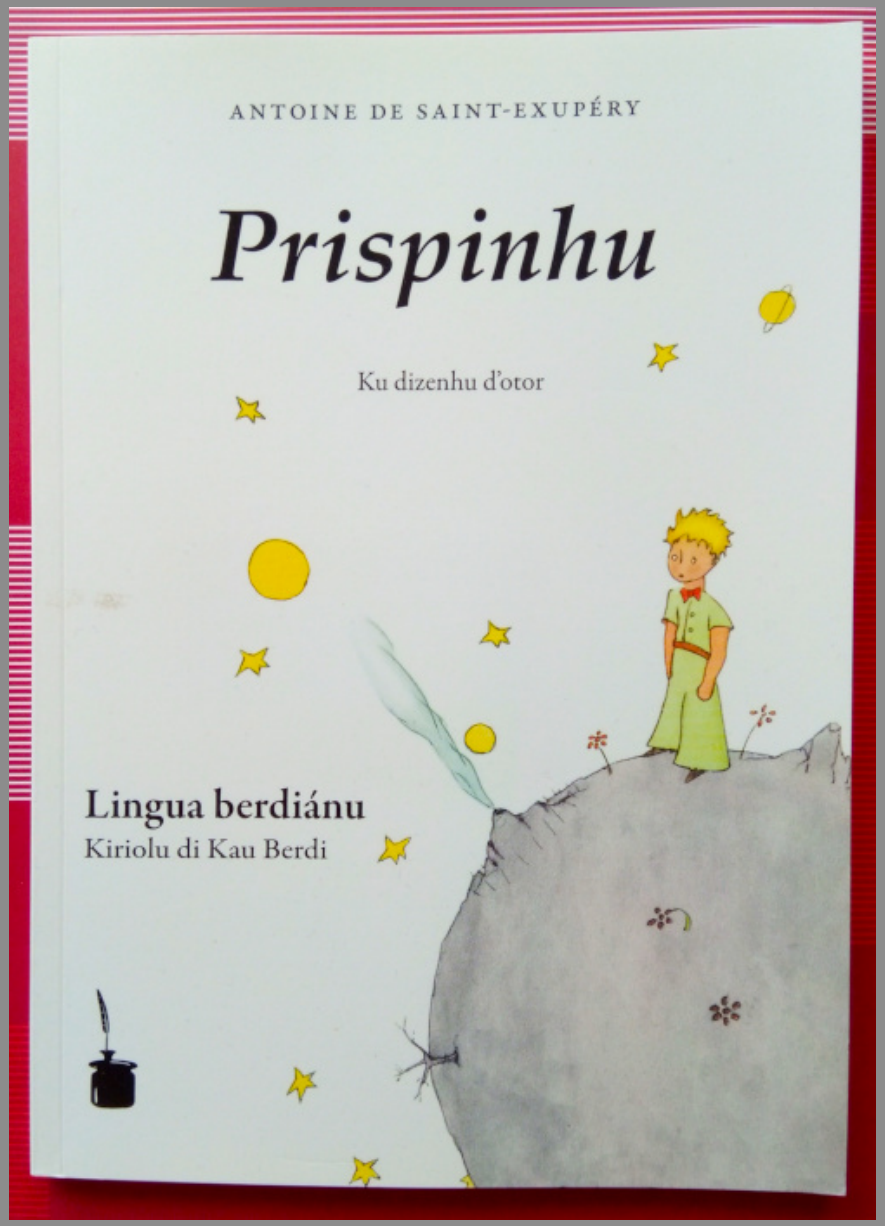
Document alignment

- also: text alignment, document linking
- correspondence in most cases given[1]
- documents might be ordered (e.g. plenary sessions with an indicated date) but typically are not (e.g. books)
- metadata varies from source to source
- typically 1:1 correspondences
- there might be null alignments[2] (depending on the document collection)
- translation direction can be indicated on this level[3]

[1] (Tiedemann 2012)
[2] units that have no counterpart
[3] but, for example, turn-based translation in Europarl

Document alignment – top layer: book





Document alignment – subordinate layer: chapter

XI

La seconde planète était habitée par un vaniteux:

– Ah! Ah! Voilà la visite d'un admirateur! s'écria de loin le vaniteux dès qu'il aperçut le petit prince.

Car, pour les vaniteux, les autres hommes sont des admirateurs.

– Bonjour, dit le petit prince. Vous avez un drôle de chapeau.

– C'est pour saluer, lui répondit le vaniteux. C'est pour saluer quand on m'acclame. Malheureusement il ne passe jamais personne par ici.

– Ah oui? dit le petit prince qui ne comprit pas.

– Frappe tes mains l'une contre l'autre, conseilla donc le vaniteux.

Le petit prince frappa ses mains l'une contre l'autre. Le vaniteux salua modestement en soulevant son chapeau.

– Ça, c'est plus amusant que la visite au roi, se dit en lui-même le petit prince. Et il recommença de frapper ses mains l'une contre l'autre. Le vaniteux recommença de saluer en soulevant son chapeau.

Après cinq minutes d'exercice le petit prince se fatigua de la monotonie du jeu:

– Et, pour que le chapeau tombe, demanda-t-il, que faut-il faire?

38

XI

Na sugundu planéta, éra un vaidós ki ta morába la. Di lonji, sim-e odja prispinhu, vaidós fla:

- Abé-Mariâ! Dja parsi dimirador!

Pamódi, pa vaidós, tud'algen k'é ka el é si dimirador.

Prispinhu fla-l:

- Bon diâ, fórti txapeu stránhu!

Vaidós kudí-l si:

- É pa N ta tra. É pa N ta tra algen óra k-ês ta da-m pálmu. Más paxénxa, a-li nunka ka ta pása nungen.

Prispinhu ka ntendi, e purgunta:

- É simé?

Vaidós fla-l si:

- Nhu bá ta da ku mó na kunpanheru.

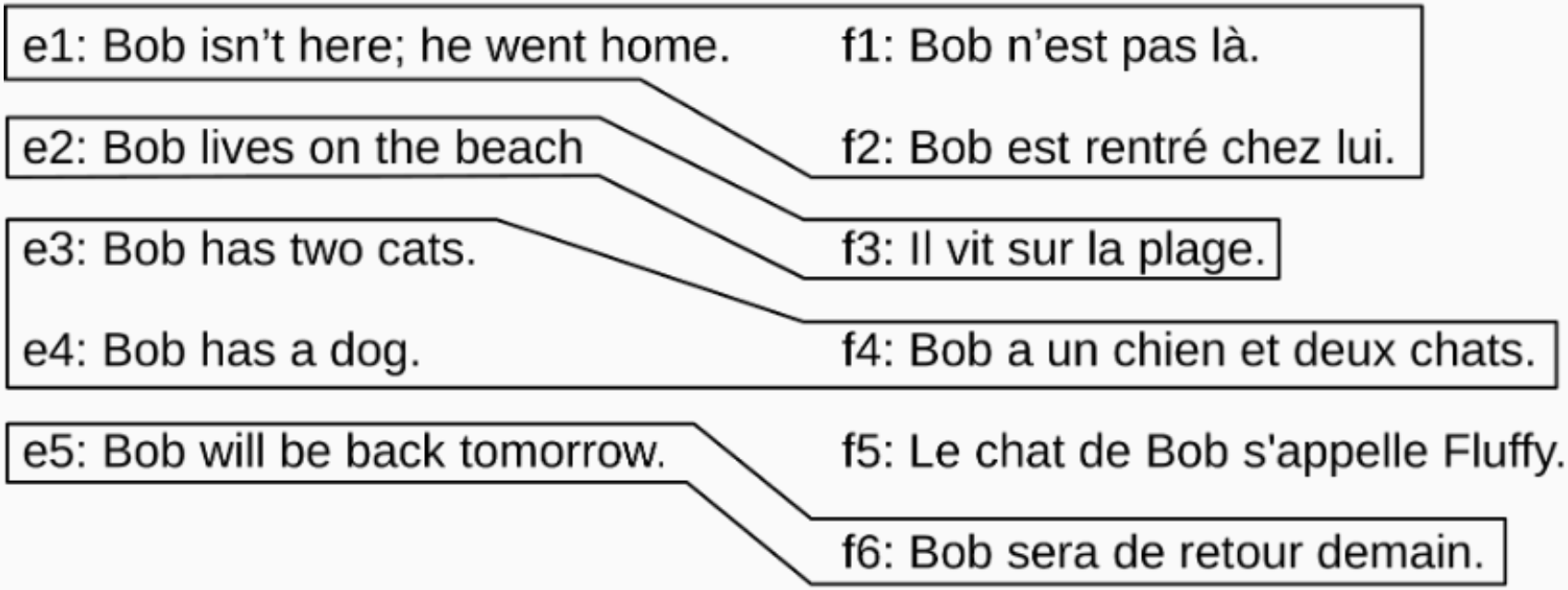
Prispinhu da ku mó na kunpanheru. Vaidós tra-l txapeu ku ár mudéstu.

