University of Zurich^UZH

Parallel Corpora — Tools and Applications

PaCor 2021 Workshop

Johannes Graën

Friday 25th June, 2021

About this workshop

Parallel corpora
What are parallel corpora (technically)?

Alignment & alignment tools
What is alignment and how can it be obtained?

Corpus exploration & visualization
What can we do with tools based on aligned parallel corpora?

Utilization for linguistics & language learning
How can we make use of (word-)aligned parallel corpora?

About me

Universität Stuttgart

upf. Universitat Pompeu Fabra *Barcelona*

JUGEND FORSCHT

GÖTEBORGS UNIVERSITET

University of Zurich UZH

FNSNF

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

Remarks on workshop format

Corpora & tools

- Corpora and tools are available online
- Alignment tools use different input and output format
- Training on large corpora requires substantial resources
- Training on small corpora yields suboptimal models
- All state-of-the-art software runs on command line only
- ⇒ Hands-on exercises with those tools won’t work

More interactive parts

- Two group exercises to experiment with (online) tools
- Questions, comments and discussions desired
- Please use the ”raise hand” feature or the chat

Introduction

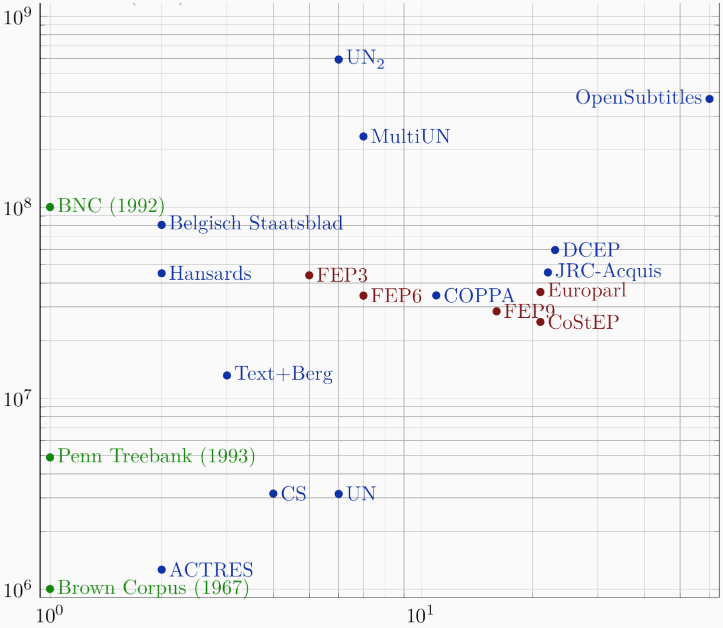
Some characteristics

- collections of produced language in two languages
- ”language”: text, speech, video recordings (but also multimodal)
- ”languages”: typically two different natural languages
- parallel corpora of the same language include learner corpora (with ”corrections”), different translations of the same original work (e.g. the Bible), etc.
- bitexts: parallel text corpora
- multiparallel corpora: parallel corpora in more than two languages
- comparable corpora: same topics but not translations of each other (e.g. Wikipedia articles)

Compilation of parallel corpora

- find a sufficiently large number of translated texts
- sources are often multilingual/multinational institutions
- licenses are less problematic when the translations are funded by public money (vs. e.g. literary works)
- once the material is available it needs to be processed:
  1. document alignment/linkage
  2. sentence alignment
  3. word alignment
- each processing step depends on its predecessor
- not all of them might be necessary for a given purpose
- … but word-aligned parallel corpora allow us to get the most out of the material

Parallel corpora in two dimensions (log. tokens/log. languages)



Selected resources

Collections (data available)

- OPUS – collection of translated texts[1]
- PaCoCo – Zurich Parallel Corpus Collection[2]
- Parallel corpora in CLARIN

Other resources

- PaGes – German/Spanish parallel Corpus, semi-automatically sentence-aligned
- CLUVI – manually sentence-aligned parallel corpora, several corpora related to Galician

[1] (Tiedemann 2009, 2012)

[2] (Graën, Kew, Shaitarova, and Volk 2019)

