Parallel Corpus Examples for Language Learning Applications

Johannes Graën

Thursday 16th May, 2019

GÖTEBORGS UNIVERSITET

Universitat Pompeu Fabra Barcelona

Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Context

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 (Cobb and Boulton 2015)
- Combination of NLP methods + CALL $\rightarrow$ ICALL applications

Outline

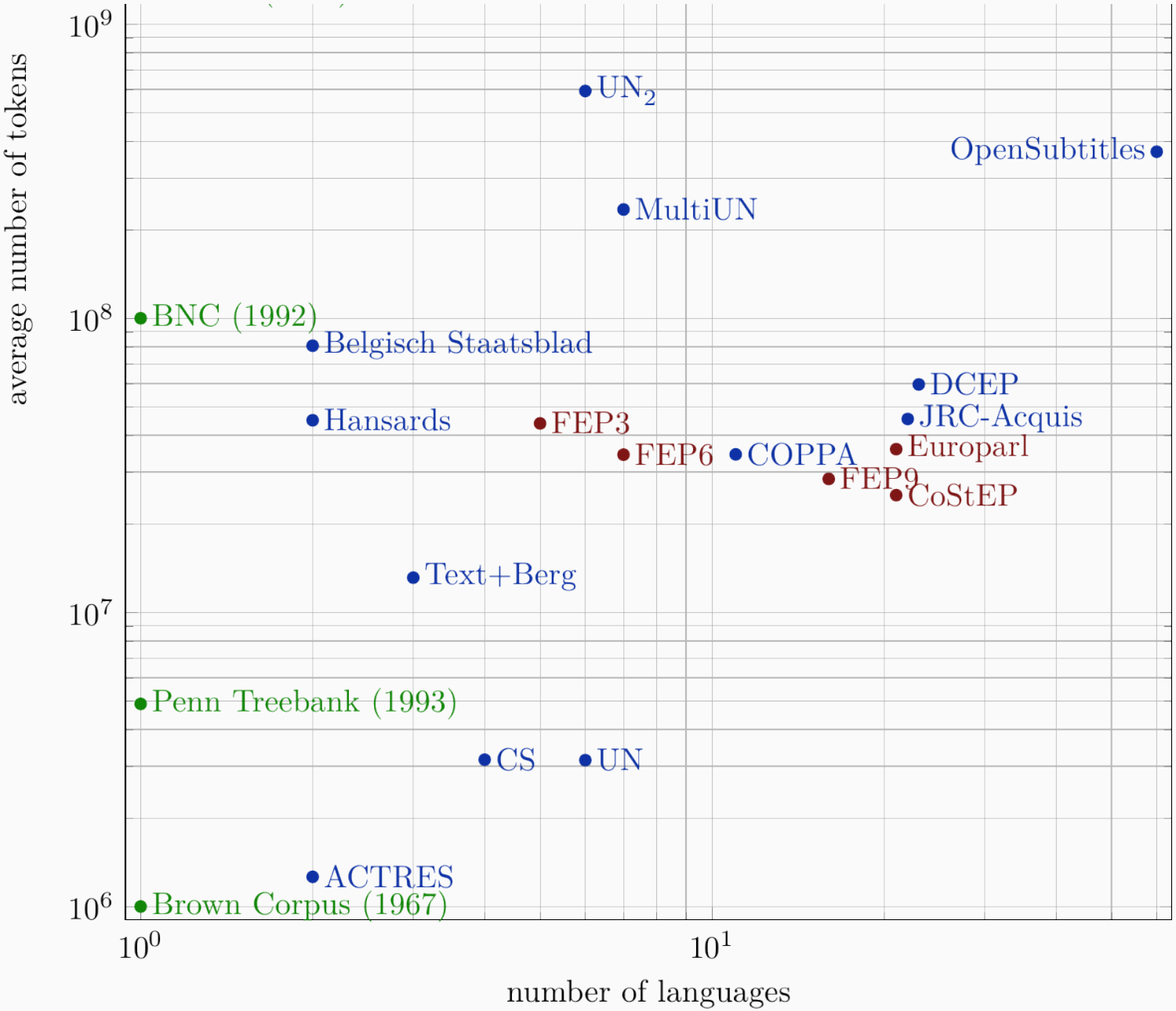
1. Parallel Corpora
2. Alignment
3. Applications Directed to Language Learners

Parallel Corpora

Parallel Corpora

- Collection of translated texts (or speech)
- Wikipedia $\neq$ parallel
- We have used Europarl for corpus linguistic studies

Parallel corpora



Comparing languages: 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 [...]"

$\Rightarrow$ variable article use, zero articles (Callegaro 2017)

Translation direction

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

Alignment

Alignment

- Correspondence relation between units in parallel corpora (documents, sentences, words/tokens, ...)
- Also: process of finding those correspondences
- Initial purpose: training statistical machine translation models
- Pure statistics $\rightarrow$ no inherent linguistic properties

Multiparallel word alignment: linguistic perspective

Simple word alignment


Die rumänische Gesellschaft ist erwachsen geworden.