42

Sentence alignment

- variable use of punctuation marks: better split sentences and perform alignment on segments
- often ordered, no overlapping alignment links (monotonicity)
- frequently 1:1 correspondences (depends on text type and mode of translation)
- null alignments may occur when information is added or omitted during translation (strongly dependent on text type)

Sentence alignment – (toy) example



(Thompson and Koehn 2019)

1:1 alignments

	English	German	Spanish
1	Of course, I have said it often before, I am no lover of capitalism.	Selbstredend bin ich, wie schon häufig gesagt, kein Freund des Kapitalismus.	Aunque por supuesto, como ya he dicho en otras muchas ocasiones, no soy un seguidor del capitalismo.
2	Capitalism is not an object of my affection, it is simply a means to an end.	Der Kapitalismus hat nicht meine Sympathie, er ist lediglich Mittel zum Zweck.	No es una de mis predilecciones, es simplemente un medio para conseguir un fin.
3	In any case, I do often like to distinguish between capitalism and liberalism.	Auf jeden Fall pflege ich oft zwischen Kapitalismus und Liberalismus zu unterscheiden.	En cualquier caso, a menudo me gusta hacer una diferencia entre el capitalismo y el liberalismo.
4	Clearly, my socialist friends are keen to combine these, yet the two things are not the same.	Meine sozialistischen Freunde werfen natürlich gerne beide zusammen, sie sind aber nicht das Gleiche.	Está claro que mis amigos socialistas tienden a combinarlos, pero se trata de dos cosas distintas.
5	Even I have to say it.	Das möchte ich doch einmal klarstellen.	Aunque tenga que decirlo.

1:n alignments

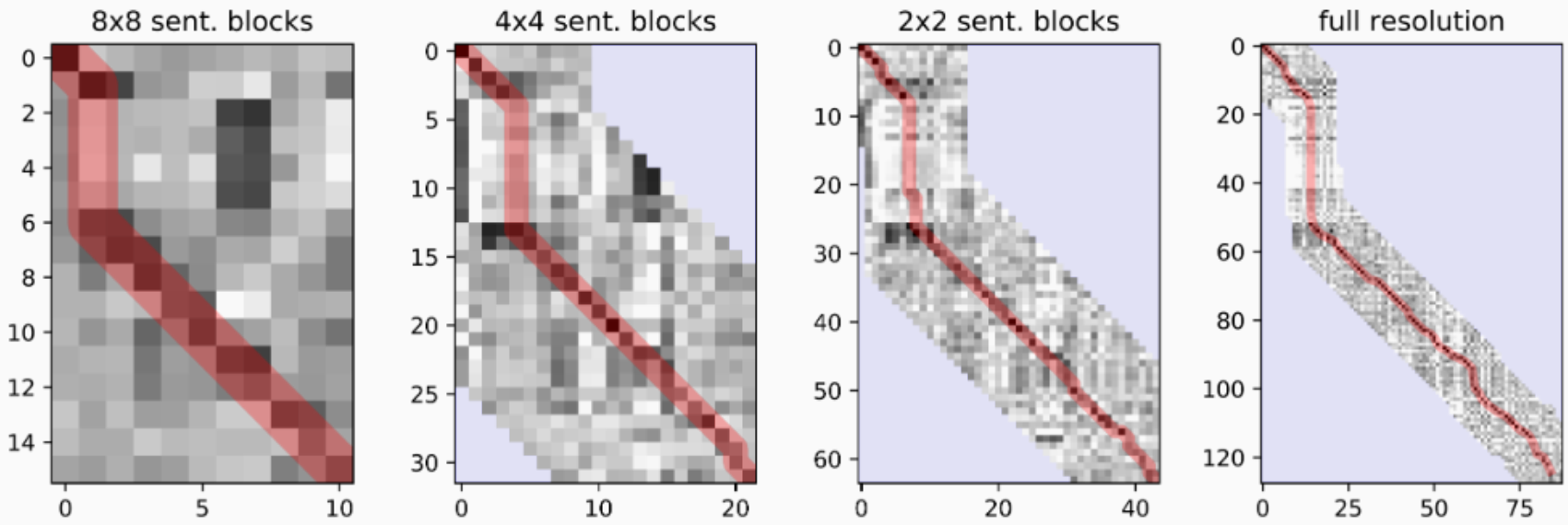
	English	German	Spanish
1	I hear MEPs who, I think, still believe in the effectiveness, honour and values of Europe, as well as feeling a certain pride in being European.	Europaabgeordnete, die meiner Meinung nach doch Grundsätze wie Effizienz und Ehre sowie die Wertvorstellungen Europas hochhalten und einen gewissen Stolz empfinden, Europäer zu sein – diese Abgeordneten höre ich ständig lamentieren und ein Sündenbekenntnis ablegen, dass an alledem im Grunde Europa schuld sei.	He escuchado las intervenciones de diputados al PE que, desde mi punto de vista, aún creen en la eficacia, el honor y los valores de Europa y que además sienten cierto orgullo de ser europeos.
2	I hear them constantly complaining and apologising.		Les he oído quejarse y pedir disculpas de un modo constante.
3	Basically this is all meant to be Europe's fault.		Todo esto significa esencialmente que es culpa de Europa y no puedo aceptarlo.
4	I do not accept that.	Dem stimme ich nicht zu.	

Added/omitted information

	English	German	Spanish
1	We are currently working on a PNR package.	Wir arbeiten derzeit an einem Fluggastdatensatz-paket (Passenger Name Record, PNR).	En estos momentos, estamos trabajando sobre el paquete de registro de nombres de los pasajeros (PNR).

Sentence alignment – monotonicity

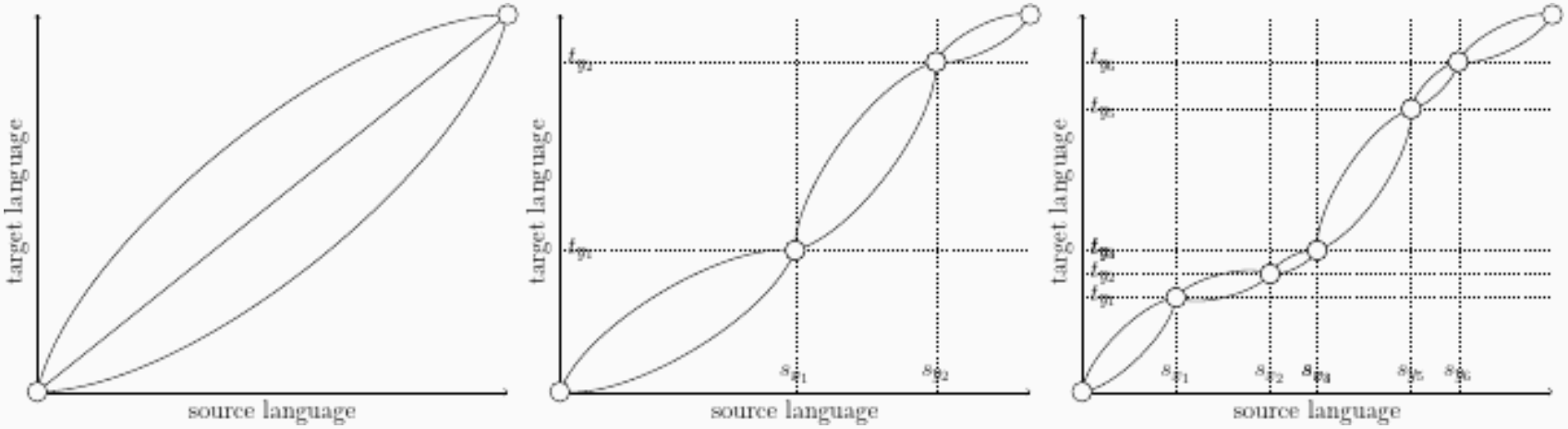
Under the assumption of monotonicity and infrequent null alignments, we can find the overall alignment around the diagonal:



(Thompson and Koehn 2019)

Sentence alignment – monotonicity

The same idea but as "iterative refinement" approach (anchors):



(Tiedemann 2012)

Sentences aligned – and now?

Wenn ihre Katze Bier trinkt , ist dies vielleicht der Grund , warum sie krank ist .

If her cat is drinking beer , then that is probably what is making the cat ill .

Si su gato bebe cerveza , probablemente sea eso lo que enferma al gato .

Jos hänen kissansa juo olutta , kissa tulee sairaaksi .

Si son chat boit de la bière , c' est probablement cela qui le rend malade .

Se il suo gatto beve birra , probabilmente è per quello che sta male .

Als haar kat bier drinkt , wordt hij daar waarschijnlijk ziek van .

Se o gato da senhora deputada bebe cerveja , provavelmente é isso que o traz doente .

