Exploiting multiparallel corpora as measure for semantic relatedness to support language learners

Johannes Graën Gerold Schneider

Institute of Computational Linguistics
University of Zurich, Switzerland

20th April, 2018

Outline

Motivation

Corpus Material

Methods

Evaluations

Demo

Motivation

- Computational Linguistics and Learner Error research have made impressive progress recently, but they have not reached their collaborative potential yet (Granger and Lefer 2016)
- Verb-prefix constructions and phrasal verbs are difficult to acquire for language learners (Gilquin, Granger, et al. 2011) E.g. *DE schlagen* $\leftrightarrow$ *DE vorschlagen*
- False Friends are a frequent and difficult problem for language learners. Most resources are in the form of dictionaries (Varela 2011), which are open and incomplete
- On the other hand, not all occurrences of "false friends" are incorrect. E.g.: *ES firma* $\leftrightarrow$ *PT firma*
- There are now resources allowing us to compare real translations, e.g. Linguee, but
  - they require a considerable amount of reading, do not offer nice aggregations or visualizations
  - they do not specifically target language learners

Outline

Motivation

Corpus Material

Methods

Evaluations

Demo

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

[1] `http://pub.cl.uzh.ch/purl/costep`

Source

Europarl (version 7)

- Comprises transcript of the European Parliament sittings
- Contains numerous errors
- Has originally been compiled for training SMT systems
- Provides (reliable) alignment at the level of individual sittings

CoStEP (Corrected & Structured Europarl Corpus; Graën, Batinic, and Volk (2014))[1]

- Bases on the Europarl corpus
- Has undergone extensive cleaning
- Comprises ≈87 % of the original corpus material
- Provides alignment of speaker turns and additional speaker information (manually added)

[1] http://pub.cl.uzh.ch/purl/costep

Our Corpus

Version 9

- ≈150,000 speaker turns from **CoStEP** in 16 languages; altogether ≈450 million tokens
- **Tokenization** with our own multilingual tokenizer Cutter;[2] sentence segmentation based on tokenization tags
- Part-of-speech tagging and **lemmatization** with the TreeTagger and its featured language models
- Pairwise **sentence alignment** with hunalign and **word alignment** with four word aligners (Berkeley Aligner, GIZA++, fast_align and efmaral)
- For this application, we use only bidirectional alignments supported by all four aligners in 12 languages

[2] `http://pub.cl.uzh.ch/purl/cutter`

Outline

Motivation

Corpus Material

Methods

Evaluations

Demo

Lemma distribution matrix

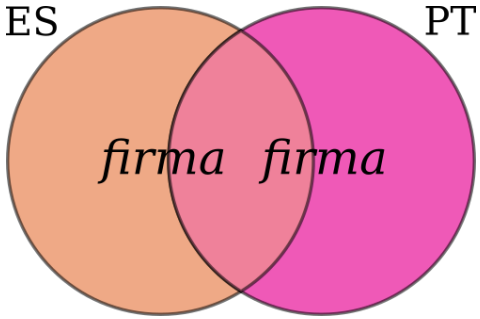
- Based on word alignment and lemmatization
- Reflects the probability of a lemma λ_s in the source language to be aligned with a lemma λ_t in the target language. E.g.:

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.

Alignment overlap

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



- We calculate frequencies for common lemmas:

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

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

- The overlap measure takes into account the absolute frequency:

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

Outline

Motivation

Corpus Material

Methods

Evaluations

Demo

Quantitative Evaluation

Based on a dictionary of "false friends" how well does our method detect them? How many prototypical translations does it incorrectly label as false friend?

- We use 2 online resources: `http://mentalfloss.com/article/57195/50-spanish-english-false-friend-words` and `https://en.wiktionary.org/wiki/Appendix:False_friends_between_English_and_Spanish`
- 64 items (cut a few trivial or contested cases)
- Overlap thresholds of 25% and 50% macro overlap

Examples and Precision

ES	EN false friend	EN Trans=good friend
actual	actual	current
asistir	assist	attend
campo	camp	countryside
compromiso	compromise	obligation
decepción	deception	disappointment
introducir	introduce	insert
éxito	exit	success
suceso	success	event
recordar	record	remember
vaso	vase	glass

Threshold	Prec(false friend)	Prec(good friend)
25%	88.9% (40/45)	70.0% (15/18)
50%	80.7% (46/57)	83.3% (21/30)