Romanian society has grown up.


La sociedad rumana ha madurado.

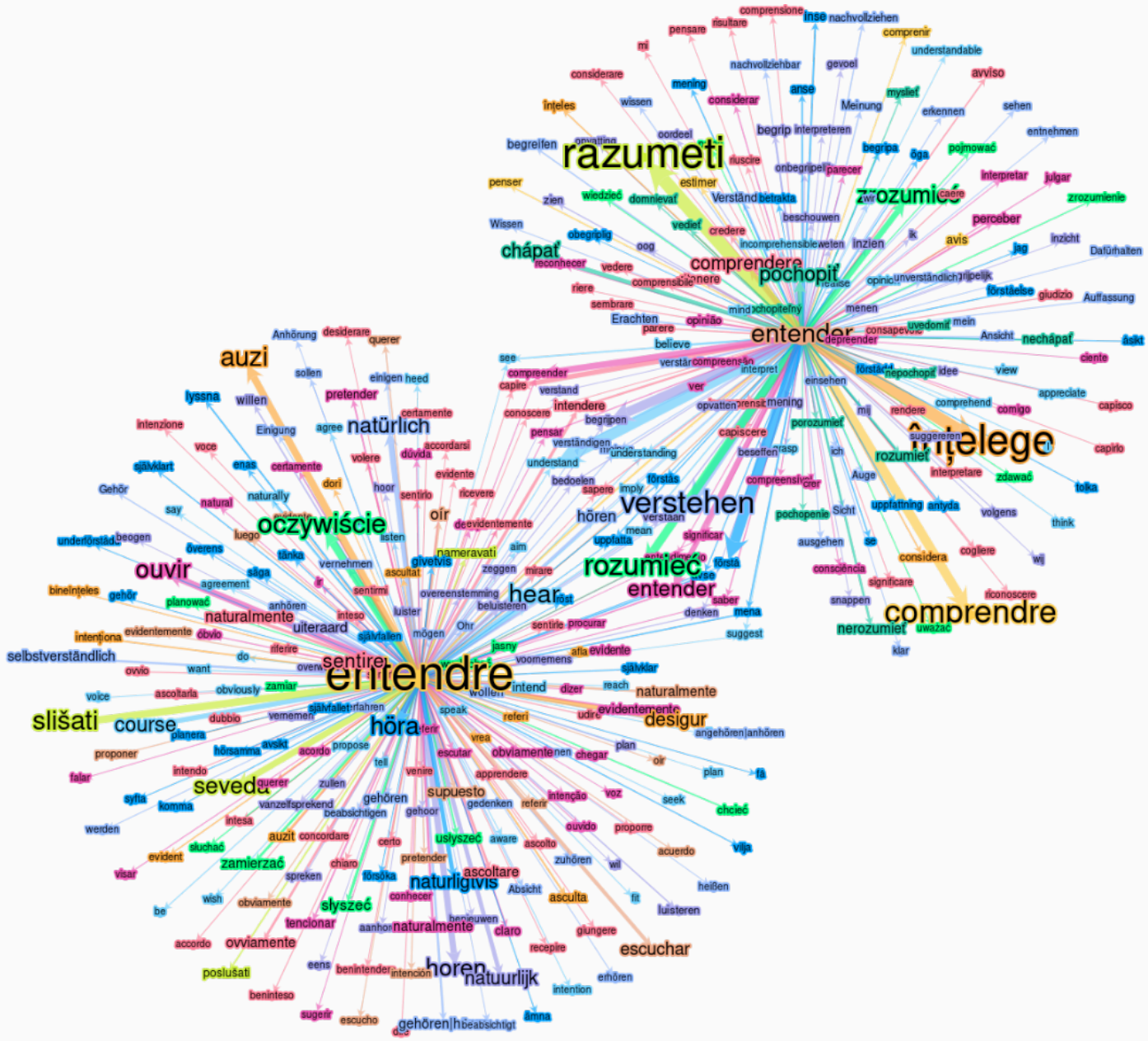

La société roumaine a mûri.


De Roemeense samenleving is volwassen geworden.