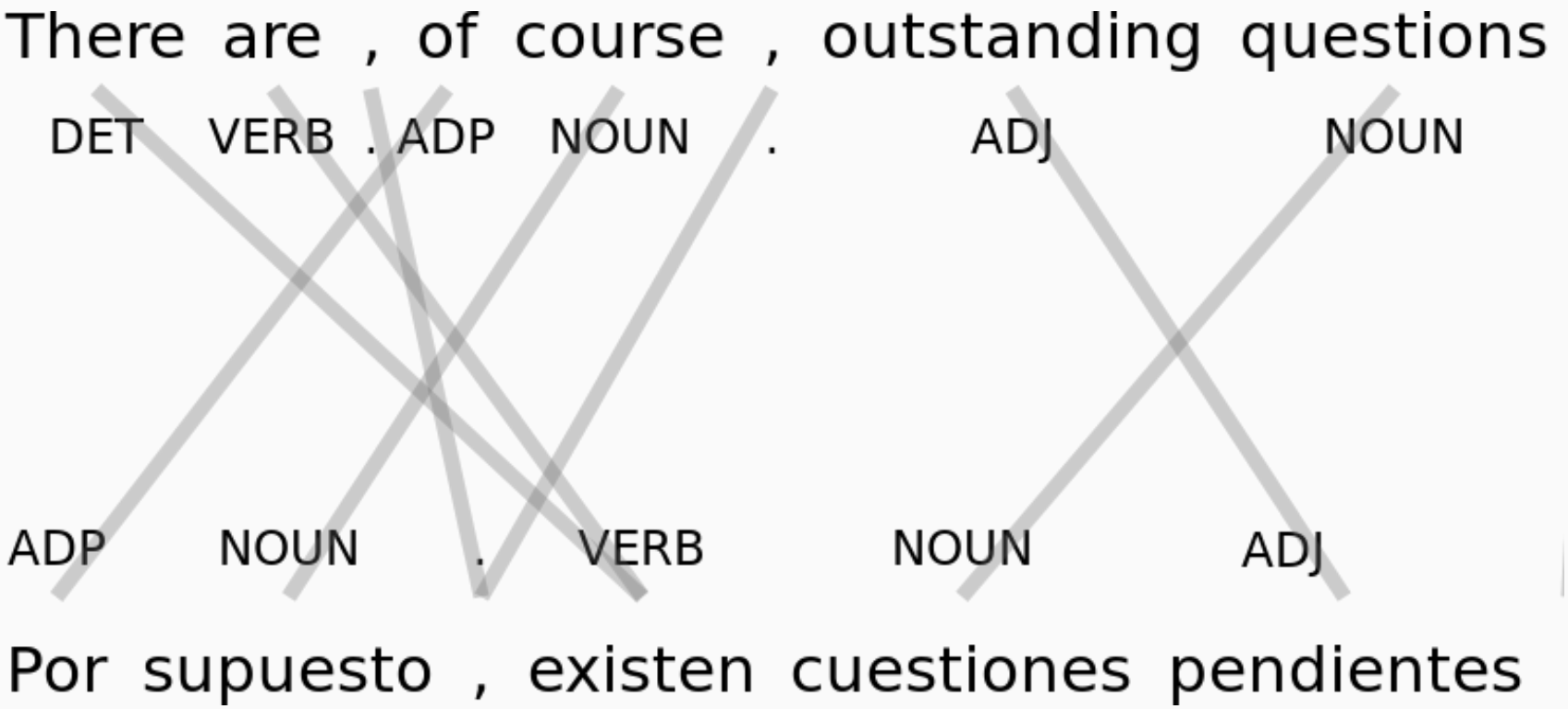
Om hennes katt dricker öl så är det troligen det som gör katten sjuk .

Caroline Jackson, 22/10/1997

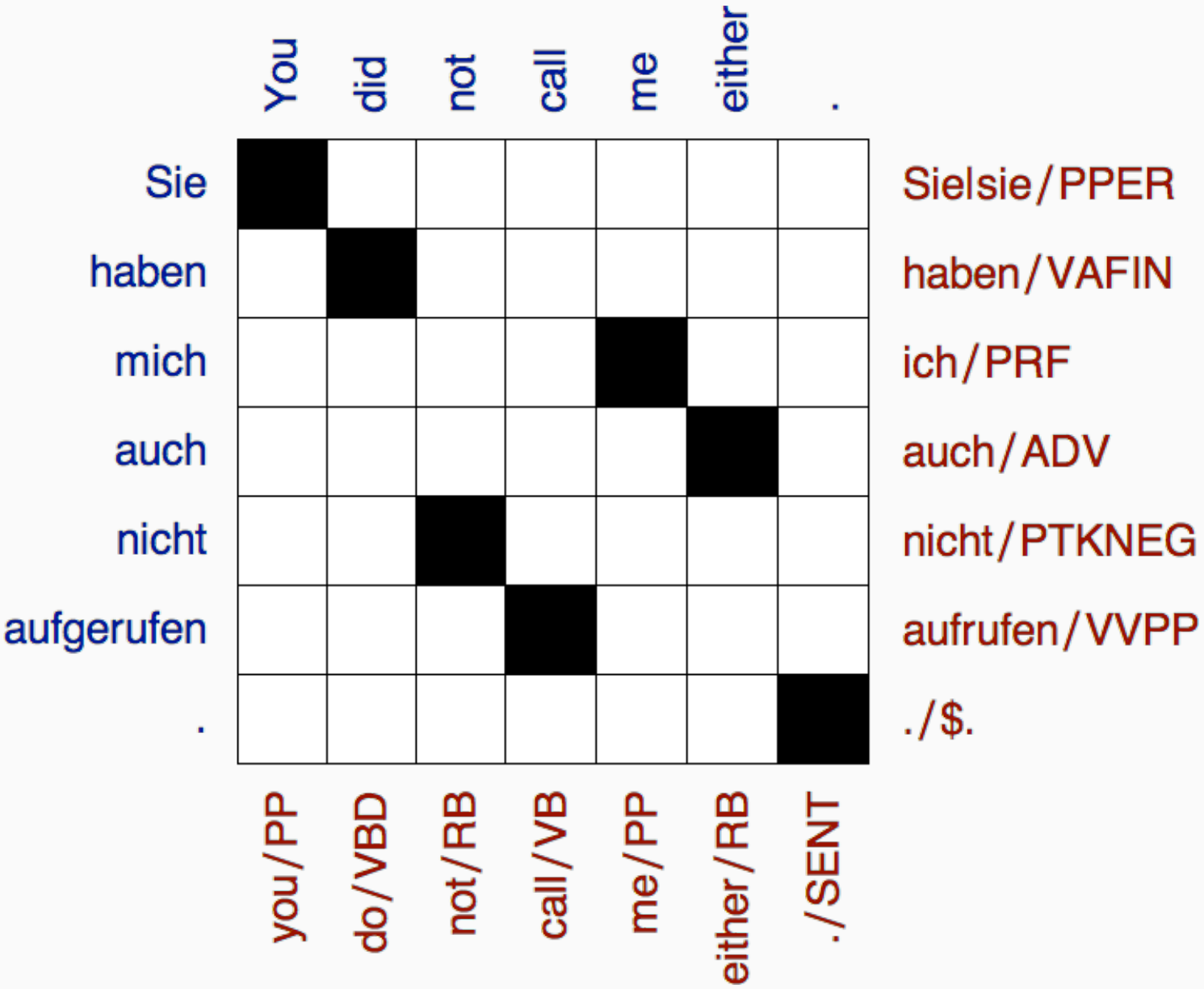
Word alignment

- aligns actually any token provided (requires tokenization)
- no order can be assumed; alignment only by chance monotonic
- 1:1 alignments most frequent, but many different ratios observable
- the concept of ”words” may differ drastically between typologically less-related languages
- null alignments are frequent (e.g. function words)
- $\Rightarrow$ word alignment comes with a significant error rate and only the aligned word may be of little help for particular applications

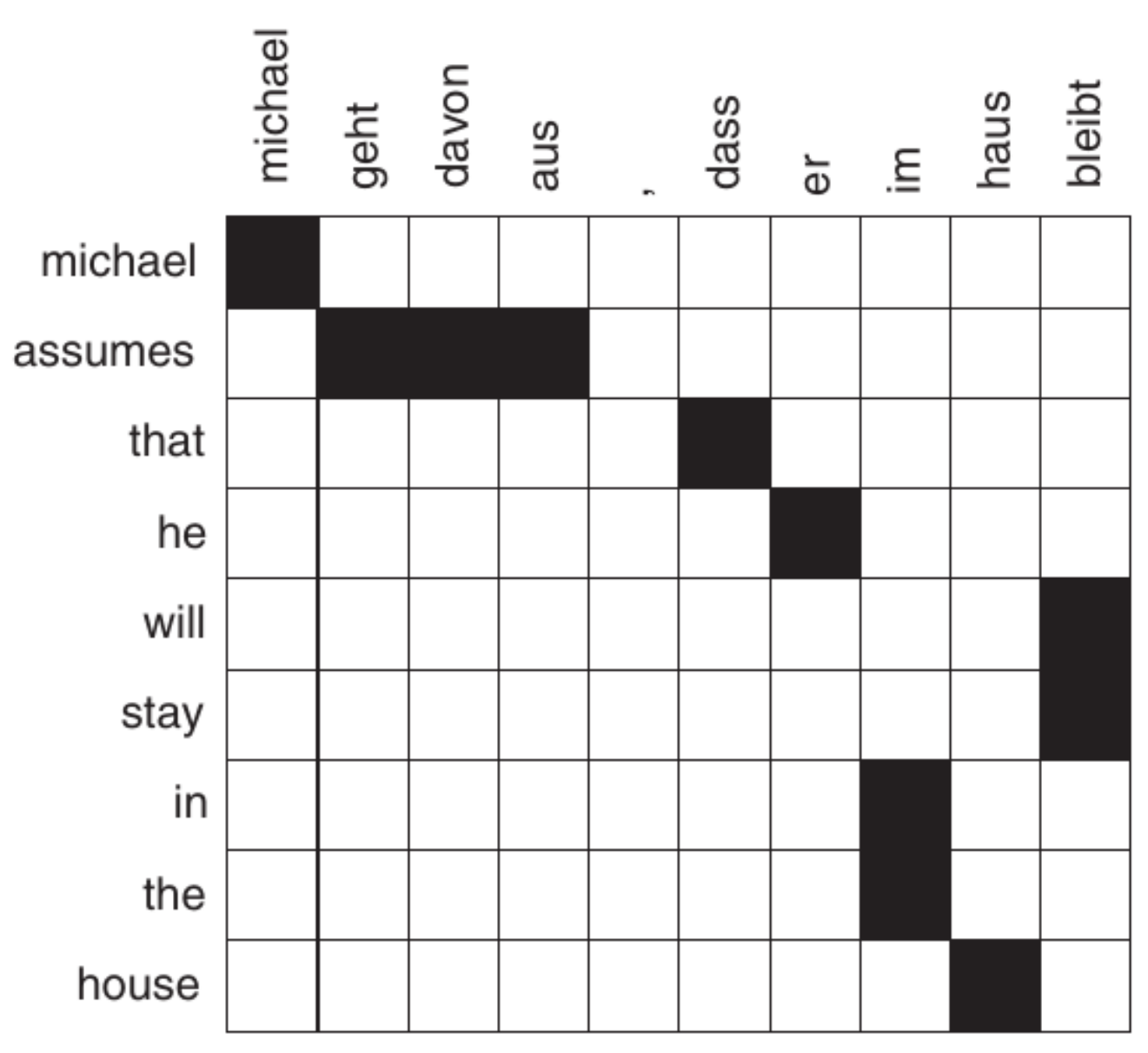
Alignment links



Alignment matrix (1)



Alignment matrix (2)



Multiparallel word alignment

Ein neues politisches Klima entsteht nach und nach.

