Learning to Align across Languages:

Toward Multilingual FrameNet

Collin Baker and Luca Gilardi
collinb,lucag@icsi.berkeley.edu

International Computer Science Institute, Berkeley, California

International FrameNet Workshop 2018.05.12
at LREC, Miyazaki, Japan

Overview

Paths to FrameNets

Basic objectives and resources

- Practical vs. theoretical orientation
  - Berkeley FrameNet: Will Frame Semantics work in general? Will it help working lexicographers?
  - Spanish, German, Japanese FN, etc.: Will these frames work for ES, DE, JP? Are others needed?
  - Specific domains: Will FS work for sports texts, legal texts, environmental issues, dialog, argumentation, etc.? FN Brasil: Can we build an app for tourism?

Paths to FrameNets

Basic objectives and resources

- Practical vs. theoretical orientation
  - Berkeley FrameNet: Will Frame Semantics work in general? Will it help working lexicographers?
  - Spanish, German, Japanese FN, etc.: Will these frames work for ES, DE, JP? Are others needed?
  - Specific domains: Will FS work for sports texts, legal texts, environmental issues, dialog, argumentation, etc.? FN Brasil: Can we build an app for tourism?
- Lexicon vs. annotated corpus
  - Berkeley FN: build “Dictionary of the Future", annotate as documentation of usage; choose clear, simple examples
  - Later, began “full-text" annotation
  - Given a corpus, how well is it covered by FN? Annotate regardless of sentence complexity.

Paths to FrameNets

Existing Resources

- Existing corpora vs. DYI
  - BFN: British National Corpus, American National Corpus
  - SALSA: Tiger news corpus; no changes to parses
  - Japanese FN: Balanced Corpus of Contemporary Written Japanese (BCCWJ)
  - Spanish FN: built corpus from news, books, with a publisher, emphasizing New World Spanish
- Target language lexicons (national, free, commercial, bilingual)
- NLP tools: wide variation in target language tools: POS taggers, parsers, linked resources

Paths to FrameNets

Translation as part of the methodology

- **Korean FN:** (Hahm *et al.* 2014) First hired translators to translate ∼4,000 sentences from BFN, then fixed any errors. Later also translated from Japanese FN and projected annotation.
- **Hebrew FN:** (Hayoun & Elhadad 2016) Annotators chose English FN LUs in 200 frequent frames, found Hebrew translations (only LUs, not sentences).
- **Finnish FN:** (Lindén *et al.* 2017) The “annotated parts" of 80,000 BFN sentences were translated into Finnish in an early stage of development.

Coverage and Content

- The Team: ratio of Linguists:Computer Scientists
- General vs. domain-specific
- Frames and LUs vs. frames and word forms
- Spans vs. Parse nodes vs. head words
- Annotate FEs. . . and what else?
- Frame relations?

Coverage and Content

Coordination among FrameNets

- Distance from ICSI FN model
  - SALSA: "proto-frames"
  - French FN: merged frames (cf. Ruppenhofer *et al.* (2010)
  - FN Brasil: tourism frames
- Hard to update to new releases from ICSI just from the diff files
- How to feed back into Berkeley FN?
- Coordination between other FNs (including projection)

Coverage and Content

Availability of data produced

- Licensing and access methods
  - Licensing issues
  - Request to download/access website
  - Free to browse
  - Free to download
  - Available through web API
- Data types available
  - Frame names & descriptions
  - Lexical Units & definitions
  - Annotations w/o text
  - Annotated text
  - Frame & FE relations
  - Non-finalized or “problem" data

Linguistic Issues

Valences vs. constructions

- Define new frame or hypothesize construction?
- Parallel development of Construction Grammar
- Annotation of copulas and supports, even NPs, presupposes constructions
- Negatives and conditionals
- Sufficiency frame (*just enough water to prevent them from sticking*)
- Causatives in English and Japanese, among others $\Rightarrow$ reorganization of frames?

Linguistic Issues

Metaphor and Metonymy

- FrameNet policies on marking of metaphor:
  - fully lexicalized $\Rightarrow$ target frame (Attack.attack vs. Judgement_communcation)
  - productive or nonce $\Rightarrow$ source frames + label as metaphor
  - Metaphorical mappings were implicit (new F-F relation)
- Metonymy is rampant in FN, but is not annotated
- MetaNet project: frames $\neq$ FrameNet frames
- Metaphorical mappings are 1st class objects in MetaNet

Alignment across languages

Why align?