Outline

Motivation

Corpus Material

Methods

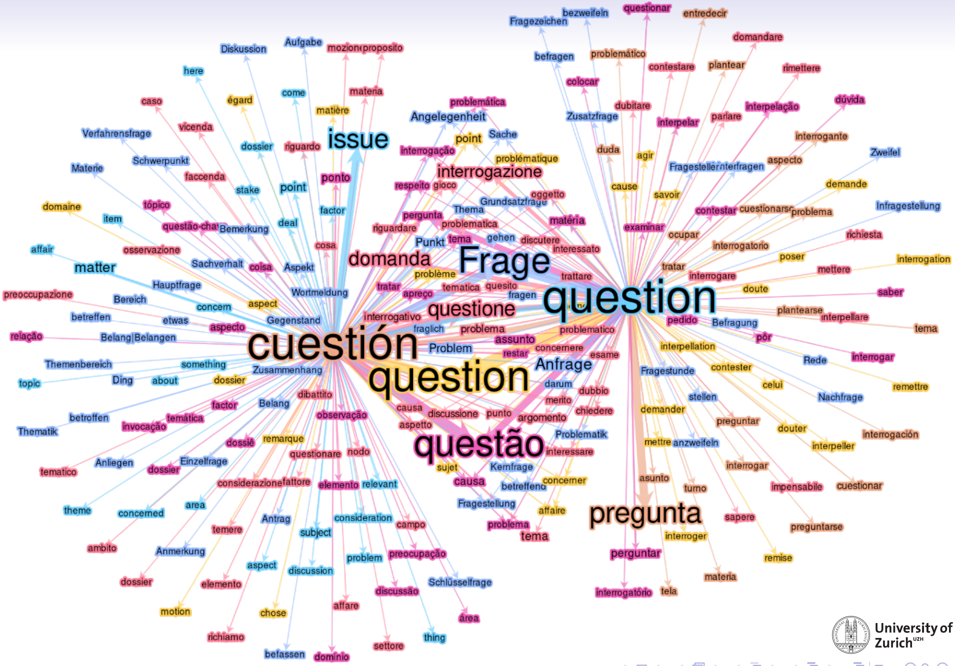
Evaluations

Demo

Demo[3]

1. *DE lösen* $\leftrightarrow$ *DE auslösen*
2. *DE steigen* $\leftrightarrow$ *DE ansteigen*
3. *ES tirar* $\leftrightarrow$ *FR tirer*
4. *ES entender* $\leftrightarrow$ *FR entendre*
5. *EN annoy* $\leftrightarrow$ *EN disturb* $\leftrightarrow$ *EN bother*

[3] `http://pub.cl.uzh.ch/purl/alignment_overlap`



University of Zurich[UZH]

References

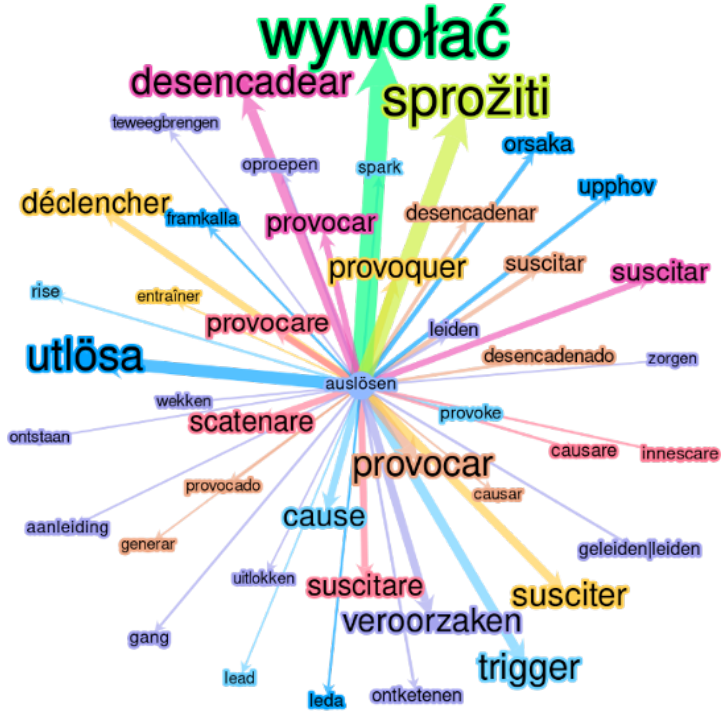

Johannes Graën, Dolores Batinic, and Martin Volk (2014). "Cleaning the Europarl Corpus for Linguistic Applications". In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*. (Hildesheim). Stiftung Universität Hildesheim, pp. 222–227