A new political climate is gradually emerging.

Un nuevo clima político está emergiendo gradualmente.

Uusi poliittinen ilmapiiri on vähitellen muotoutumassa.

Un cadre politique nouveau voit progressivement le jour.

Stopniowo wyłania się nowy klimat polityczny.

Își face apariția treptat un nou climat politic.

Postopoma nastaja novo politično vzdušje.

Ett nytt politiskt klimat håller gradvis på att växa fram.

Hierarchical alignment – nested alignment units

Wir möchten nicht die Katze im Sack kaufen.

Nous ne voulons pas acheter chat en poche.

Não estamos interessados em comprar gato por lebre.

We are not interested in buying a pig in a poke.

Vi är inte intresserade av att köpa grisen i säcken.

Alignment tools – documents

- if correspondence is not already known, resort to a comparison of metadata, document size, cognates, etc. to identify the most likely set of corresponding documents
- use language identification if the source material might be unclean (e.g. data collected from the internet)

Alignment tools – sentences

Features used include:

- sentence length
- lexical correspondence (possibly induced from the data)
- cognates and extra-linguistic data (e.g. numbers, URLs)

Alignment tools – sentences

Popular and state-of-the-art sentence aligners:

- Hunalign (Varga et al. 2005)
  uses an (induced) dictionaries and sentence lengths
- Gargantua (Braune and Fraser 2010)
  designed for asymmetrical parallel corpora[4]
- Bleualign (Sennrich and Volk 2010)
  based on machine translation
- Vecalign (Thompson and Koehn 2019)
  based on bilingual sentence embeddings

[4] many null alignments

Alignment tools – words

Popular and state-of-the-art word aligners:

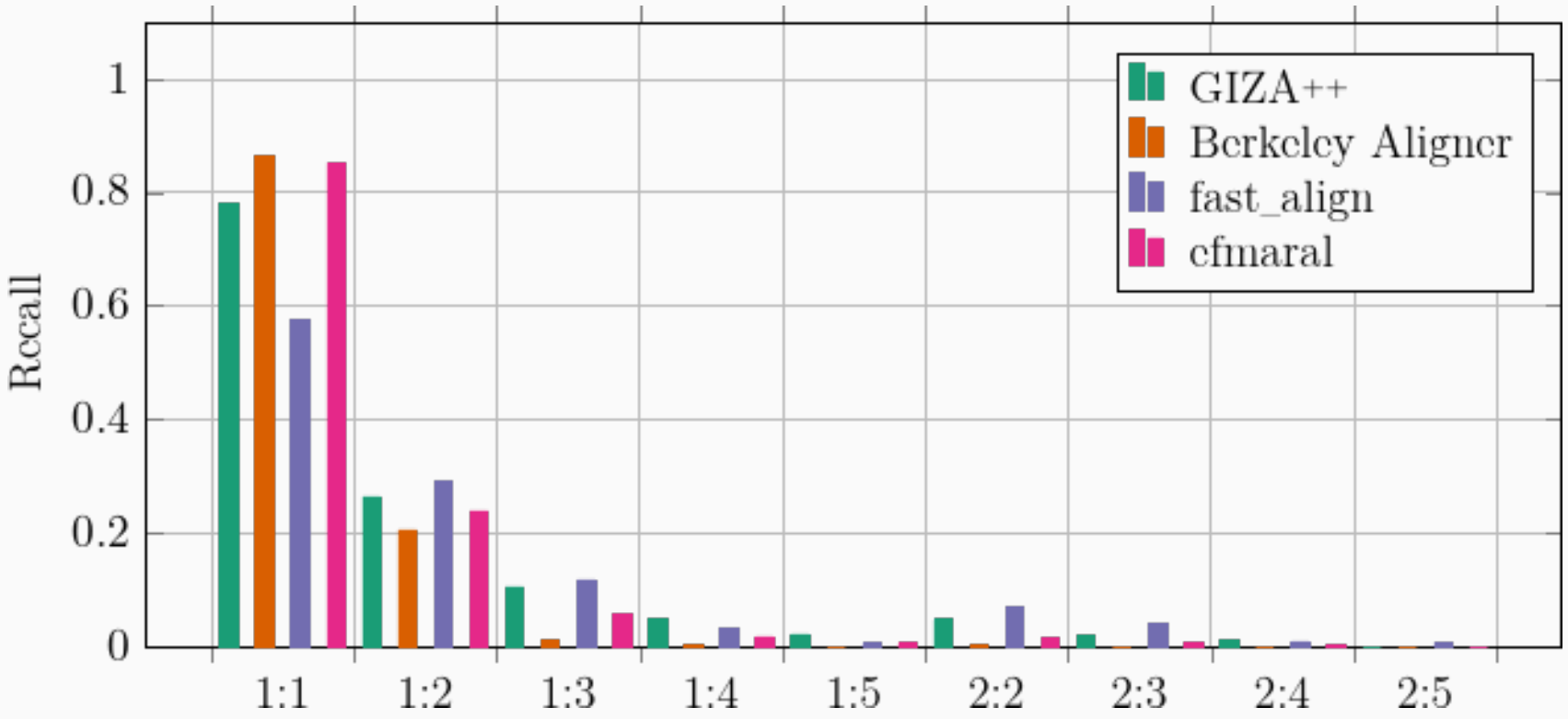
- GIZA++ (Och and Ney 2003)
  implements the "IBM models", asymmetric models
- BerkeleyAligner (Liang, Taskar, and Klein 2006)
  adds a probability threshold and a (symmetric) HMM
- fastalign (Dyer, Chahuneau, and Smith 2013)
  very fast, but less reliable
- efmaral/eflomal (Östling and Tiedemann 2016)
  uses Bayesian mathematical; can store trained model
- SimAlign (Sabet, Dufter, Yvon, and Schütze 2020)
  based on bilingual word embeddings
- AWESOME (Dou and Neubig 2021)
  based on bilingual word embeddings

Alignment tools – words

- most word aligner generate unidirectional alignments, i.e. 1:n alignments[5]
- those aligners need to train two models (one for each direction) and results need to be symmetrized
- combining the output of several aligners (e.g. by majority vote) can lead to better results (better precision, lower recall)

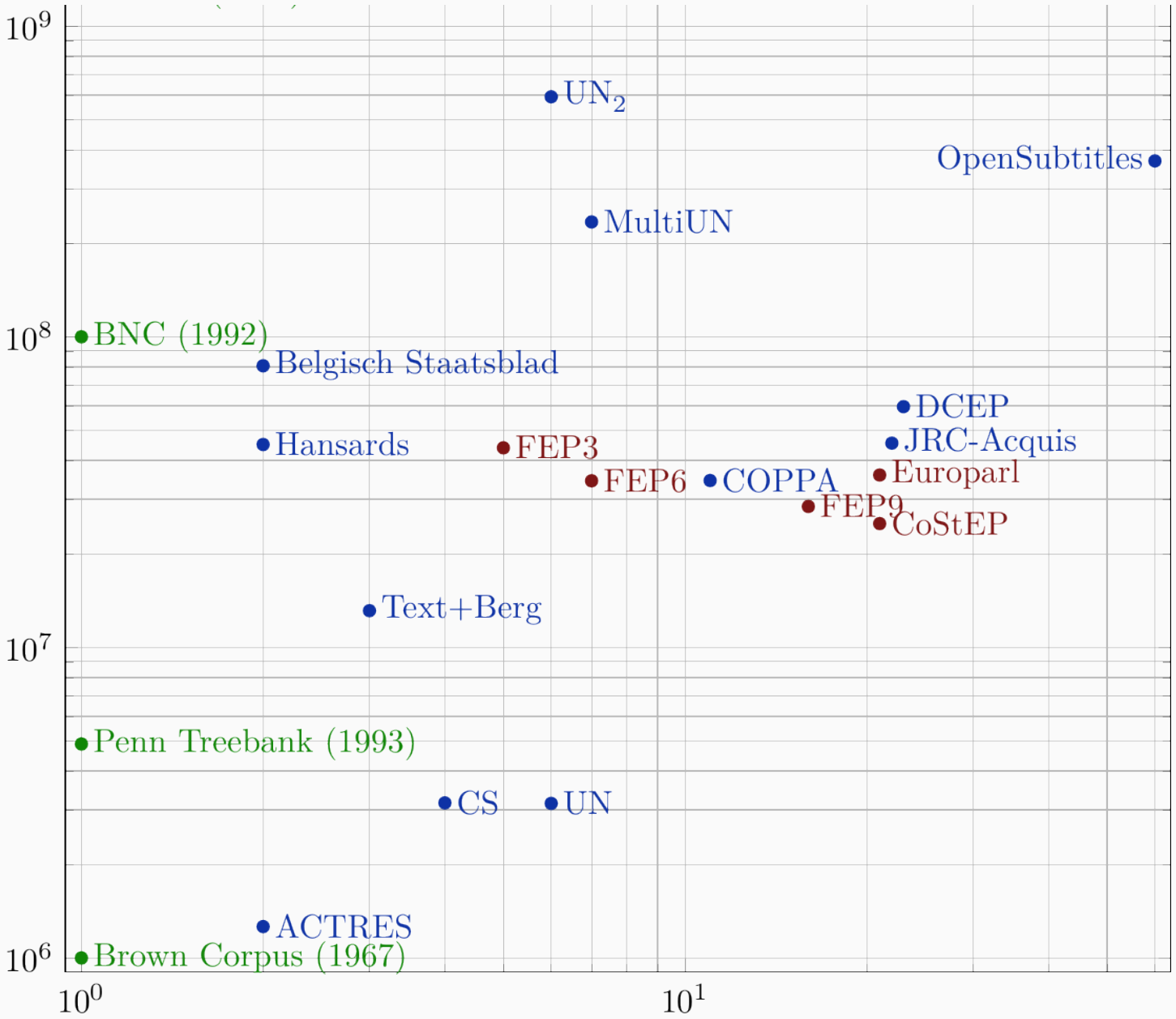
[5] fot the case "potencia visual" ↔ "sight", 'potencia' and 'visual' can both be aligned to sight when going from Spanish to English, the other way round, 'sight' can only be aligned to one of them)