PaCoCo – motivation

Parallel corpora in various formats with heterogeneous metadata

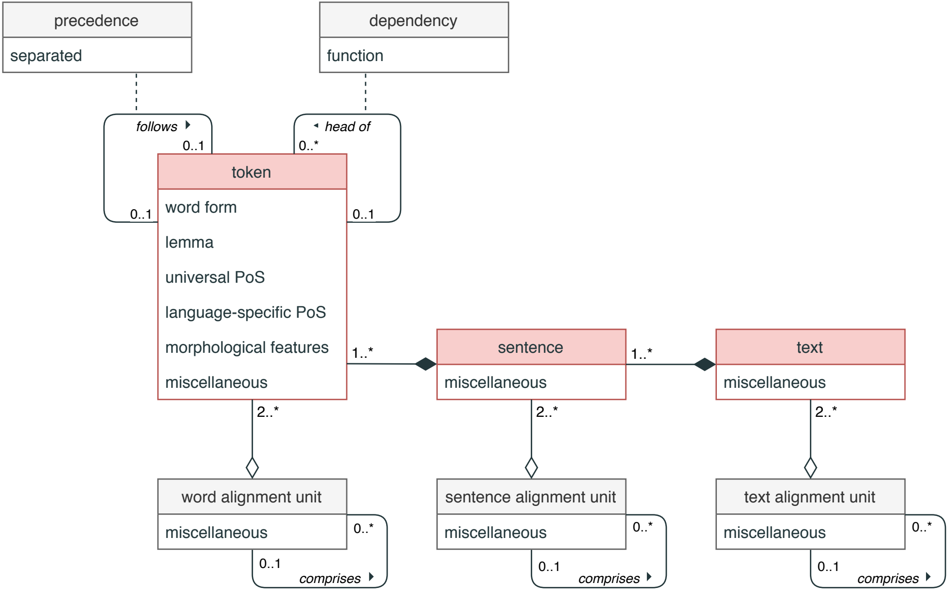
- 10 different corpora
- 7 different formats
  - tree-structure (XML, TEI)
  - tabular (.tsv, .csv)
- non-standardized
- heterogeneous metadata and annotations
- difficult to **combine**, **extract** & **exploit** the data at our finger tips

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



Questions
Comments
Discussion

Alignment

Definition

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

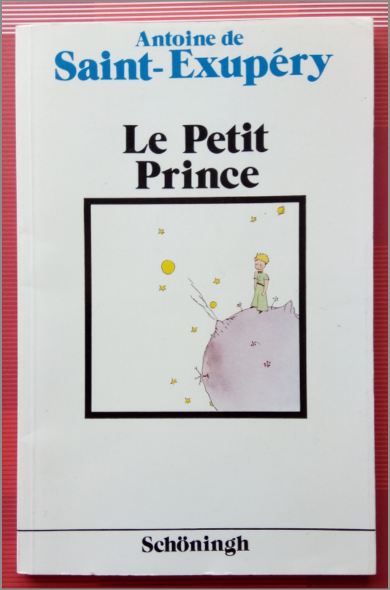
- also: text alignment, document linking
- correspondence in most cases given[3]
- documents might be ordered (e.g. plenary sessions with an indicated date) but typically are not (e.g. books)
- metadata varies from source to source
- typically 1:1 correspondences
- there might be null alignments[4] (depending on the document collection)
- translation direction can be indicated on this level[5]

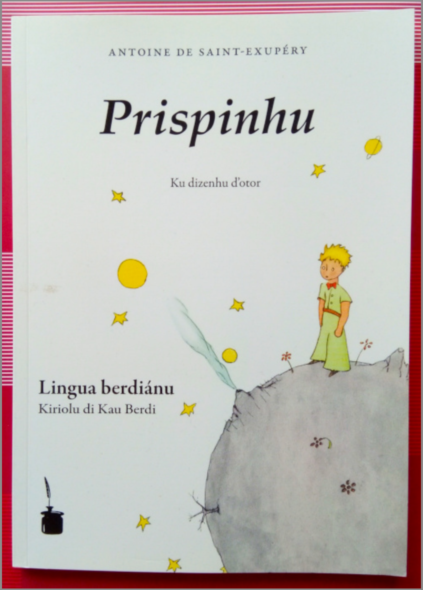
[3] (Tiedemann 2012)

[4] units that have no counterpart

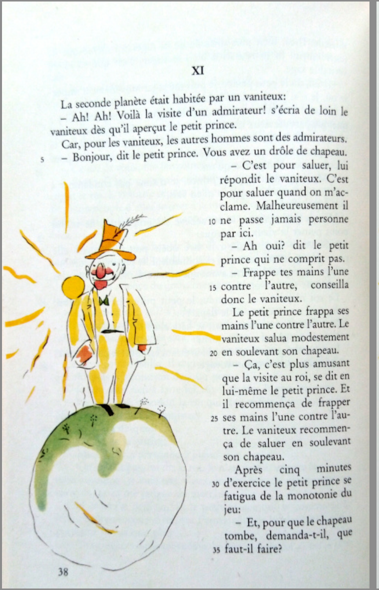
[5] but, for example, turn-based translation in Europarl

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.

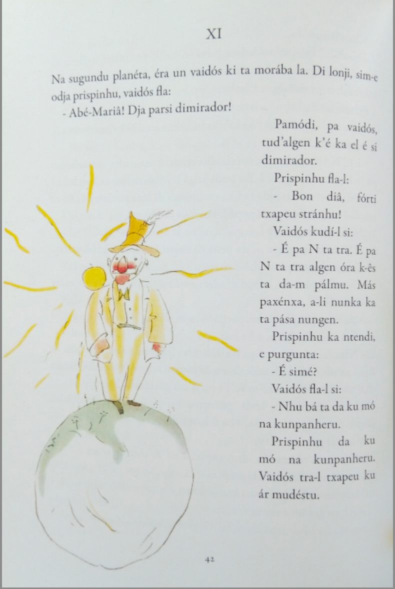
– Frappe tes mains l'une contre l'autre, conseilla donc le vaniteux.

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?