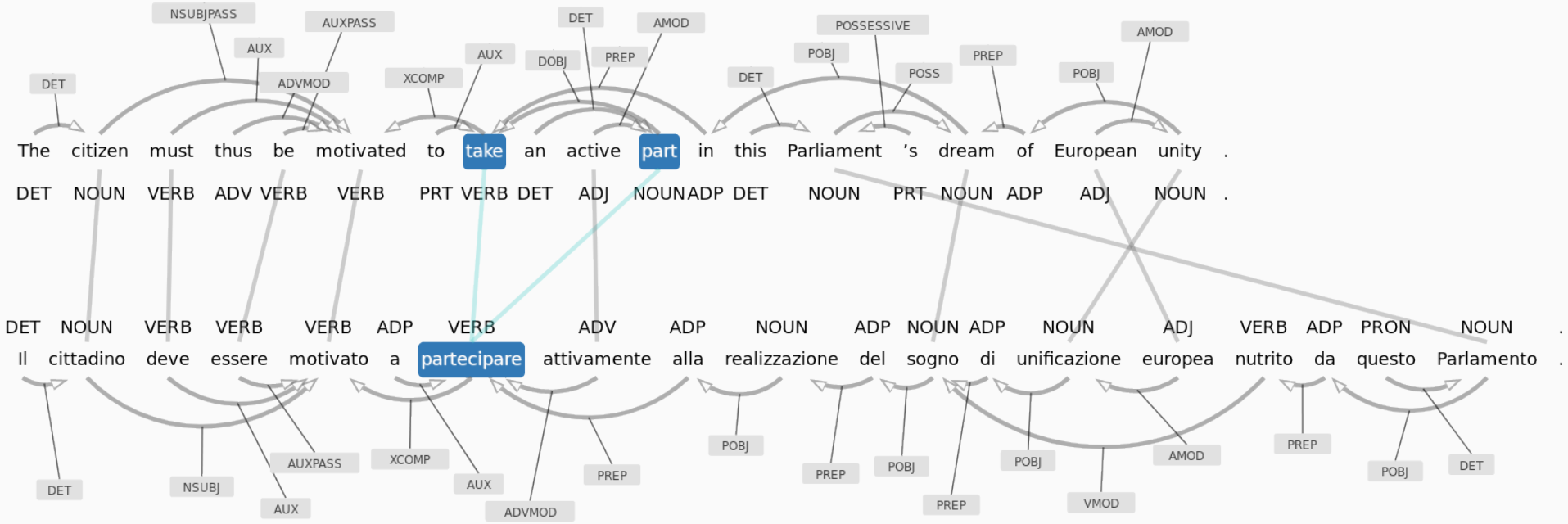

A sociedade romena amadureceu.


Det rumänska samhället har blivit vuxet.

Alignment distribution (overlap)



Approach: alignment combined with annotation



- Language-wise annotation: part-of-speech tagging and parsing
- Alignment as interlingual (correspondence) annoation

Applications Directed to Language Learners

Learner proficiency levels: CEFR

- CEFRLex project[1]: lists of words & expressions extracted from graded textbook corpora for different languages (Dürlich and François 2018; François et al. 2016; Gala et al. 2013)
- Single words vs. multi-word expressions (e.g. pull + back)
- Lexical vs. grammatical proficiency

[1] http://cental.uclouvain.be/cefrlex/

Good dictionary examples

- Following the idea of GDEX for lexica (Kilgarriff et al. 2008)
- Linguistic criteria (sentence length, vocabulary frequency, ...)
- Two dimensions: linguistic complexity and dependency on context

Sentence candidates for exercise generation

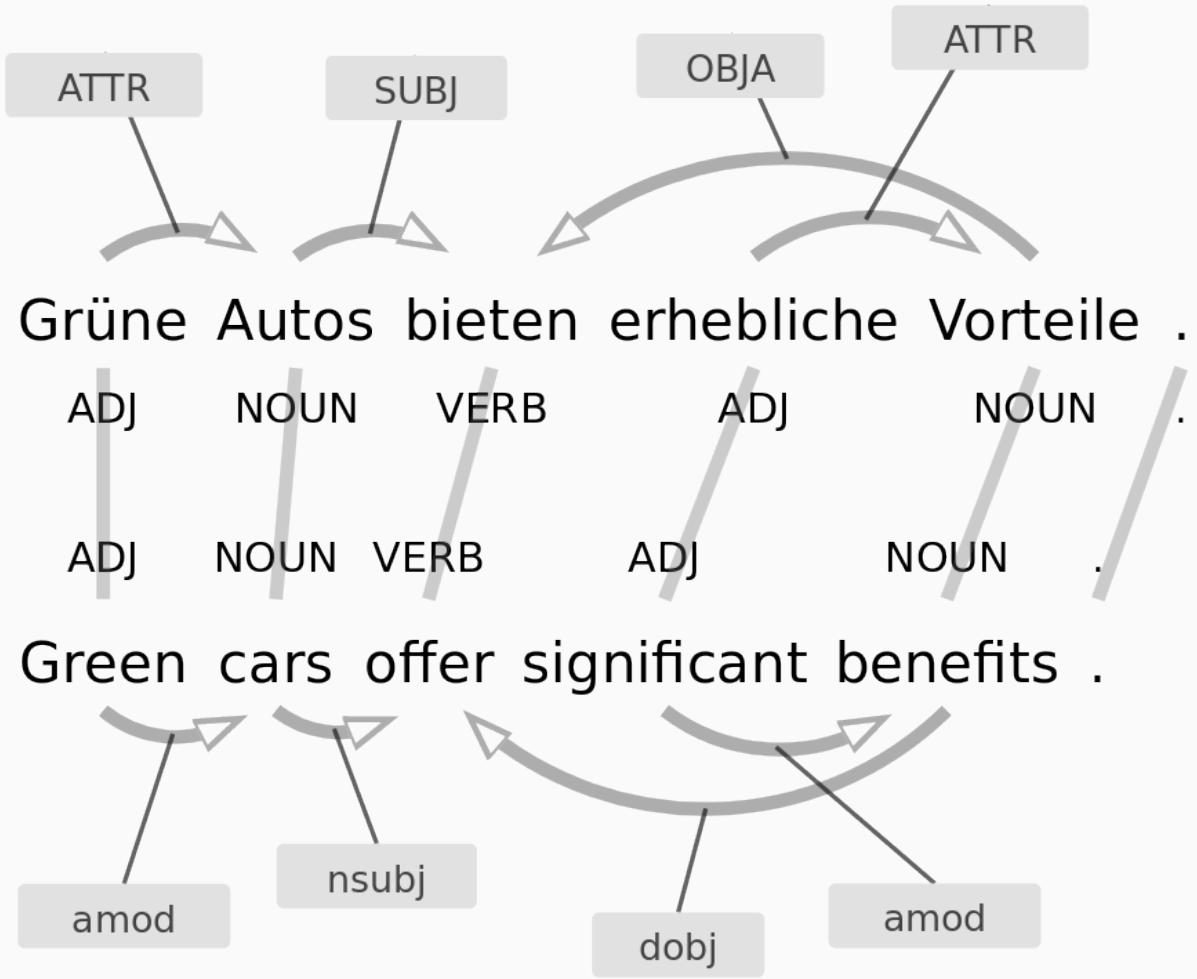
Nr	Criterion	Nr	Criterion
	Search term		**Additional structural criteria**
1	*Absence of search term*	13	Negative formulations
2	Number of matches	14	*Interrogative sentence*
3	*Position of search term*	15	*Direct speech*
	Well-formedness	16	*Answer to closed questions*
4	*Dependency root*	17	Modal verbs
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