Alignment types found by different aligners



Existing parallel corpora

Some parallel corpora (log #tokens/log #languages)



Zurich Parallel Corpus Collection (PaCoCo)

Parallel corpora in various formats with heterogeneous metadata

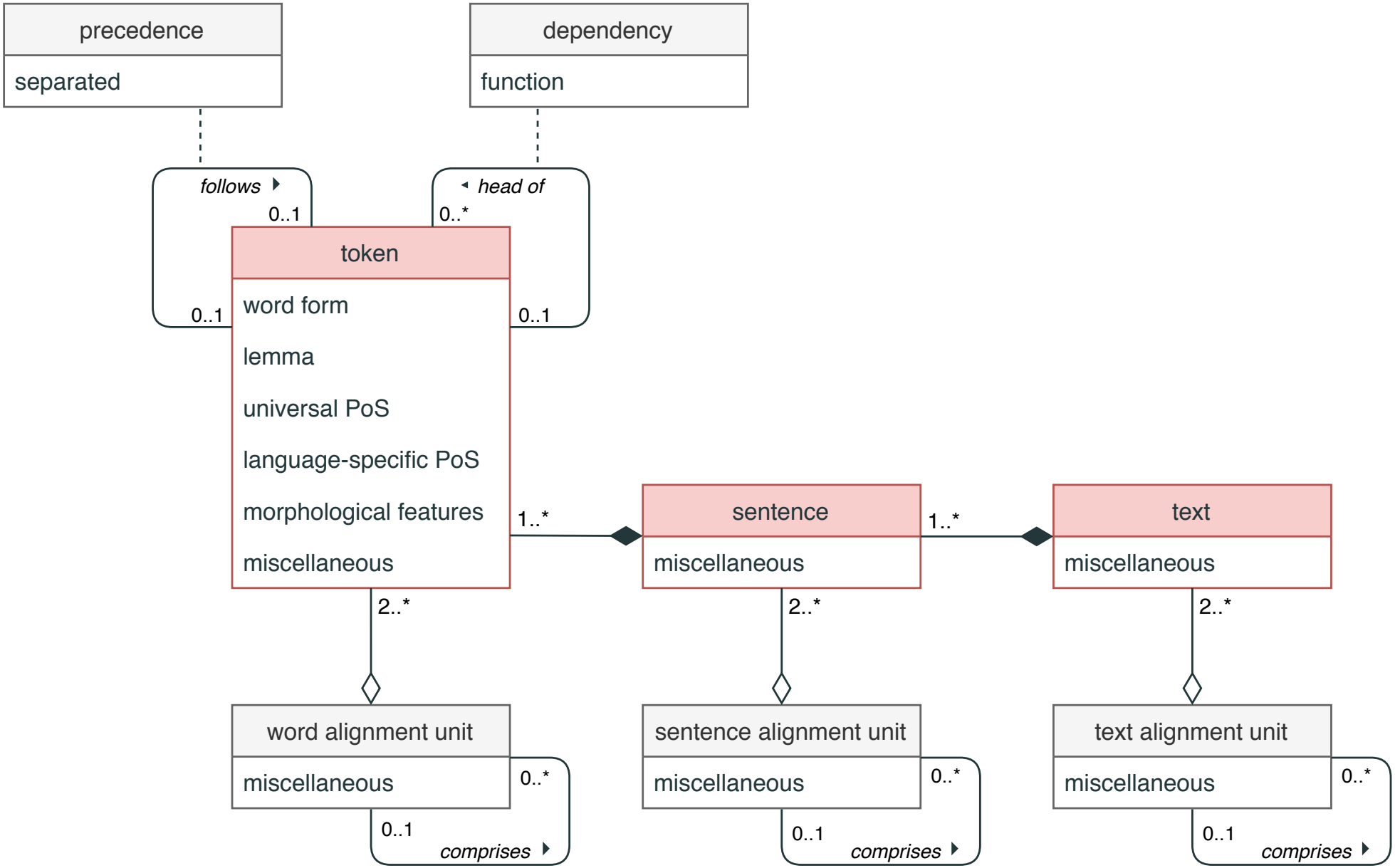
- 10 different corpora
- 7 different formats
  - tree-structure (XML, TEI)
  - tabular (.tsv, .csv)
- non-standardized
- heterogeneous metadata and annotations
- difficult to **combine**, **extract** & **exploit** the data at our finger tips

PaCoCo – corpora

	languages	tokens	years	alignment
Sparcling	de, en, es, fr, it + 11	454.7m	15	word
SLC	de, fr	11.4m	—	word
Rumantsch Grischun	de, rm	0.9m	—	word
Swatchgroup Geschäftsbericht	de, gsw	0.2m	—	word
Medi-Notice	de, fr, it	58.9m	—	word
Text + Berg	de, fr, it, rm, gsw, en	52.6m	150	sentence
CS Bulletin	de, en, es, fr, it	61.6m	120	sentence
Horizons	de, en, fr	2.9m	14	document

https://pub.cl.uzh.ch/purl/PaCoCo

PaCoCo – data model



Linguistic Applications

Contrastive analysis: variable article use

- *de* “In unseren einzelnen Mitgliedstaat und gemeinsam als **Europäische Union** müssen wir [...]”
- *en* “In our individual Member States, and collectively as **the European Union**, we must [...]”
- *es* “En nuestros respectivos Estados miembros y, de manera colectiva, en **la Unión Europea** debemos [...]”
- *it* “Sia nei singoli Stati membri che collettivamente, come **Unione europea**, dobbiamo esercitare [...]”
- *pt* “Em cada um dos nossos Estados-Membros, e colectivamente enquanto **União Europeia**, temos que [...]”
- *sv* “I våra enskilda medlemsstater, och samfällt som **Europeiska unionen**, måste vi [...]”

Elena Callegaro (2017). “Parallel Corpora for the Investigation of (Variable) Article Use in English: A Construction Grammar Approach”. PhD thesis. University of Zurich