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:

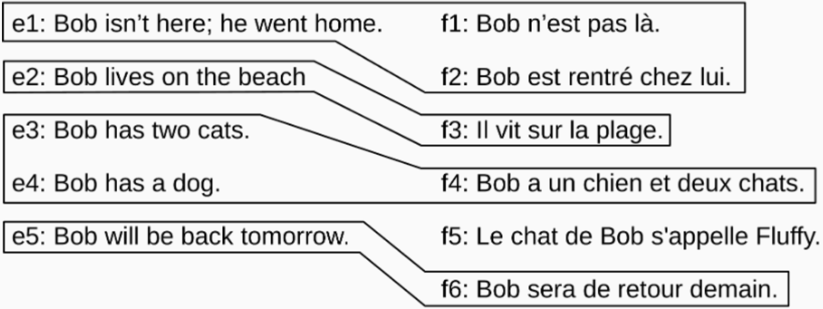
- Nhu bá ta da ku mó na kunpanheru.

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

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



e1: Bob isn’t here; he went home.

e2: Bob lives on the beach

e3: Bob has two cats.

e4: Bob has a dog.

e5: Bob will be back tomorrow.

f1: Bob n’est pas là.

f2: Bob est rentré chez lui.

f3: Il vit sur la plage.

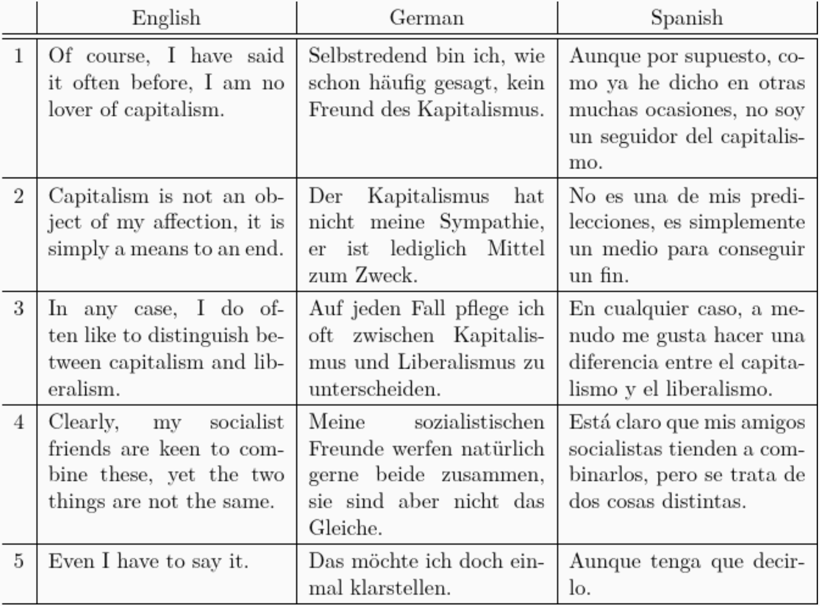
f4: Bob a un chien et deux chats.

f5: Le chat de Bob s'appelle Fluffy.

f6: Bob sera de retour demain.

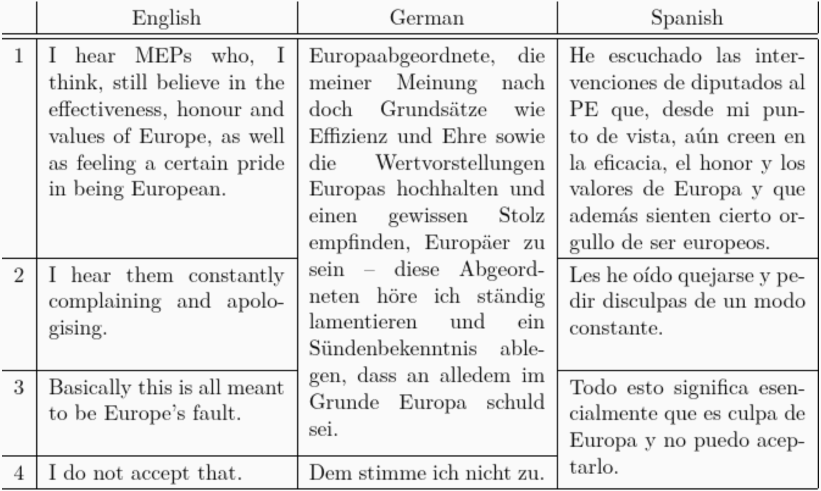
(Thompson and Koehn 2019)

1:1 alignments



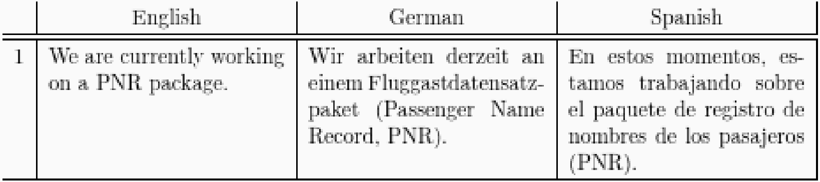
	English	German	Spanish
1	Of course, I have said it often before, I am no lover of capitalism.	Selbstredend bin ich, wie schon häufig gesagt, kein Freund des Kapitalismus.	Aunque por supuesto, como ya he dicho en otras muchas ocasiones, no soy un seguidor del capitalismo.
2	Capitalism is not an object 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.	