Criteria for HitEx framework (Pilán et al. 2016)

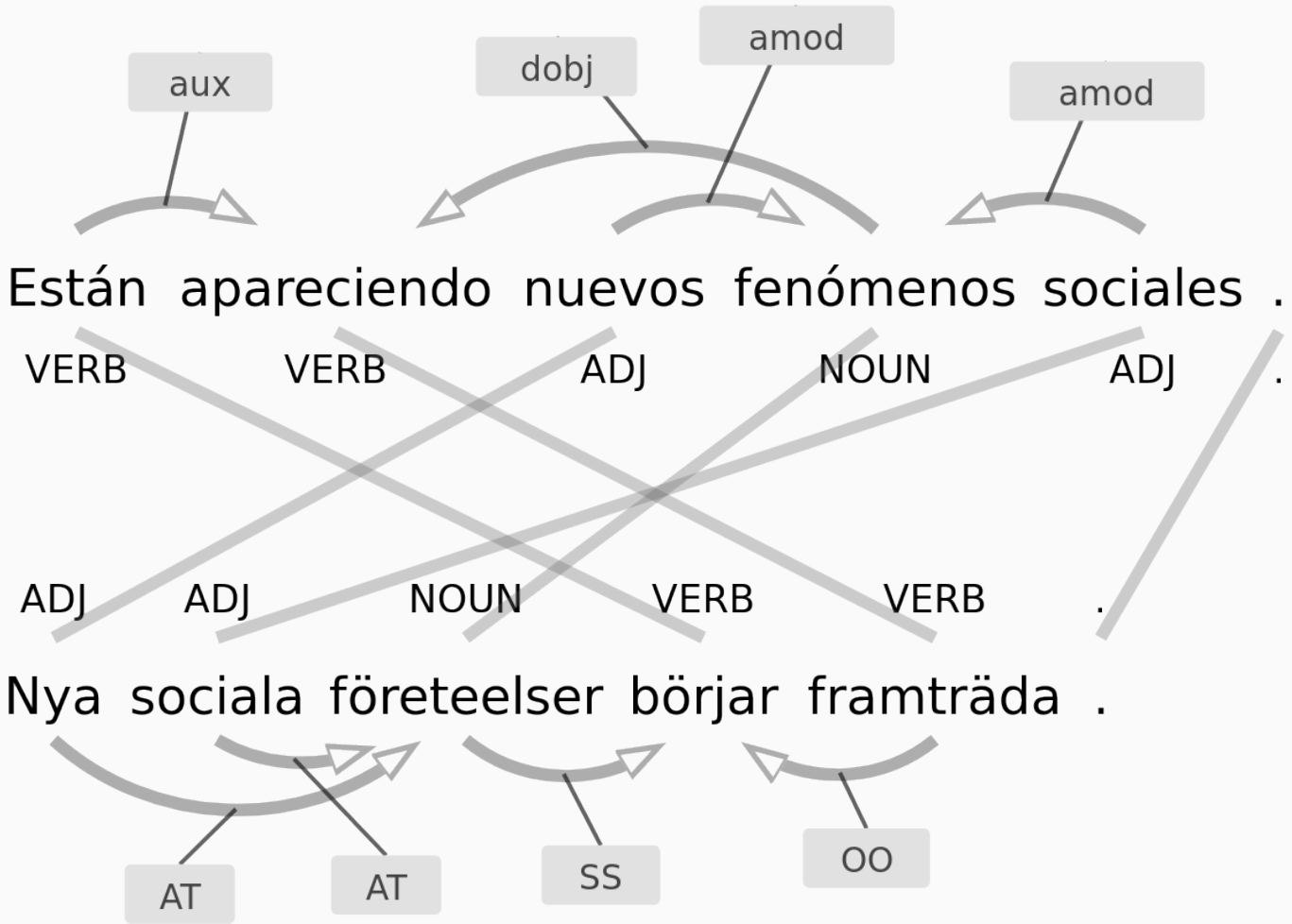
Candidates for parallel sentences

- 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

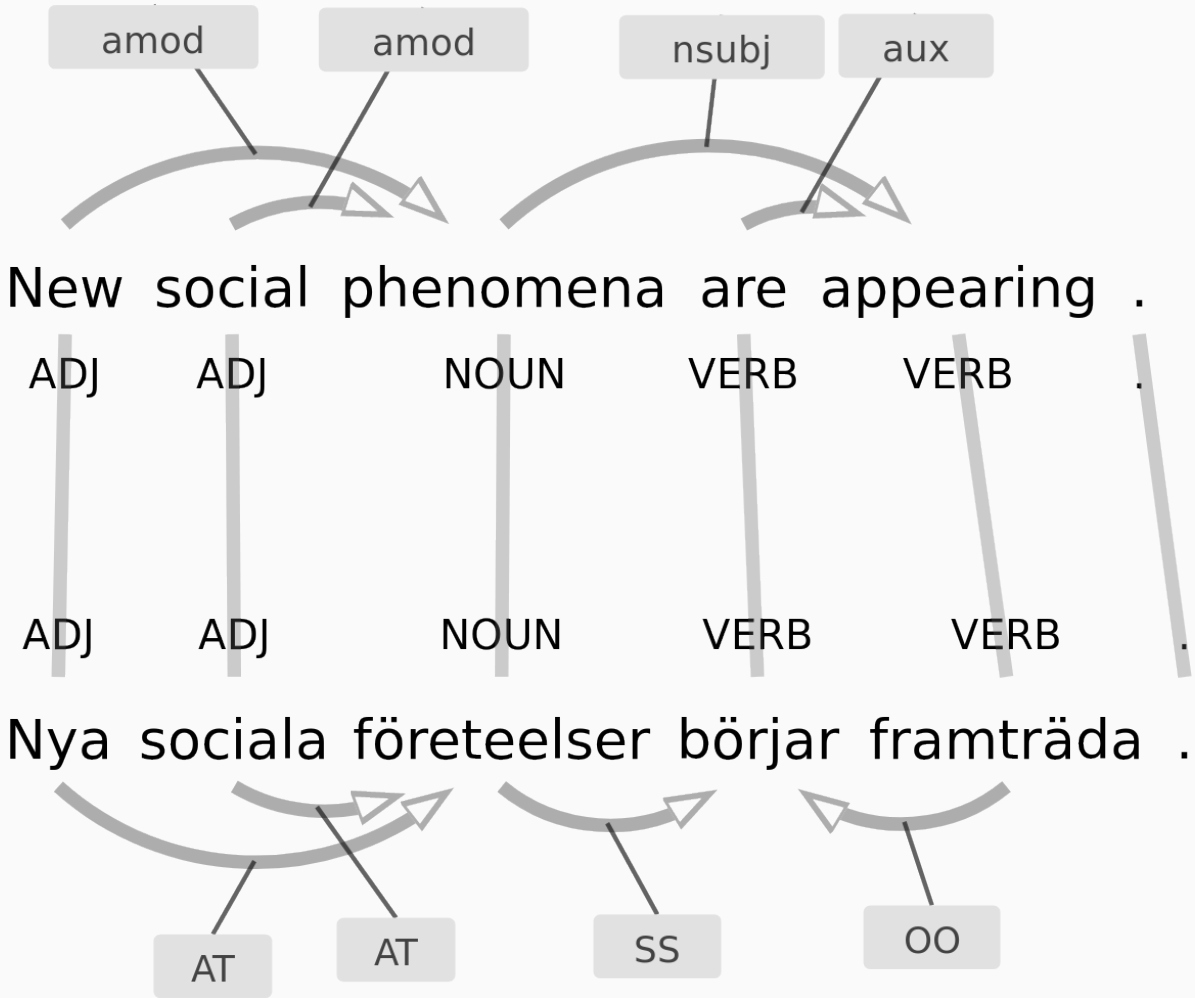
Straight correspondence $\leftrightarrow$ literal translation