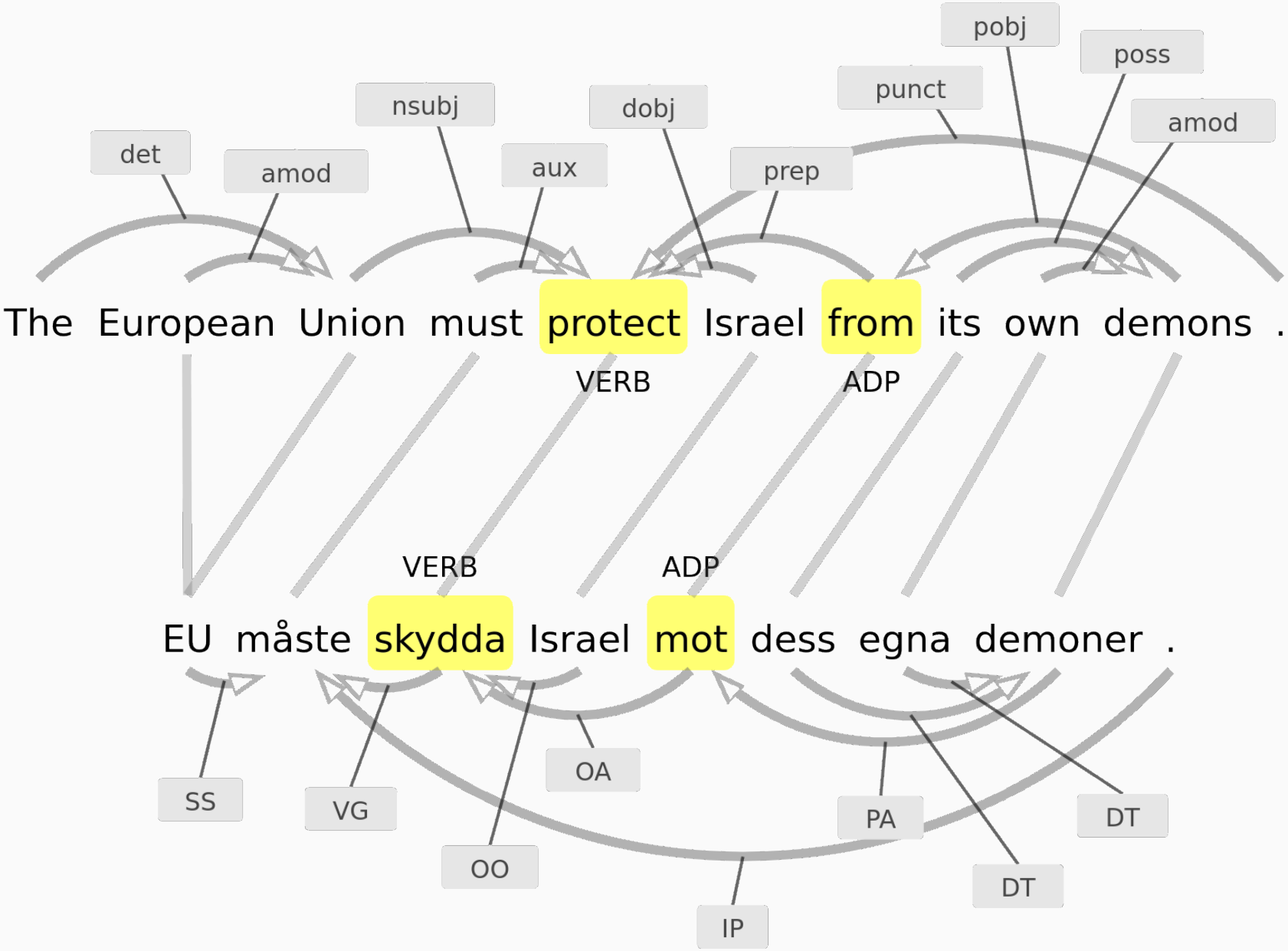
Multilingual phraseology: discontinuous constructions

1	When does the Council intend to reach a decision on the establishment of this future observatory?
	När kommer rådet att fatta beslut om att inrätta detta framtida organ?
2	It has attempted to reallocate budgetary resources from the Progress programme to the microfinance facility before the European Parliament has reached a decision.
	Den har försökt omfördela budgetresurser från Progressprogrammet till instrumentet för mikrokrediter innan Europaparlamentet har fattat ett beslut.
3	Furthermore, the decision-making process itself can be unclear, as the convention submits proposals and the Intergovernmental Conference has to reach decisions.
	Dessutom kan det bli oklart kring själva beslutsfattandet, eftersom konventet lägger fram förslag och regeringskonferensen måste fatta beslut.
4	When the matter comes before Parliament, therefore, we often have to reach our decisions very quickly if we want to make the internal market a reality for the citizens of Europe.
	Kommer ärendet sedan till parlamentet, måste vi ofta fatta mycket snabba beslut, eftersom vi vill öppna den gemensamma marknaden för medborgarna.

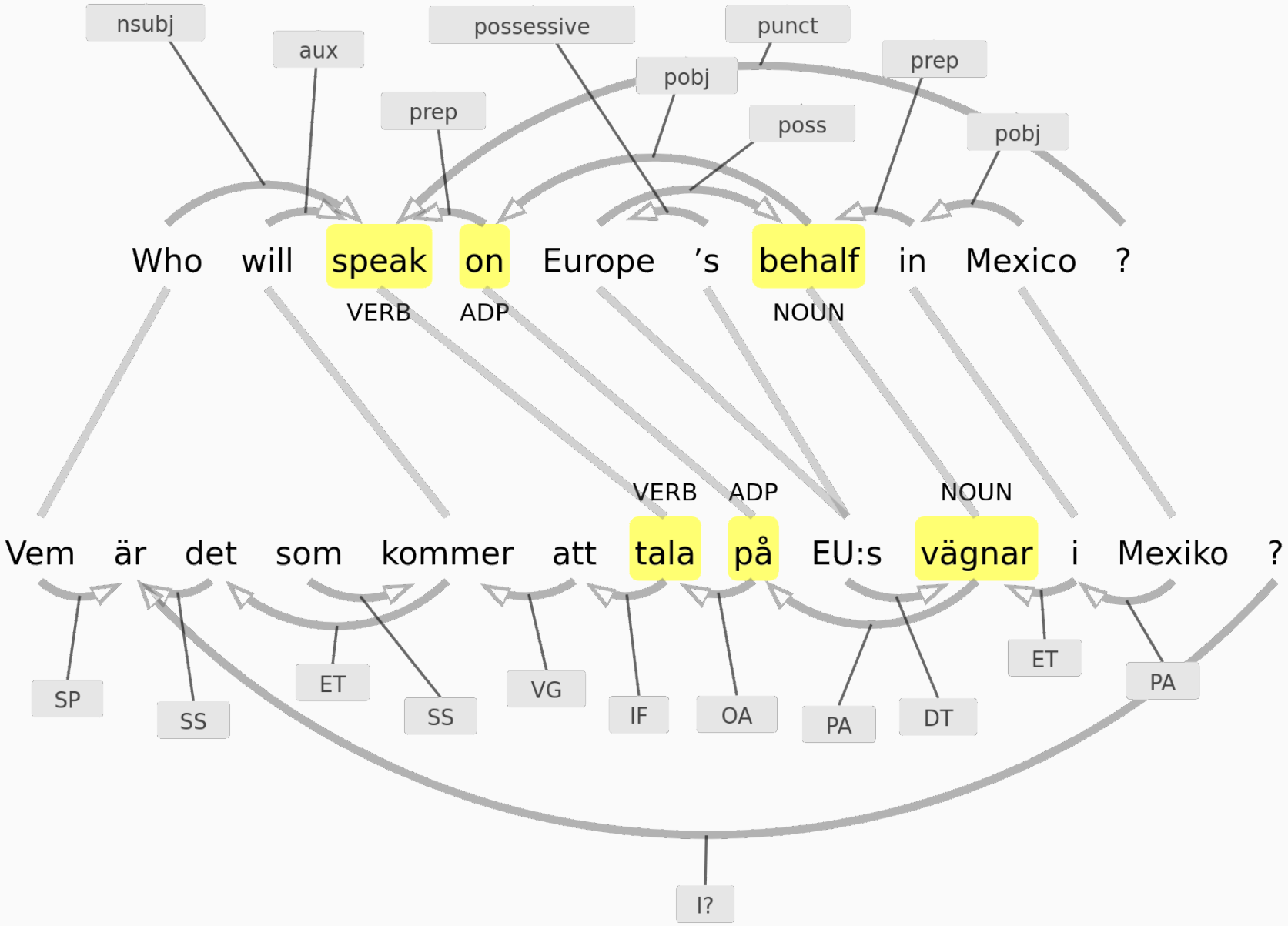
Constellations

- syntactic parsing combined with word alignment
- statistical association measures help identifying elements of surprise
- in the case of support verb construction this is the verb correspondence
- $\Rightarrow$ high surprisal is an indicator for idiomaticity
- *https://pub.cl.uzh.ch/purl/constellations*