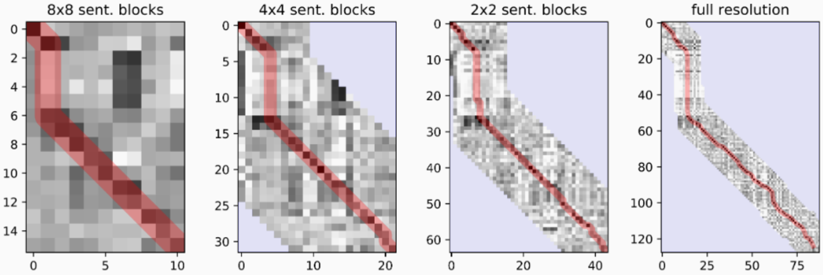
Added/omitted information



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

Sentence alignment – monotonicity

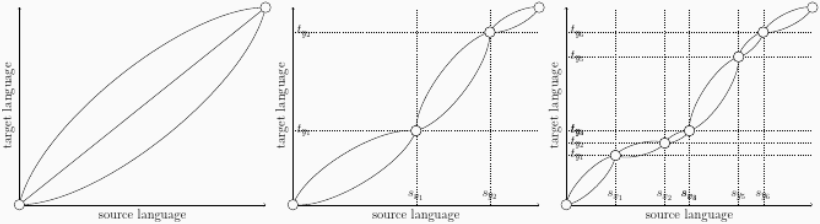
Under the assumption of monotonicity and infrequent null alignments, we can find the overall alignment around the diagonal:



(Thompson and Koehn 2019)

Sentence alignment – monotonicity

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

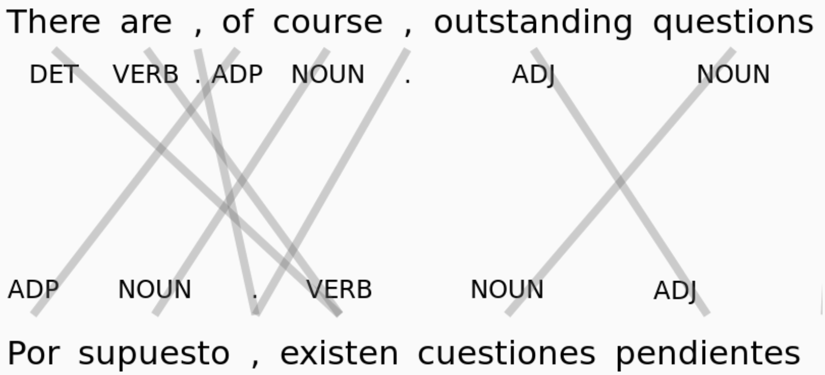


(Tiedemann 2012)

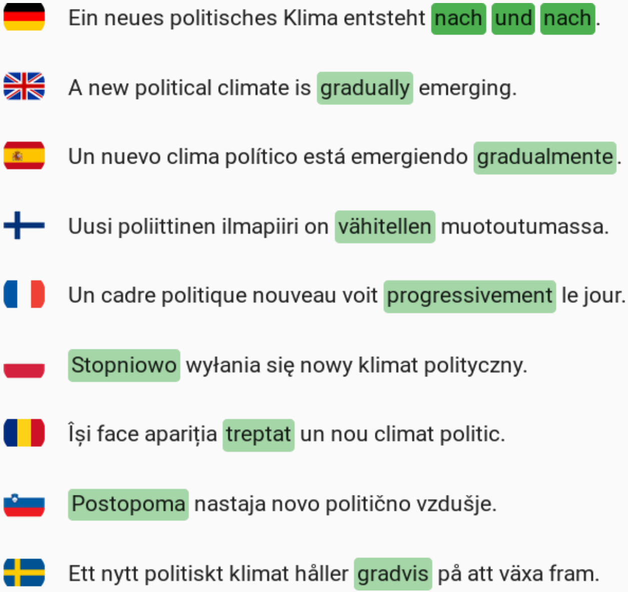
Word alignment

- 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)
- ⇒ word alignment comes with a significant error rate and only the aligned word may be of little help for particular applications

Alignment links


There are , of course , outstanding questions
DET VERB . ADP NOUN . ADJ NOUN
ADP NOUN . VERB NOUN ADJ
Por supuesto , existen cuestiones pendientes

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.

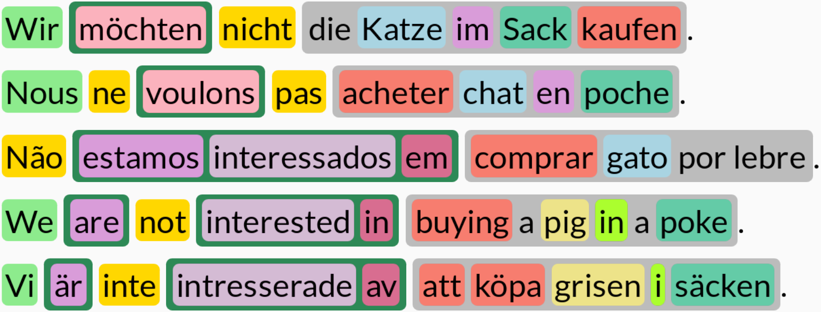
Postopoma nastaja novo politično vzdušje.