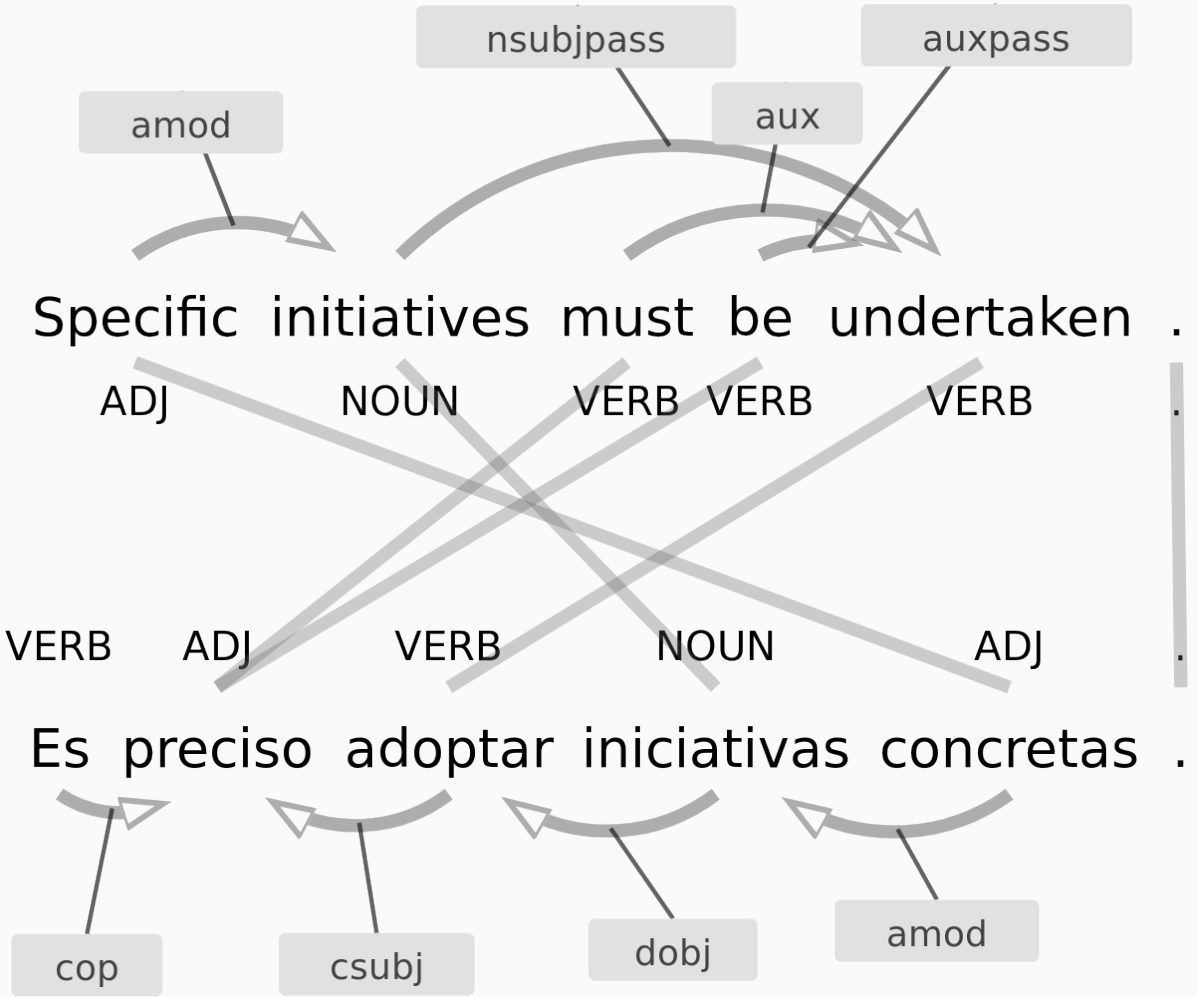
Diverging syntactical structure (es/sv)