Constellations



Constellations

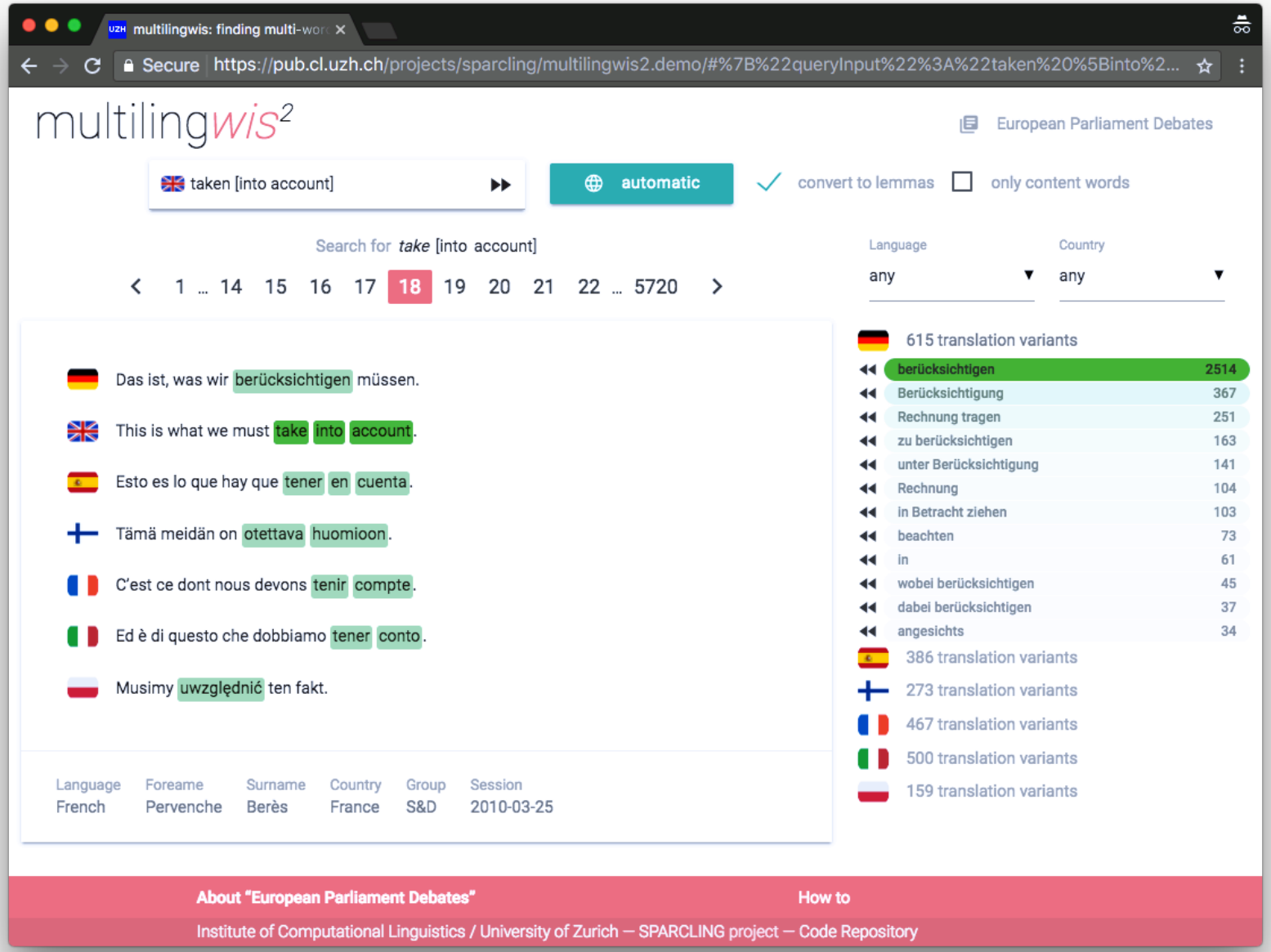


Constellations: support verb constructions

rank	German		English		Italian		count
1	annehmen	Gestalt	take	shape			39
2	darstellen	Präzedenzfall	set	precedent			10
3	bekämpfen	Armut	reduce	poverty			4
4	schaffen	Präzedenzfall	set	precedent			78
5	haben	Vorrang	take	precedence			47
1	schaffen	Abhilfe			porre	rimedio	36
2	schaffen	Präzedenzfall			costituire	precedente	23
3	gewinnen	Oberhand			prendere	sopravvento	8
4	machen	Mühe			prendere	briga	9
5	schaffen	Klarheit			fare	chiarezza	6
1			take	look	dare	occhiata	21
2			take	precedence	dare	precedenza	4
3			send	condolence	esprimere	condoglianza	5
4			take	precedence	avere	precedenza	92
5			have	illusion	fare	illusione	20

Tools

Multilingwis (Multilingual Word Information System)



Multilingwis – Translation equivalents

184 translation variants

berücksichtigen	395
Rechnung tragen	89
Berücksichtigung	50
Rechnung	26
zu berücksichtigen	21
unter Berücksichtigung	17
in Betracht ziehen	15
in	13
beachten	10
in berücksichtigen	6
Rechnung zu tragen	6
Rücksicht nehmen	6

140 translation variants

tener en cuenta	386
en cuenta	162
tomar en cuenta	67
tomar en consideración	58
tenerse en cuenta	17
en	16
cuenta	9
a	8
en consideración	8
tener en consideración	7
toma en consideración	7
se en cuenta	6

- order of lemmas maintained
- less frequent = more (alignment) errors

Multilingwis – Examples

Es gibt wirklich besorgniserregende Entwicklungen, denen schon früher hätte Rechnung getragen werden sollen.

There are some really worrying facts which should have been taken into account earlier.

Hay datos realmente preocupantes que debían haber sido tomados en consideración anteriormente.

On olemassa joitakin todella huolestuttavia tietoja, jotka olisi pitänyt ottaa aiemmin huomioon.

Il existe des données vraiment préoccupantes qui auraient dû être préalablement prises en considération.

Vi sono dati realmente preoccupanti che avrebbero dovuto essere tenuti in considerazione molto prima.

Original language	Forename	Surname	Country	Group	Session
Spanish	Pedro	Marset Campos	Spain	EUL/NGL	1998-09-16

- ordered by (common) length
- metadata shown (if available)
- interactive visualization (directed alignment links)
- option to filter for particular translation equivalents

Multilingwis

https://pub.cl.uzh.ch/purl/multilingwis

Semantic overlap by alignment frequencies

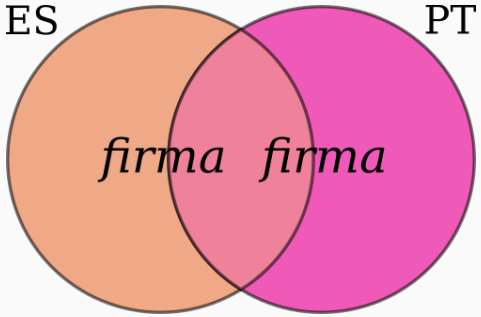
- Based on word alignment and lemmatization
- Reflects the probability of a lemma λ_s in the source language to be aligned with a lemma λ_t in the target language. E.g.:

relative frequency		absolute frequency	
$p_a(EN\ cow \mid ES\ vaca)$	= 0,82	$f_a(EN\ cow \mid ES\ vaca)$	= 305
$p_a(EN\ cattle \mid ES\ vaca)$	= 0,12	$f_a(EN\ cattle \mid ES\ vaca)$	= 44
$p_a(EN\ beef \mid ES\ vaca)$	= 0,01	$f_a(EN\ beef \mid ES\ vaca)$	= 4

- The probabilities (= relative frequencies) of all possible lemmas λ_i in the target language (i.e. the elements of the entire corresponding row) sum up to 1 by definition.

Coinciding alignments

- Two lemmas can be aligned with the same (foreign) lemma:



- We calculate frequencies for common lemmas:

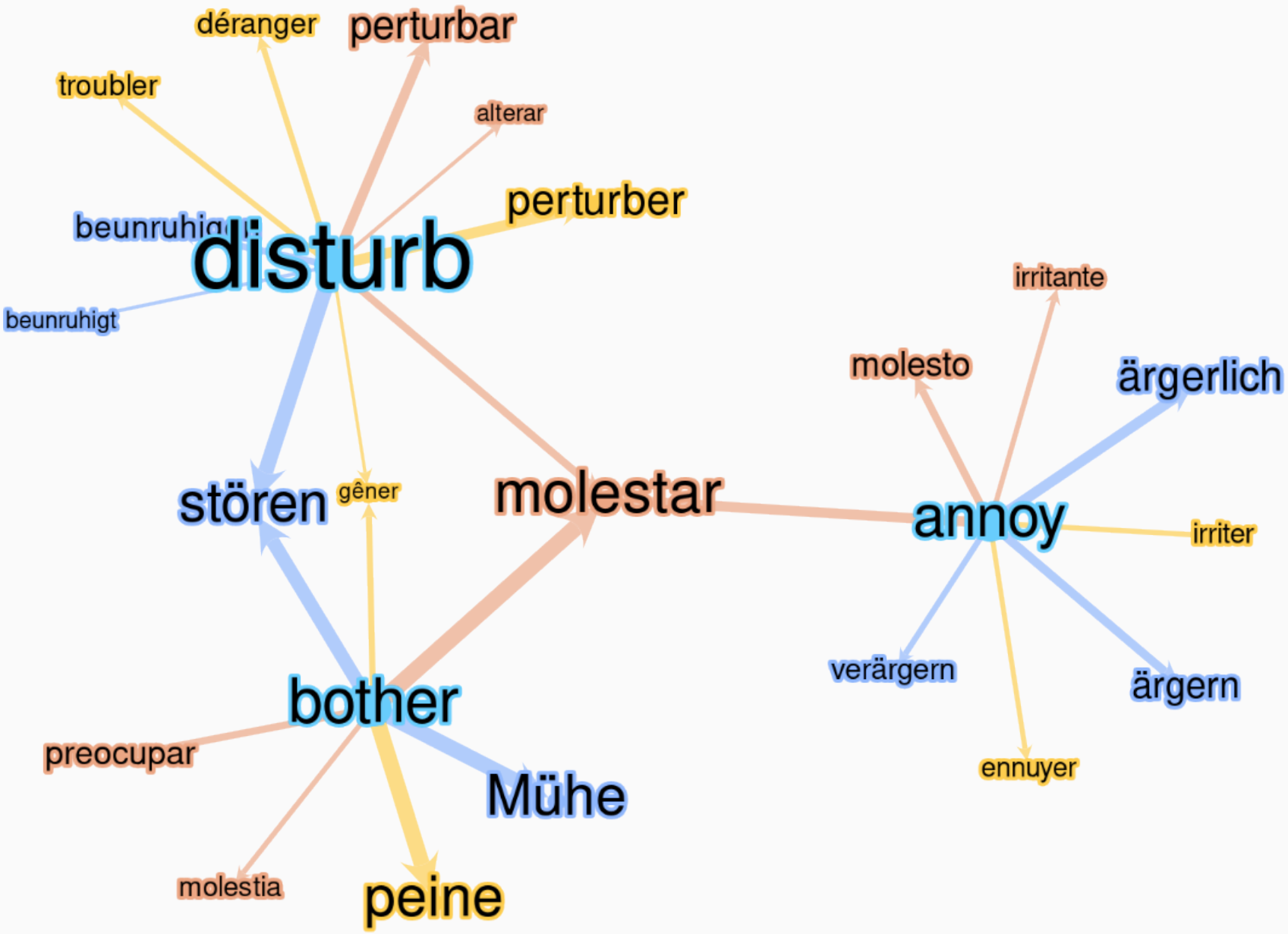
$$f_{\cap}(\lambda_1, \lambda_2|\lambda_x) = min\,(f_a(\lambda_1, \lambda_x), f_a(\lambda_2, \lambda_x)) \tag{1}$$

$$p_{\cap}(\lambda_1, \lambda_2|\lambda_x) = min\,(p_a(\lambda_x|\lambda_1), p_a(\lambda_x|\lambda_2)) \tag{2}$$