Ett nytt politiskt klimat håller gradvis på att växa fram.

Linguistically motivated word alignment

	English	German	Swedish
$\mathcal{G}_1$	Can	Können	
$\mathcal{G}_2$	we	wir	vi
$\mathcal{G}_3$	afford	uns, erlauben	
$\mathcal{G}_4$	to	zu	att
$\mathcal{G}_5$	risk	riskieren	riskera
$\mathcal{G}_6$	that, kind, of	diese	den
$\mathcal{G}_7$	relationship	Beziehung	förbindelsen
$\mathcal{G}_8$	Can, we, afford	Können, wir, uns, erlauben	Har, vi, råd
$\mathcal{G}_9$	to, risk	zu, riskieren	att, riskera
$\mathcal{G}_{10}$	that, kind, of, relationship	diese, Beziehung	den, förbindelsen

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)

Questions
Comments
Discussion

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[6]
- Bleualign (Sennrich and Volk 2010)
  based on machine translation
- Vecalign (Thompson and Koehn 2019)
  based on bilingual sentence embeddings

[6] many null alignments

Alignment tools – words

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

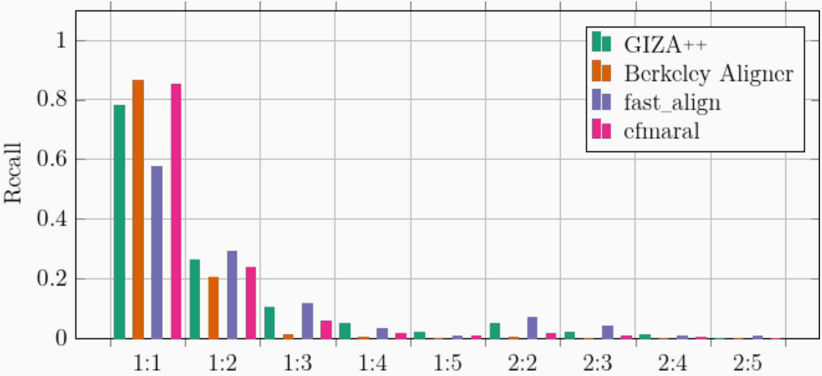
- GIZA++ (Och and Ney 2003)
  implements the ”IBM models”, asymmetric models
- BerkeleyAligner (Liang, Taskar, and Klein 2006)
  adds a probability threshold and a (symmetric) HMM
- fastalign (Dyer, Chahuneau, and Smith 2013)
  very fast, but less reliable
- efmaral/eflomal (Östling and Tiedemann 2016)
  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[7]
- 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)

[7] 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 of different aligners



Manual alignment (demo)

Our ”hierarchical alignment tool” for multiparallel corpora:

- Sentences
- Words
  ⇒ resulting multilingual alignment units

Questions
Comments
Discussion

Exploration & Visualization

Parallel concordancer

For the general public

- Linguee
- Glosbe
- TAUS (discontinued?)
- Tradooit

Special user groups

- PaGeS
- CLUVI

Multiparallel concordancer

Special user groups

- Multilingwis[8]
- ParCourE – Bible corpus explorer (typology)

[8] (Clematide, Graën, and Volk 2016; Graën, Sandoz, and Volk 2017)

Translation equivalents

- valid options to translate a word or a phrase
- can be derived from alignment statistics
- relative alignment frequencies (probabilities) are unidirectional[9]

absolute frequency		relative frequency	
$p_a(\textit{EN cow} \mid \textit{ES vaca})$	$= 305$	$f_a(\textit{EN cow} \mid \textit{ES vaca})$	$= 0.82$
$p_a(\textit{EN cattle} \mid \textit{ES vaca})$	$= 44$	$f_a(\textit{EN cattle} \mid \textit{ES vaca})$	$= 0.12$
$p_a(\textit{EN beef} \mid \textit{ES vaca})$	$= 4$	$f_a(\textit{EN beef} \mid \textit{ES vaca})$	$= 0.01$

[9] the most frequent translation of 'liberty' is 'libertad', but the most frequent translation of 'libertad' is 'freedom' (visual)

Alignment overlap – indirect support for translation equivalents

- we use a corpus with many languages; 12 of them are word-aligned to each other
- a translation equivalent can be supported by lemmas of third languages that are frequently aligned with both of the words in question
- the absence of such lemmas challenges the reliability of the original translation equivalent
- ⇒ we created an interface to explore those relations:[10] *`https://pub.cl.uzh.ch/purl/alignment_overlap`*

[10] (Graën and Schneider 2020)

Alignment overlap – use cases

- false friends (different languages)

  (es) entender & (fr) entendre

  (en) deception & (es) decepción

  (en) assist & (es) asistir

- quasi-synonyms (same language)

  (es) solicitar & (es) pedir & (es) rogar

  (de) steigen & (de) ansteigen

  (en) disturb & (en) bother & (en) annoy

- translation errors

  (en) July & (es) julio

Group work

Compare the different concordancers with regard to:

1. search options
2. query performance
3. presentation of results
4. correctness of data displayed
5. other helpful features

Use both single words and multiword expressions.