Sylviane Granger and Marie-Aude Lefer (2016). "From general to learners' bilingual dictionaries: Towards a more effective fulfilment of advanced learners' phraseological needs". In: *International Journal of Lexicography*, pp. 279–295

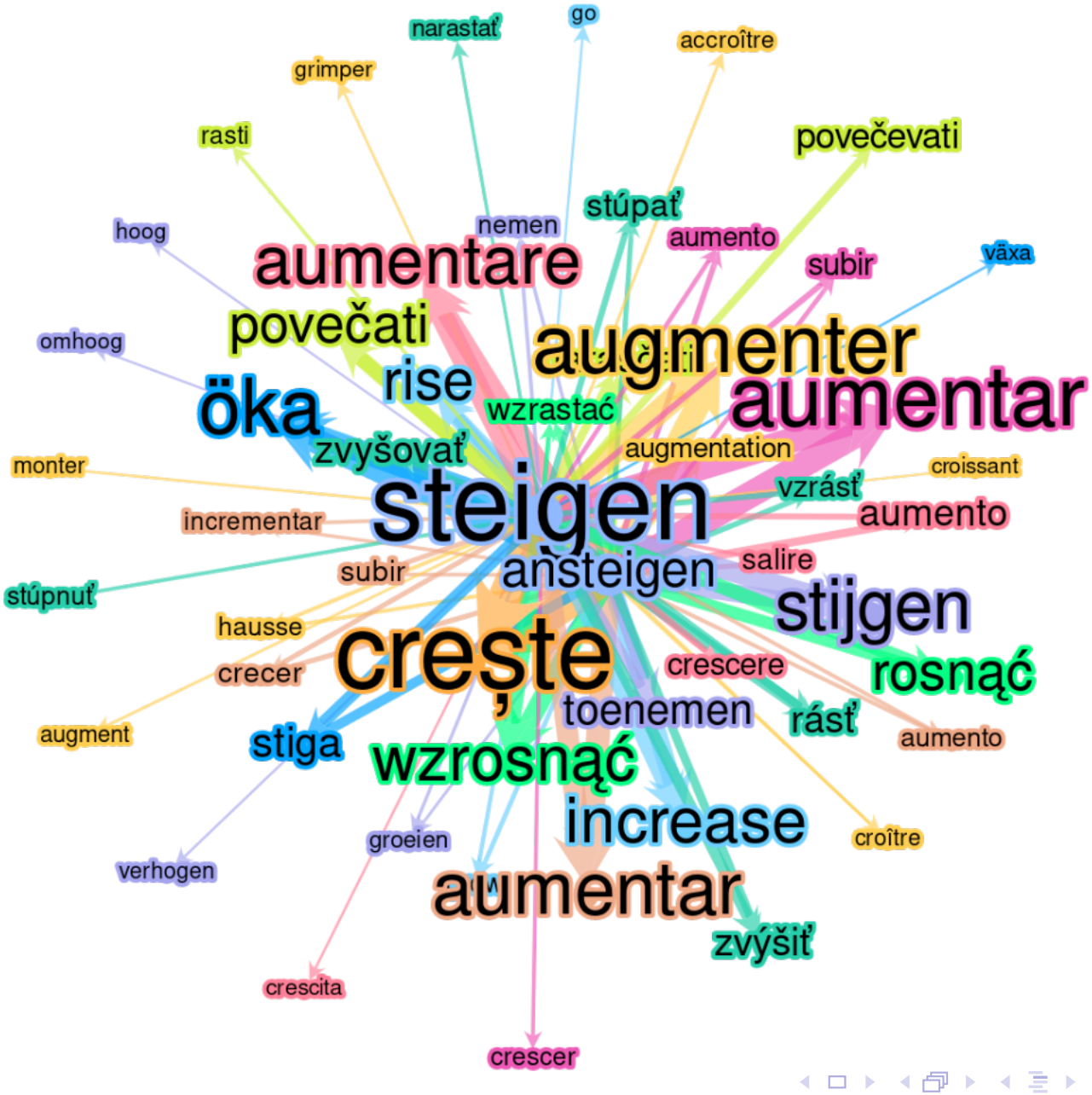

Gaëtanelle Gilquin, Sylviane Granger, et al. (2011). "From EFL to ESL: evidence from the International Corpus of Learner English". In: *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, pp. 55–78