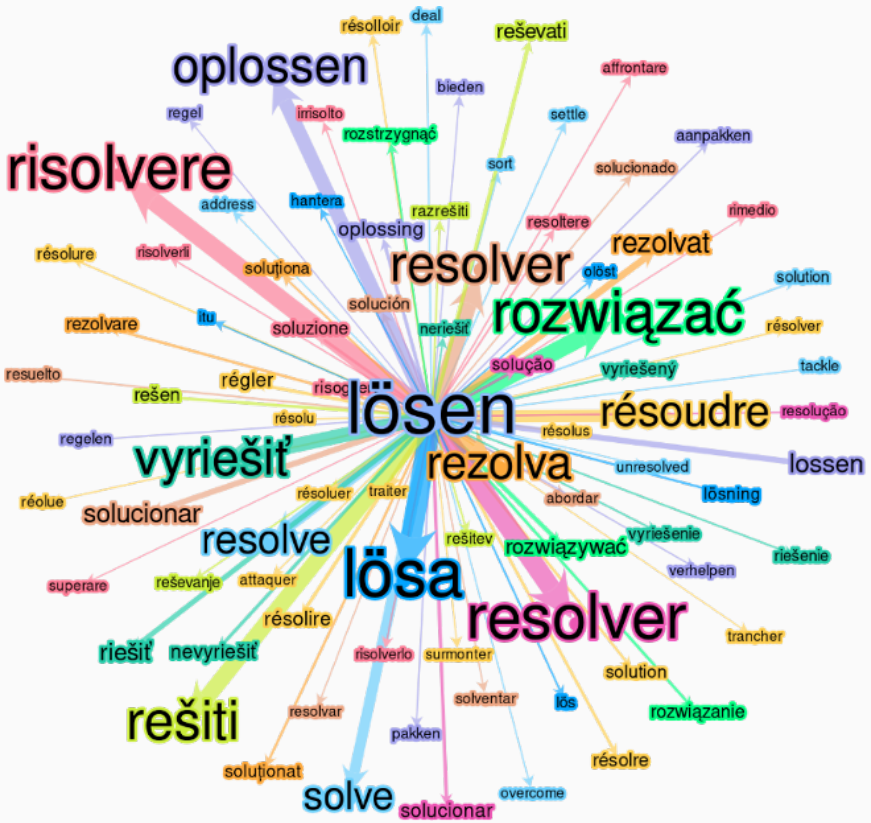
- The overlap measure takes into account the absolute frequency:

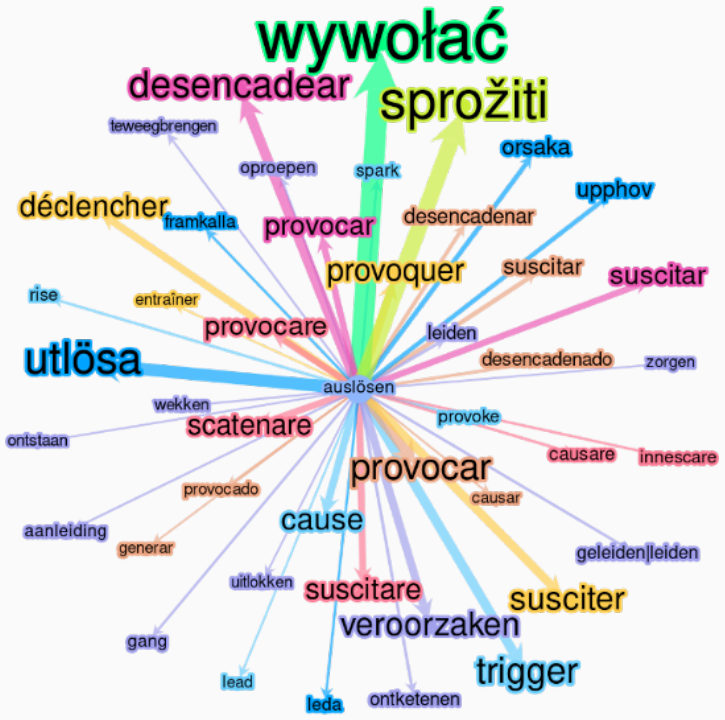
$$O_a(\lambda_1, \lambda_2) = \frac{\sum_{\lambda_x} log(f_{\cap}(\lambda_1, \lambda_2|\lambda_x) + 1) \cdot p_{\cap}(\lambda_1, \lambda_2|\lambda_x)}{\sum_{\lambda_x} log(f_{\cap}(\lambda_1, \lambda_2|\lambda_x) + 1) + \epsilon} \tag{3}$$

Semantic overlap by alignment frequencies



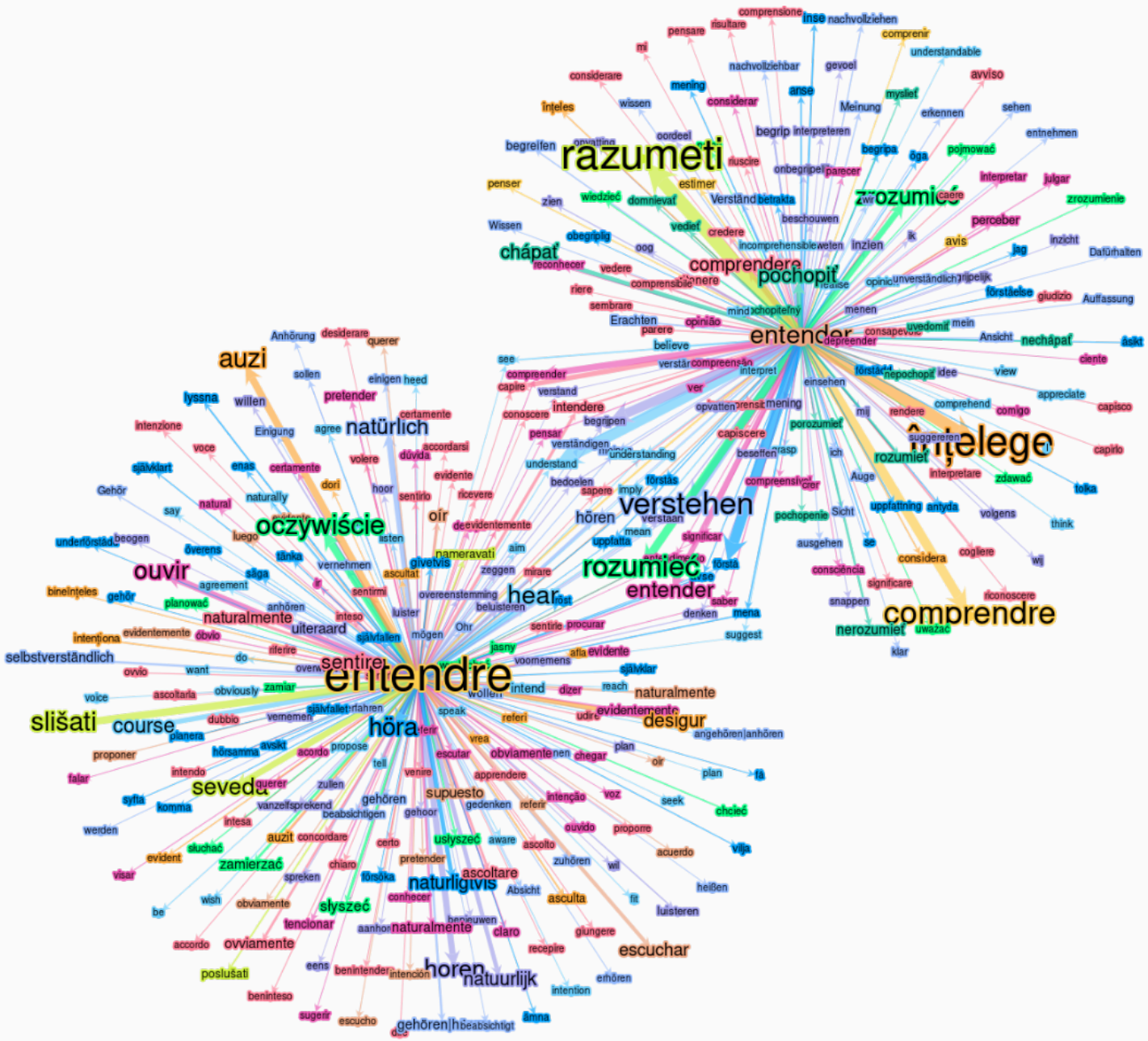
Example: no overlap (de: lösen ↔ de: auslösen)





Example: mostly overlapping (de: steigen ↔ de: ansteigen)

Example: (not so) false friends



Example: false friends (more than two lemmas)

Alignment overlap tool

https://pub.cl.uzh.ch/purl/alignment_overlap

PaCLE

- search tool for language learning from parallel corpora
- parallel sentences from OpenTitles (currently)
- users can approve/disapprove and correct examples (crowdsourcing)

https://demo.spraakbanken.gu.se/johannes/PaCLE/

Discussion

Corpus data types (sketch from 2015)