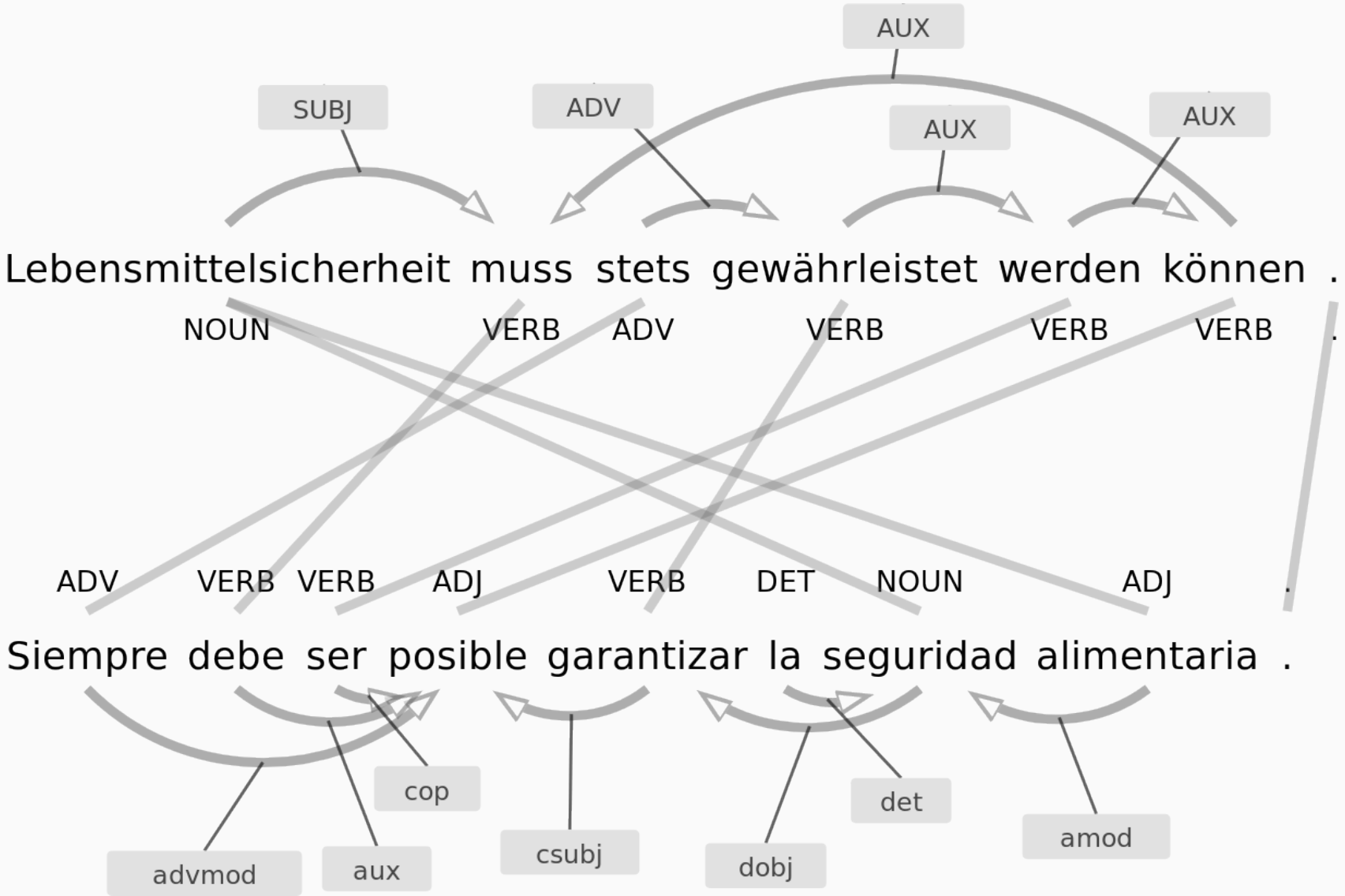
But similar structure in similar languages (en/sv)



Less literal translation ↔ more advanced