- Linguistic research
  - Comparative lexical semantics
  - Comparing valences/constructions across languages
- Implications for translation (human or machine)
  - Frames as an inter-language (Boas 2009)
  - Is translation "frame preserving"?
  - When and when not? Are frame shifts regular?
  - Can such knowledge help translation? (Čulo 2013)

Alignment across languages

Alignment by frames

- Q: Since so many projects use Berkeley FN frames, why not just line up the frames?
- A: Because of divergences from the Berkeley frames.
  - Is this the same frame? Same name? Translated name? Same ID?
  - Is this a "revised" frame? (changes in FEs, definitions)
  - Is this a new frame? (or in a later ICSI FN version?)
  - If new or revised, what is the relation to existing ICSI FN frames? Degree of relatedness? Type of relation?

Alignment across languages

Alignment by lexical units

- Frames can be aligned by measuring overlap of LUs across languages
- Sources for translation equivalents
  - Bilingual dictionaries, i.e. Multilingual WordNet, BabelNet, UBY
  - Derive from parallel corpora, preferably sentence-aligned
- Partial solution to polysemy–one sense per translation (cf. Carpuat (2009)).

Distributional approaches

Introduction

- Firth (1957) and Harris (1954)
- Count words in windows near target word
- Different spans: word2vec vs. GloVe
- Taking syntax into account: constituents or dependents
- Word forms vs. lemmas

Distributional approaches

Interpreting vectors

- Sparse binary vectors $d = 10,000+$
- Reduction to "standard" lengths $d = 40 \sim 300$
- Vector arithmetic, analogic reasoning
- Searching for meaning: QVEC (Tsvetkov *et al.* 2015) and SynEval (Köhn (2015),Köhn (2016))
- There are many ways to compose word vectors into vectors for phrases and sentences, including language models.

Trial Alignment

Multilingual distributional vector representations

- Our research on aligning is based on Hermann & Blunsom (2014) and Conneau *et al.* (2017)
- We use parallel corpora, through a supervised model but we could also have learned the vectors from monolingual corpora.
- We begin by formalizing the simple idea of translating LUs.
- Basically, we will say that a frame *X* is *aligned with* a frame *Y* when there are *enough* LUs associated with each that are good translations of each other.

Trial Alignment

Preparing the mappings

- For the first trial, we used frames and LUs from BFN 1.7 (the *source*) and Spanish FN (the *target*).
- We chose 200k word forms from the FN LUs for each language and retrieved vector representations for each of them from Facebook's `fastText.cc` project (Bojanowski *et al.* 2016)
- Learned a linear map between them, yielding a joint-space representation, (cf. Conneau *et al.* (2017).
- Removed source and target LUs for which no translation was found

Trial Alignment

Alignment Algorithm

- For each word form of each LU in each *source* frame f_S, retrieve from the vector model the N closest *target* translations
- For each of those translations, find the all the *target* frames $f_T^{(i)}$ in which they appear
- Each of those $f_T^{(i)}$ casts a "vote" for a mapping from f_S to $f_T^{(i)}$
- If there are fewer than k votes from a given LU to a given target frame, they are discarded. We have found by experimentation that with N= 10 and k = 5, we can avoid most incorrect mappings.

