University of Zurich[UZH]

Department of Computational Linguistics

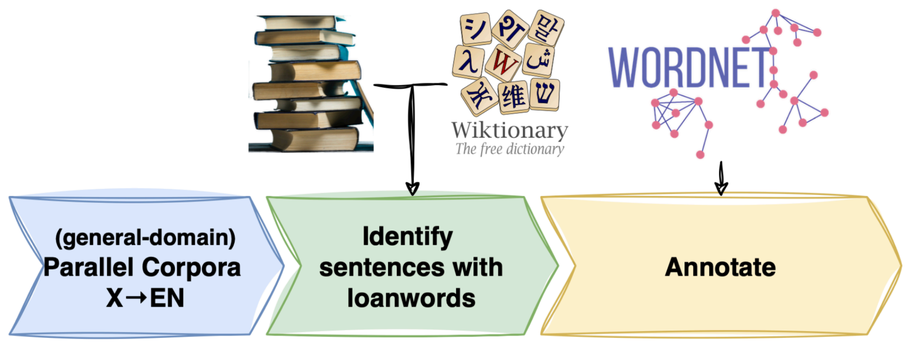
ACL 2025 VIENNA JULY 27 - AUGUST 1

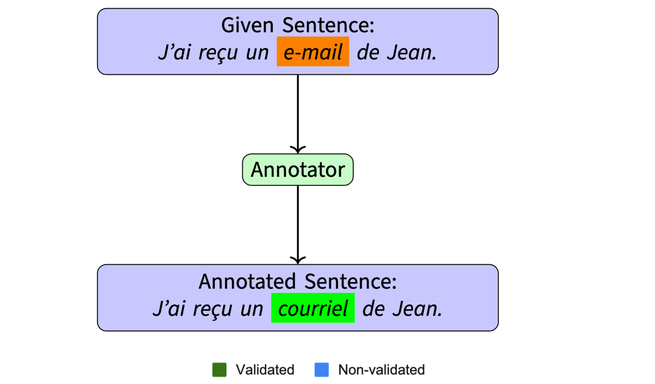
~~A Contrastive Multilingual Dataset for Evaluating Loanwords~~
A Matchwise Manytongued Givenshoard for Weighing Loanwords

Sina Ahmadi, Micha David Hess, Elena Álvarez-Mellado, Alessia Battisti, Cui Ding, Anne Göhring, Yingqiang Gao, Zifan Jiang, Andrianos Michail, Peshmerge Morad, Joel Niklaus, Maria Christina Panagiotopoulou, Stefano Perrella, Juri Opitz, Anastassia Shaitarova, Rico Sennrich
University of Zurich, UNED, University of Münster, University of Bern, Sapienza University of Rome

Motivation

- Loanwords studied for decades in historical linguistics but not that much in NLP/CL
- Under-explored NLP fields: NMT, language education, low-resource NLP, code-switching
- Missing: Loanwords in context and across languages, especially for machine translation
- **Need: Contrastive multilingual datasets to evaluate how NLP systems handle borrowed vs. native vocabulary**