Maria Luisa Roca Varela (2011). "Teaching and Learning "false friends": a review of some useful tools". In: *Encuentro* (20), pp. 80–87

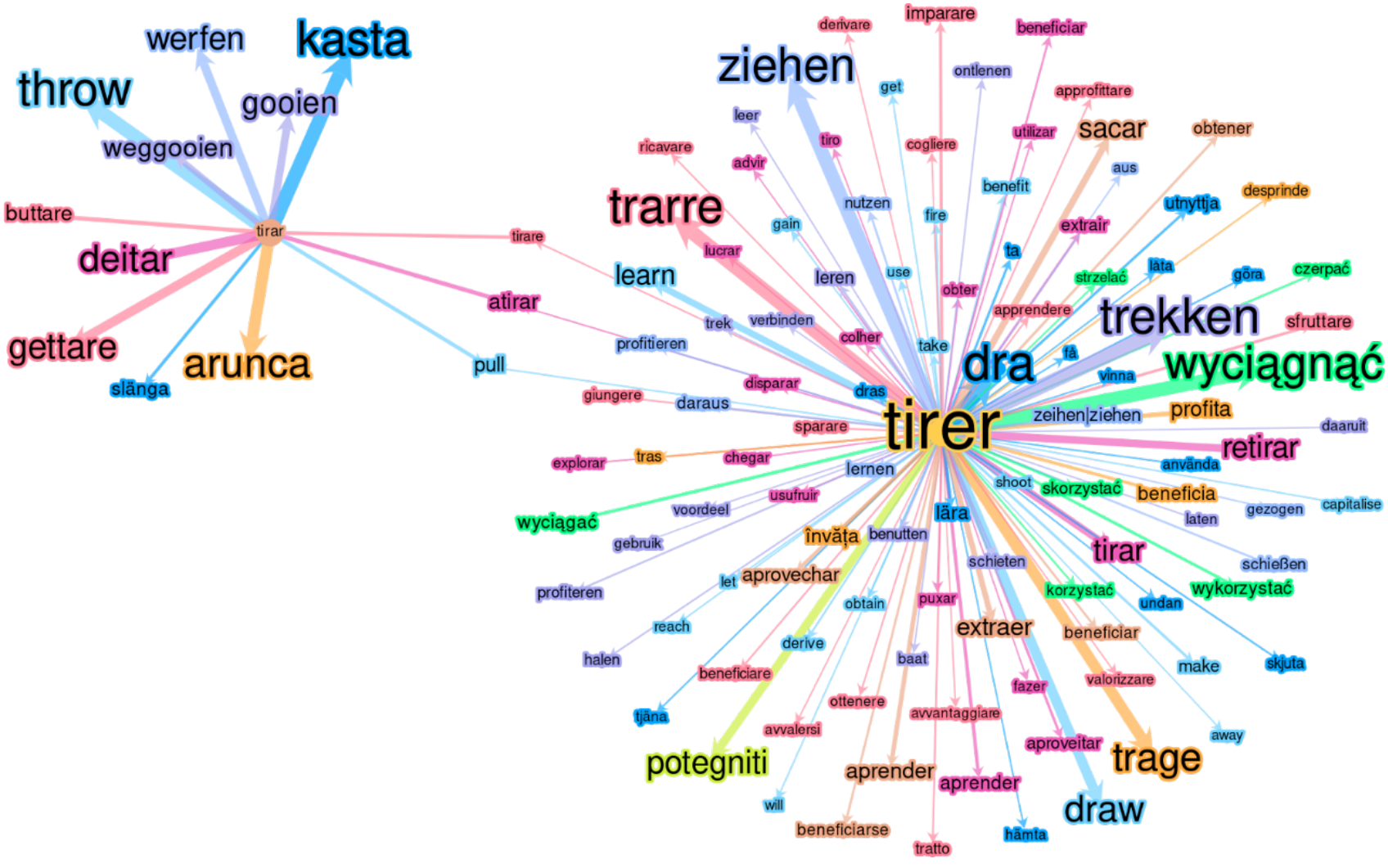