correspondences



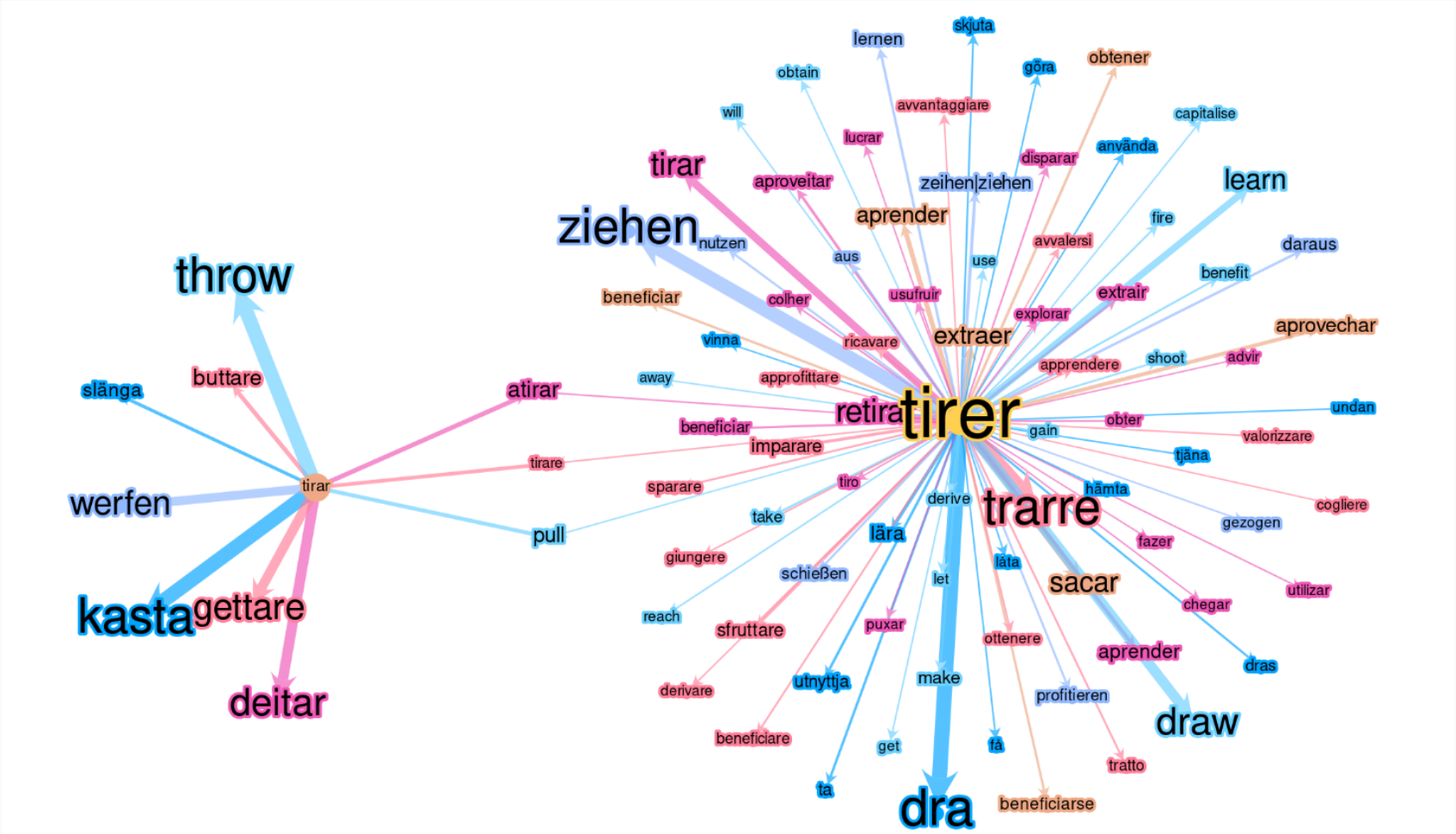
Complex structures and correspondences $\rightarrow$ high proficiency