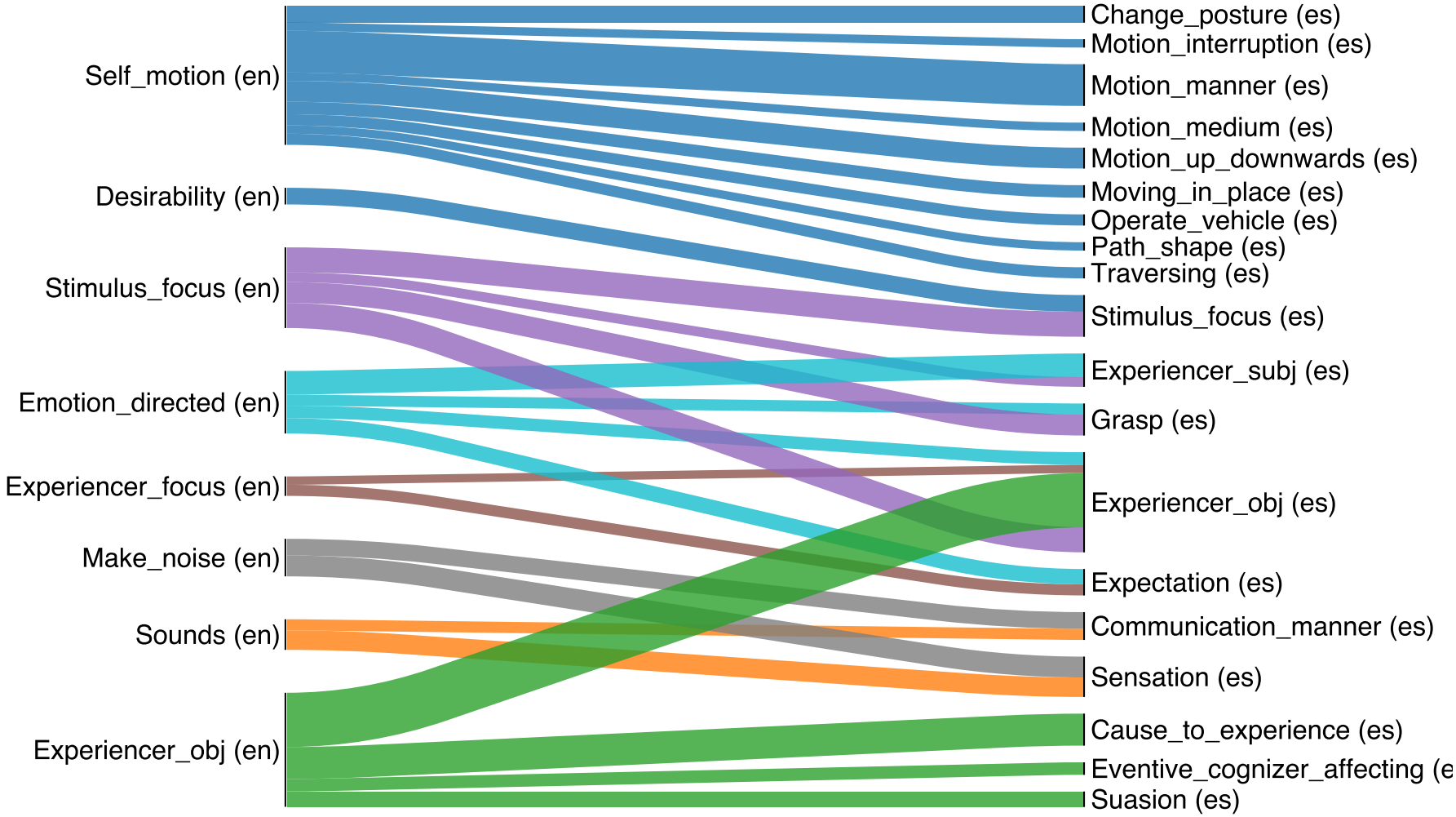

Figure: Top 20 Berkeley FN frames to Spanish FN frames

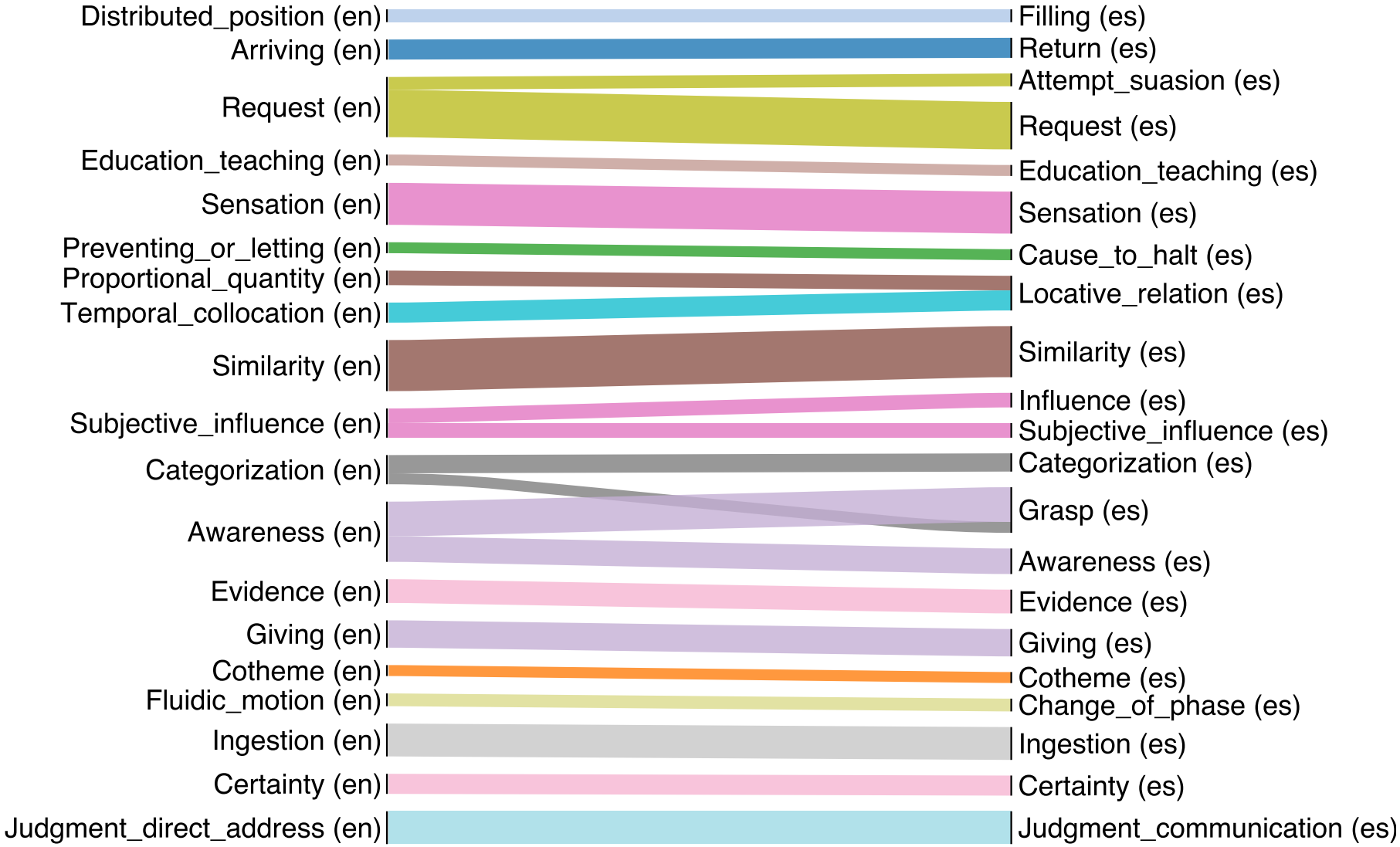

Figure: Top 80-150 Berkeley FN to Spanish FN frames

Discussion of Trial Alignment

A Few Comments

- Each of the previous slides ranks BFN frames by number of LUs.
- Many BFN LU's have no counterpart in Spanish FN (there are 13,631 LUs in BFN to 1,269 in SpFN and 1,073 BFN frames to 201 in the Spanish FN data).
- Moreover, we're comparing BFN 1.7 to Spanish FN which is derived from BFN 1.5; frames have been changed, and some LUs moved
- In the top 20 BFN frames seem to fan out more than the smaller frames, not surprisingly, given this algorithm

Discussion of Trial Alignment

Possible extensions

- In this preliminary work, we used only single-word LUs. Since multiwords constitute 12.8% of all LUs, we clearly need to find a way to handle them in the future.
- We will learn alignments from other languages to BFN; as most of the other FrameNets have fewer frames and LUs, those should be cleaner alignments
- Surprisingly, even with a dictionary of 200k word forms, vectors were not found for a some of the FN word forms; we will try again with a larger dictionary

Parallel Annotation Task

- We decided to create some parallel texts of our own, with Frame Semantic annotations, using A Web-based annotation tool developed at FN Brasil.
- Teams do the parallel annotation "independently", and evaluate in the light of the alignment.
- We are annotating the TED talk “Do Schools Kill Creativity?" using Release 1.7 frames: the task is complete for Portuguese, almost for English.
- This has revealed some missing frames: interactional, conversational frames, and deixis.
- More TED talks and other documents are planned.

Recent developments and Future Research

- Working on a better display of valences (i.e. views on the FN database)
- Use of Automatic Semantic Role Labeling!
- Crowdsourcing of frame discrimination (and FE annotation?) (Dumitrache *et al.* (forthcoming),Chang *et al.* (2015))
- Learning about ambiguity by crowdsourcing (cf. Erk *et al.* (2013))
- Frame relation induction (e.g. Virk *et al.* (2016), Botschen *et al.* (2017))

Acknoledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1629989, which we gratefully acknowledge. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

Thank you for your attention!