DE lösen ↔ DE auslösen



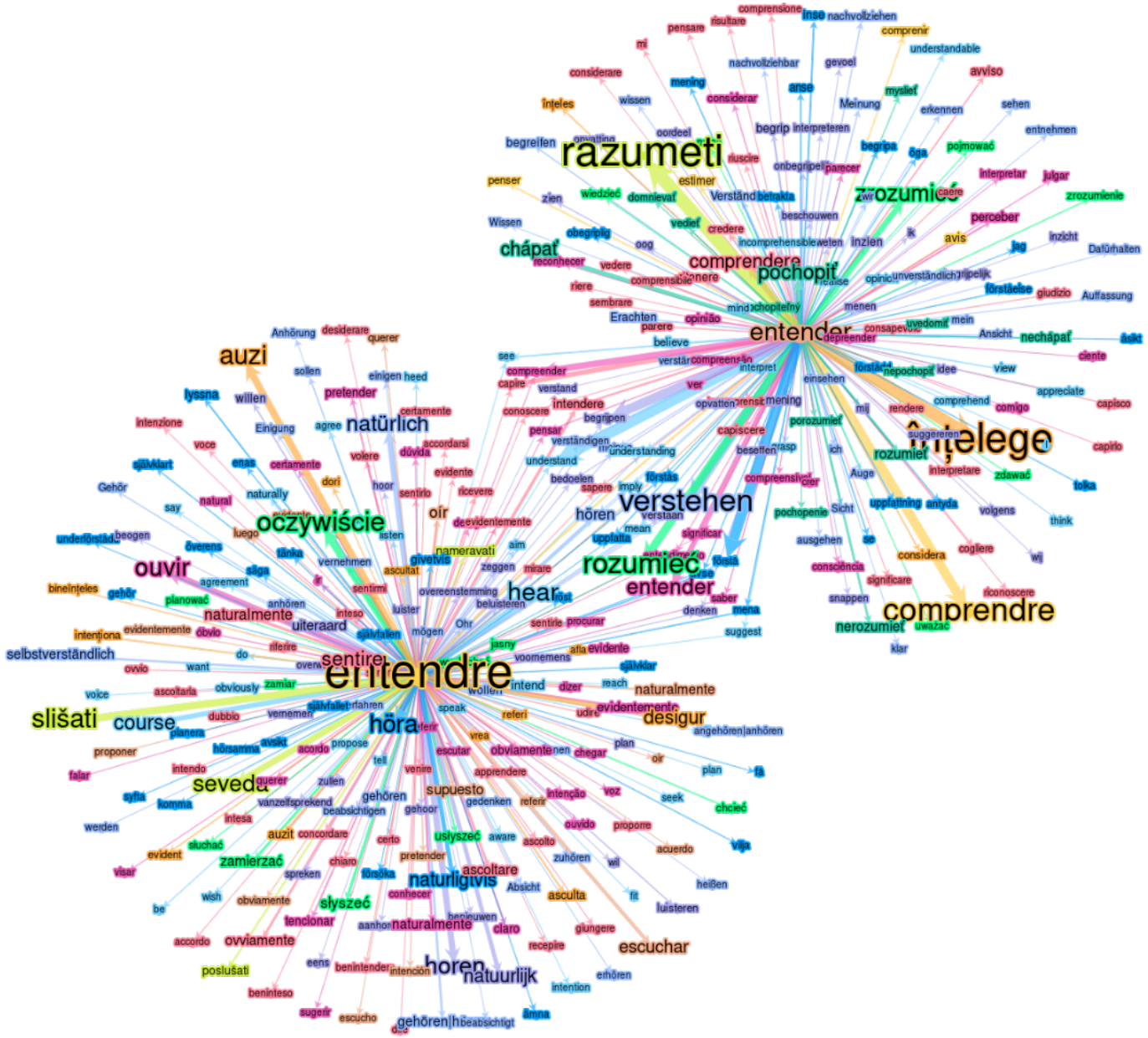
DE steigen ↔ DE ansteigen



ES tirar ↔ FR tirer



ES entender ↔ FR entendre



EN annoy ↔ EN disturb ↔ EN bother