Linguistic Applications

Contrastive analysis: variable article use

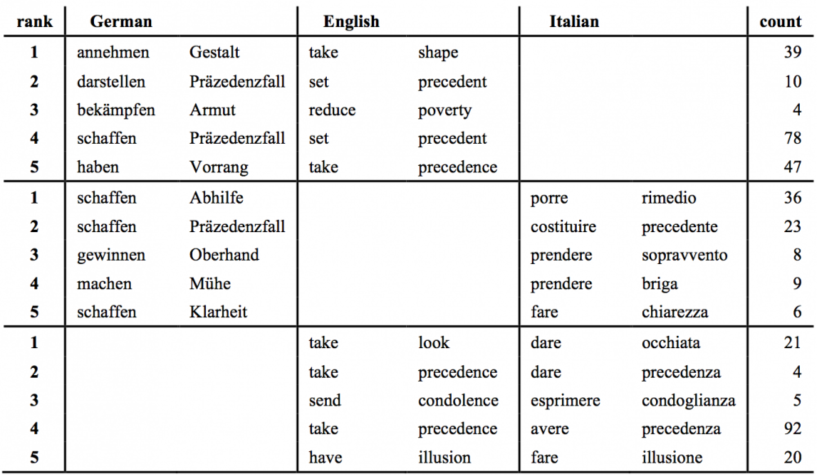
- *de* “In unseren einzelnen Mitgliedstaat und gemeinsam als **Europäische Union** müssen wir [...]”
- *en* “In our individual Member States, and collectively as **the** **European Union**, we must [...]”
- *es* “En nuestros respectivos Estados miembros y, de manera colectiva, en **la** **Unión Europea** 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 with gaps

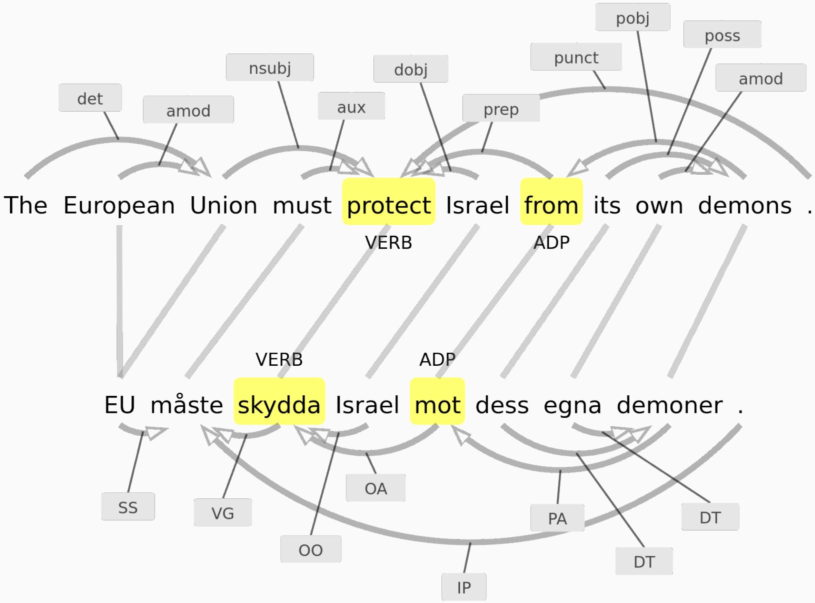
1	When does the Council intend to reach a decision on the establishment of this future observatory?
	När kommer rådet att fatta beslut om att inrätta detta framtida organ?
2	It has attempted to reallocate budgetary resources from the Progress programme to the microfinance facility before the European 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.

Primeros experimentos on support verb constructions

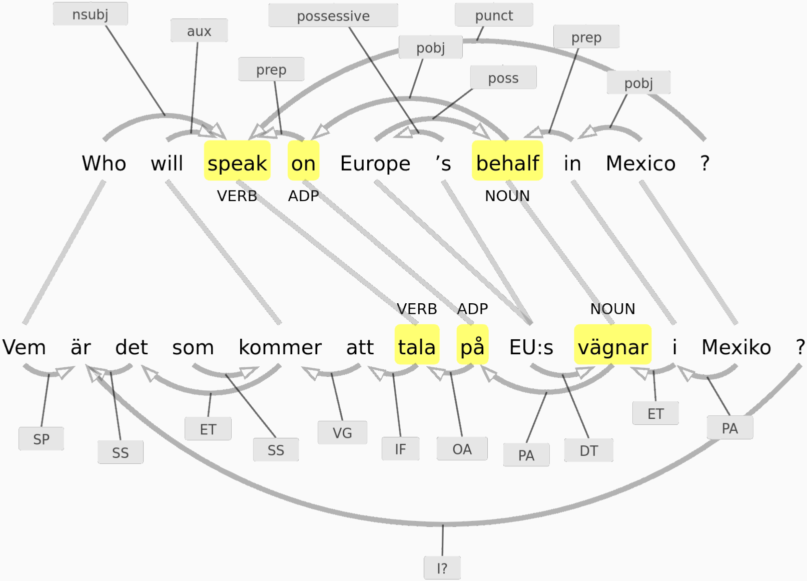


rank	German		English		Italian		count
1	annehmen	Gestalt	take	shape			39
2	darstellen	Präzedenzfall	set	precedent			10
3	bekämpfen	Armut	reduce	poverty			4
4	schaffen	Präzedenzfall	set	precedent			78
5	haben	Vorrang	take	precedence			47
1	schaffen	Abhilfe			porre	rimedio	36
2	schaffen	Präzedenzfall			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