Semantic Relatedness (False Friends)

En lugar de tener algún efecto beneficioso sobre mis constituyentes, la UE dará una bofetada a la gente de Gales si deciden **tirar** de la alfombra bajo nuestros pies.

Rather than having any sort of beneficial effect on my constituents, the EU will serve a glancing blow to the people of Wales if they decide to **pull** the rug from under our feet.



Bundle gap component

Show solution

Mellan juli och december förra året _____ bara 130 miljoner av de 700 miljoner som regeringen lagt undan .

Sen undrar jag om det inte var bola bola som _____ för köttbullar , ska försöka få tag i dottern hon vet .

" Alex , jag lyckades till slut spåra det där telefonnumret i Knightsbridge , det som _____ för restaurangens bordsbeställningar .

I veckor och månader efteråt _____ den här intervjun som sanningsvittne på att stortinget skulle få reducerad makt .

Continue

Vocabulary Multiple Choice

B1 ▾ Change level

Click to generate!

4	Among other things he ordered a big _____ of the philosopher Nietzsche.	guinea pig ▾ (guinea pig, cannabis, bacon, bottle, portrait)
3	Rör om och _____ väl .	
2	Den upplevelsen kommer jag att _____ när jag är död .	
1	Han pratade om att FN måste förändras för att _____ bättre .	kasta ▾

MultiPaVerX - Multilingual multiple choice particle verb exercise

Select language

I want to practice English particle verbs

Select mother tongue

My mother tongue is Spanish

I also speak ... Add language

Select proficiency in **English**:

Beginner Intermediate Advanced

Start

MultiPaVerX - Multilingual multiple choice particle verb exercise

+1 word | 50/50 | more context | Hints left: 2

Auseinanderbrechen

bryta ____

upp | ut | av | ihop | sönder | ned

MultiPaVerX - Multilingual multiple choice particle verb exercise

+1 word | 50/50 | more context | Hints left: 0

Auseinanderbrechen | brechen | zerstören

bryta ____

upp | ut | av | ihop | sönder | ned

References

- Callegaro, E. (2017). “Parallel Corpora for the Investigation of (Variable) Article Use in English: A Construction Grammar Approach”. PhD thesis. University of Zurich.
- Cobb, T. and A. Boulton (2015). “Classroom applications of corpus analysis”. In: *The Cambridge Handbook of English Corpus Linguistics*. Ed. by D. Biber and R. Reppen. Cambridge University Press, pp. 478–497.
- Dürlich, L. and T. François (2018). “EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language”. In: *11th International Conference on Language Resources and Evaluation (LREC)*.

References ii

- François, T., E. Volodina, I. Pilán, and A. Tack (2016). “SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners.”. In: *10th International Conference on Language Resources and Evaluation (LREC)*.
- Gala, N., T. François, and C. Fairon (2013). “Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons”. In: *E-lexicography in the 21st century: thinking outside the paper*.
- Kilgarriff, A., M. Husák, K. McAdam, M. Rundell, and P. Rychlỳ (2008). “GDEX: Automatically Finding Good Dictionary Examples in a Corpus”. In: *Proceedings of the 13th EURALEX International Congress*. Ed. by J. D. Elisenda Bernal. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

References iii

Pilán, I., E. Volodina, and L. 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.