`https://framenet.icsi.berkeley.edu`

BOAS, HANS C.
2009.
Semantic frames as interlingual representations for multilingual lexical databases.
In *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, ed. by Hans C. Boas, p. 59–100.

BOJANOWSKI, PIOTR, EDOUARD GRAVE, ARMAND JOULIN, & TOMAS MIKOLOV.
2016.
Enriching word vectors with subword information.
arXiv preprint arXiv:1607.04606 .

BOTSCHEN, TERESA, HATEM MOUSSELLY-SERGIEH, & IRYNA GUREVYCH.
2017.
Prediction of frame-to-frame relations in the framenet hierarchy with frame embeddings.

In *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP)*, 146–156.

CARPUAT, MARINE.
2009.
One translation per discourse.
In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 19–27. Association for Computational Linguistics.

CHANG, NANCY, PRAVEEN PARITOSH, DAVID HUYNH, & COLLIN BAKER.
2015.
Scaling Semantic Frame Annotation.
In *Proceedings of The 9th Linguistic Annotation Workshop*, 1–10, Denver, Colorado, USA. Association for Computational Linguistics.

CONNEAU, ALEXIS, GUILLAUME LAMPLE, MARC'AURELIO RANZATO, LUDOVIC DENOYER, & HERVÉ JÉGOU.

2017.
Word translation without parallel data.
arXiv preprint arXiv:1710.04087 .

DUMITRACHE, ANCA, LORA AROYO, & CHRIS WELTY, forthcoming.
Capturing ambiguity in crowdsourcing frame disambiguation.

ERK, KATRIN, DIANA MCCARTHY, & NICHOLAS GAYLORD.
2013.
Measuring Word Meaning in Context.
Computational Linguistics .

FIRTH, J. R.
1957.
Papers in Linguistics 1934–1951.
London: Oxford University Press.

HAHM, YOUNGGUN, YOUNGSIK KIM, YOUSUNG WON, JONGSUNG WOO, JIWOO SEO, JISEONG KIM, SEONGBAE PARK, DOSAM HWANG, & KEY-SUN-CHOI.

2014.
Toward matching the relation instantiation from DBpedia ontology to Wikipedia text: Fusing FrameNet to Korean.
In *Proceedings of the 10th International Conference on Semantic Systems*, 13–19.

HARRIS, ZELLIG.
1954.
Ditributional Structure.
Word 10.146–62.

HAYOUN, AVI, & MICHAEL ELHADAD.
2016.
The Hebrew FrameNet Project.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, & Stelios

Piperidis, Paris, France. European Language Resources Association (ELRA).

HERMANN, KARL MORITZ, & PHIL BLUNSOM.
2014.
Multilingual models for compositional distributed semantics.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 58–68.
Association for Computational Linguistics.

KÖHN, ARNE.
2015.
What's in an embedding? Analyzing word embeddings through multilingual evaluation.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2067–2073.

———.
2016.

Evaluating embeddings using syntax-based classification tasks as a proxy for parser performance.
In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 67–71.

LINDÉN, KRISTER, HEIDI HALTIA, JUHA LUUKKONEN, ANTTI O LAINE, HENRI ROIVAINEN, & NIINA VÄISÄNEN.
2017.
FinnFN 1.0: The Finnish frame semantic database.
Nordic Journal of Linguistics 40.287–311.

RUPPENHOFER, JOSEF, JONAS SUNDE, & MANFRED PINKAL.
2010.
Generating FrameNets of Various Granularities: The FrameNet Transformer.
In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, &

Daniel Tapias, Valletta, Malta. European Language Resources Association (ELRA).

TSVETKOV, YULIA, MANAAL FARUQUI, WANG LING, GUILLAUME LAMPLE, & CHRIS DYER.
2015.
Evaluation of word vector representations by subspace alignment.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2049–2054.

ČULO, OLIVER.
2013.
Constructions-and-frames analysis of translations: The interplay of syntax and semantics in translations between English and German.
Constructions and Frames 5.143–167.

VIRK, SHAFQAT MUMTAZ, PHILIPPE MULLER, & JULIETTE CONRATH.
2016.

A supervised approach for enriching the relational structure of frame semantics in framenet.
In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3542–3552, Osaka, Japan. The COLING 2016 Organizing Committee.