Constelaciones


pobj
poss
punct
amod
nsubj
dobj
det
aux
prep
amod
The European Union must protect Israel from its own demons .
VERB
ADP
VERB
ADP
EU måste skydda Israel mot dess egna demoner .
OA
SS
VG
PA
DT
OO
DT
IP

Constelaciones



Who will speak on Europe ’s behalf in Mexico ?

VERB ADP NOUN

VERB ADP NOUN

Vem är det som kommer att tala på EU:s vägnar i Mexiko ?

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
- ⇒ high surprisal is an indicator for idiomaticity
- *https://pub.cl.uzh.ch/purl/constellations*

Language Learning

Project

From parallel corpora to multilingual exercises

Making use of large text collections and crowdsourcing techniques for innovative autonomous language learning applications

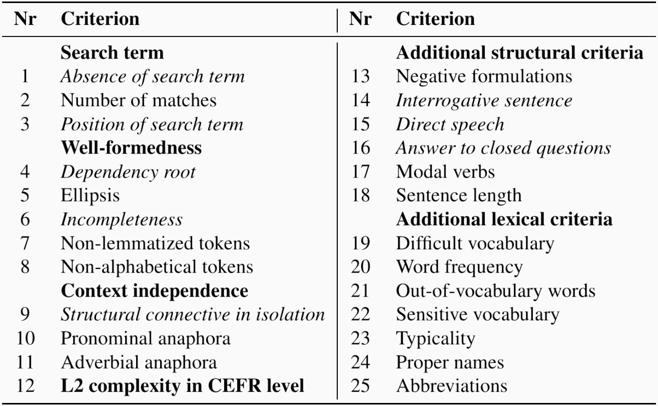
- Data-driven learning (DDL)
- Corpus-based learning is both effective and efficient
- Combination of NLP methods + CALL → ICALL applications

(Cobb and Boulton 2015)

Sketch

- Curate parallel corpora (cleaning, hierarchical alignment)
- Let teachers generate language learning exercises based on good examples
- Give (autonomous) learners access to the same functionality
- Use crowd (teachers & learners) for improvement

Sentence candidates for exercise generation (HitEx framework)



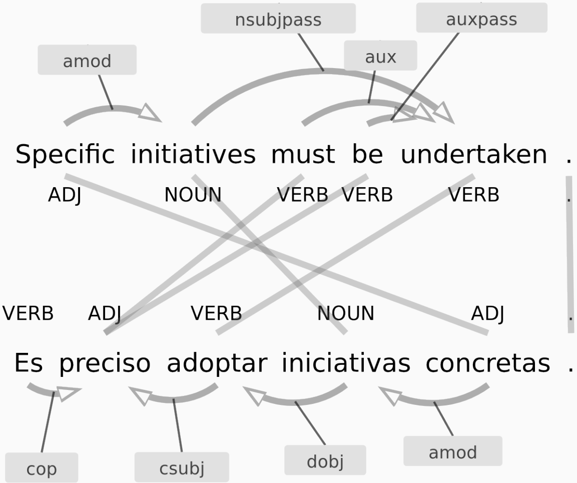
Nr	Criterion	Nr	Criterion
	Search term		**Additional structural criteria**
1	*Absence of search term*	13	Negative formulations
2	Number of matches	14	*Interrogative sentence*
3	*Position of search term*	15	*Direct speech*
	Well-formedness	16	*Answer to closed questions*
4	*Dependency root*	17	Modal verbs
5	Ellipsis	18	Sentence length
6	*Incompleteness*		**Additional lexical criteria**
7	Non-lemmatized tokens	19	Difficult vocabulary
8	Non-alphabetical tokens	20	Word frequency
	Context independence	21	Out-of-vocabulary words
9	*Structural connective in isolation*	22	Sensitive vocabulary
10	Pronominal anaphora	23	Typicality
11	Adverbial anaphora	24	Proper names
12	**L2 complexity in CEFR level**	25	Abbreviations

(Pilán, Volodina, and Borin 2016)

Candidates for parallel sentences

- Monolingual criteria plus translation characteristics
- Close (literal) or free (idiomatic) translation
- Correspondence on lexical level (single items vs. longer units)
- Similar part-of-speech and syntax or (systematic) difference between languages

Less literal translation ↔ more advanced correspondences


nsubjpass
auxpass
amod
aux
Specific initiatives must be undertaken .
ADJ
NOUN
VERB
VERB
VERB
.
VERB
ADJ
VERB
NOUN
ADJ
.
Es preciso adoptar iniciativas concretas .
cop
csubj
dobj
amod

Corpora in classroom situations

- Data-driven learning (DDL)
- also: discovery learning; discover & reconstruct
- intentional learning vs. incidential learning (context)
- direct vs. indirect corpus consultation

Language learning from corpora with moderated examples

- *Teacher* performs corpus search and compiles lists (indirect corpus consultation)
- Can deselect examples and/or mark them as defective
- These lists can be shared and serve as basis for automatic exercise generation
- Exercise links are passed to *learners*
- *Learners* can also provide feedback on single examples to teacher and system
- Results of automatic evaluation are shown to both
- Exercises with manual evaluation possible
- *Learners* can also use the application autonomously (direct corpus consultation)

Aspects of crowdsourcing

- ‘active’ vs. ‘passive’
- explicit vs. implicit
- motivation (paid, voluntary, “forced”)
- qualification (skills required, learners themselves)

PaCLE – use parallel corpora for generating language learning exercises

- tool available online: *https://demo.spraakbanken.gu.se/johannes/PaCLE/*

  example: (es) cuyo vs. (ca) el cual
- guidelines: *https://demo.spraakbanken.gu.se/johannes/PaCLE_demo/PaCLE-guidelines.pdf*

Group work

Look up the following Swedish expressions in PaCLE:

1. med tiden
2. kort och gott
3. över förväntan
4. öga mot öga
5. huvud på ett fat
6. med flit
7. sakta men säkert

- Can we derive their meaning?
- How straightforward is the translation compared to other languages?
- Is there more than one valid translation variant?
- Can we identify synonyms for the Swedish expressions?

Discussion

References

Braune, Fabienne and Alexander Fraser (2010). “Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora”. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING): Posters*. Association for Computational Linguistics (ACL), pp. 81–89.

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

- Clematide, Simon, Johannes Graën, and Martin Volk (2016). “Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora”. In: *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia computacional y basada en corpus: perspectivas monolingües y multilingües*. Ed. by Gloria Corpas Pastor. Geneva: Tradulex, pp. 447–455.
- Cobb, Tom and Alex Boulton (2015). “Classroom applications of corpus analysis”. In: *The Cambridge Handbook of English Corpus Linguistics*. Ed. by Douglas Biber and Randi Reppen. Cambridge University Press, pp. 478–497.
- Dou, Zi-Yi and Graham Neubig (2021). “Word Alignment by Fine-tuning Embeddings on Parallel Corpora”. In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith (2013). “A Simple, Fast, and Effective Reparameterization of IBM Model 2”. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics (ACL), pp. 644–649.

Graën, Johannes, Tannon Kew, Anastassia Shaitarova, and Martin Volk (July 2019). “Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection”. In: *Challenges in the Management of Large Corpora (CMLC-7)*. Ed. by Peter Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi. Leibniz-Institut für Deutsche Sprache.

- Graën, Johannes, Dominique Sandoz, and Martin Volk (2017). “Multilingwis2 – Explore Your Parallel Corpus”. In: *Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA)*. Linköping Electronic Conference Proceedings 131. Linköping University Electronic Press, Linköpings universitet, pp. 247–250.
- Graën, Johannes and Gerold Schneider (2020). “Exploiting Multiparallel Corpora as a Measure for Semantic Relatedness to Support Language Learners”. In: *Strategies and Analyses of Language and Communication in Multilingual and International Contexts*. Ed. by David Levey. Cambridge Scholars Publishing, pp. 153–167.
- Liang, Percy, Ben Taskar, and Dan Klein (2006). “Alignment by Agreement”. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics (ACL), pp. 104–111.

Och, Franz Josef and Hermann Ney (2003). “A Systematic Comparison of Various Statistical Alignment Models”. In: *Computational Linguistics* 29.1, pp. 19–51.

Östling, Robert and Jörg Tiedemann (2016). “Efficient word alignment with Markov Chain Monte Carlo”. In: *Prague Bulletin of Mathematical Linguistics* 106, pp. 125–146.

Pilán, Ildikó, Elena Volodina, and Lars Borin (2016). “Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation”. In: *Traitement Automatique des Langues* 57.3, pp. 67–91.

Sabet, Masoud Jalili, Philipp Dufter, François Yvon, and Hinrich Schütze (2020). “Simalign: High quality word alignments without parallel training data using static and contextualized embeddings”. In: *arXiv preprint arXiv:2004.08728*.

Sennrich, Rico and Martin Volk (2010). “MT-based sentence alignment for OCR-generated parallel texts”. In: *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*.

Thompson, Brian and Philipp Koehn (Nov. 2019). “Vecalign: Improved Sentence Alignment in Linear Time and Space”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1342–1348.

Tiedemann, Jörg (2009). “News from OPUS – A collection of multilingual parallel corpora with tools and interfaces”. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Vol. 5, pp. 237–248.

– (2012). “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)* (Istanbul).

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy (2005). “Parallel corpora for medium density languages”. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP)* (Borovets), pp. 590–596.