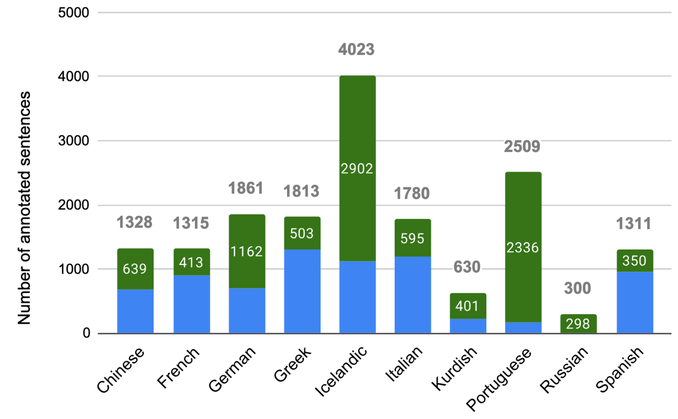




ConLoan

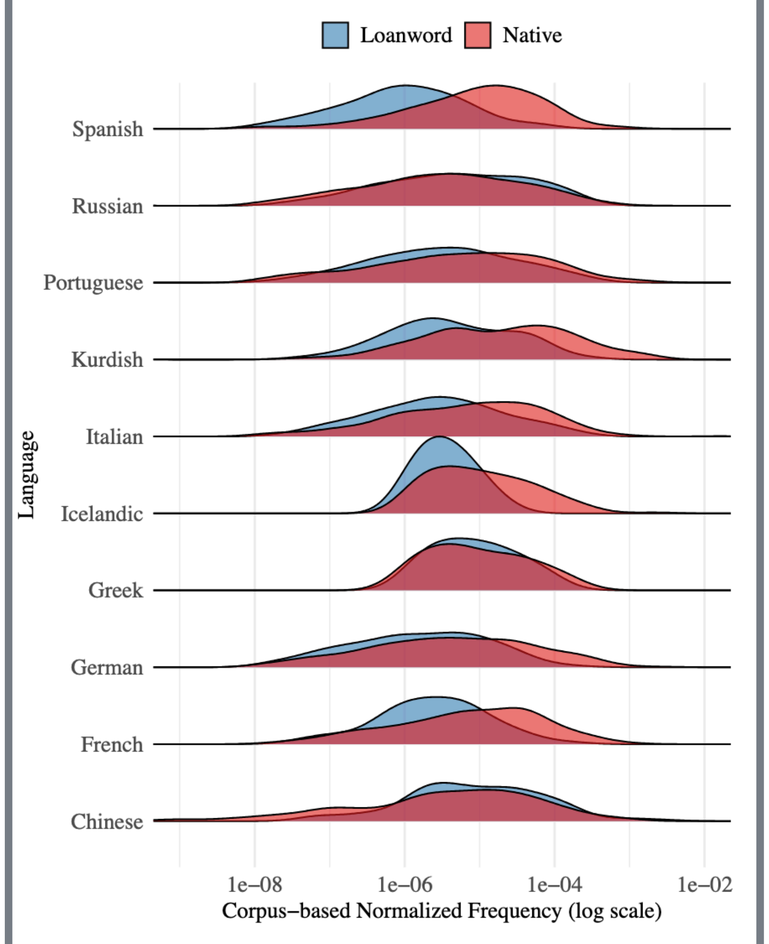
A contrastive dataset: loanwords in sentences are replaced by native alternative

- 16,870 sentences across 10 languages: **Chinese, French, German, Greek, Icelandic, Italian, Kurdish, Portuguese, Russian, Spanish**
- 56.9% contain loanwords
- 55.78% of the loanwords are replaced by native non-identical words.

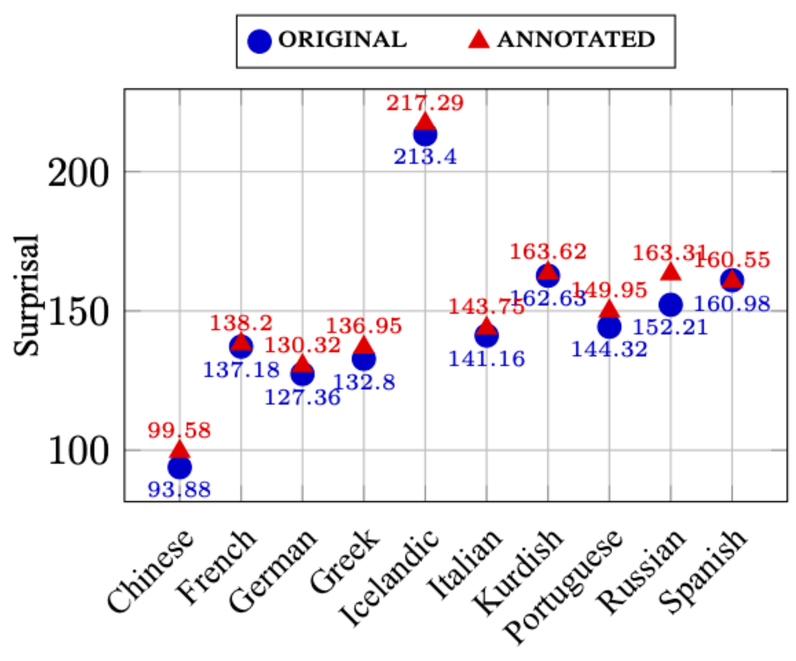


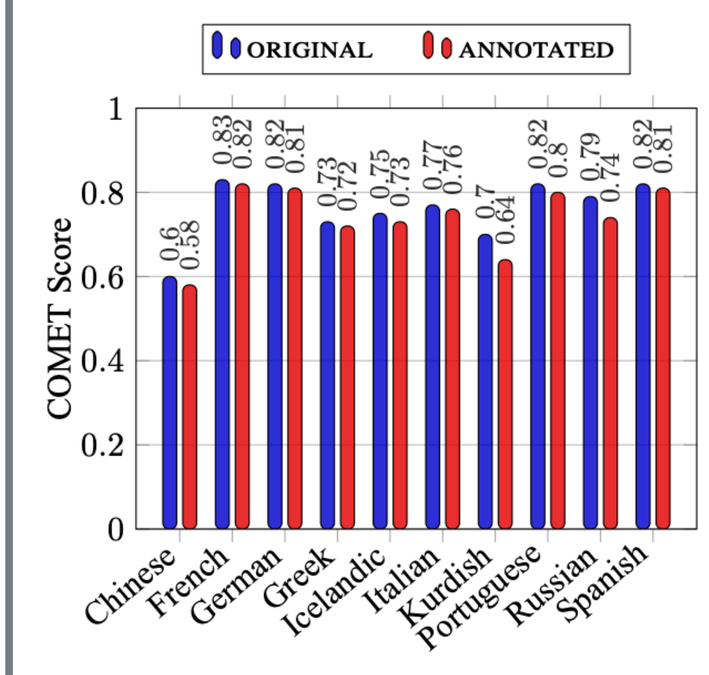
Key Findings

Finding 1: Native words are MORE frequent than loanwords in natural text



Finding 2: LLMs show LOWER surprisal for loanword sentences





Finding 3: NMT performs WORSE when translating native alternatives

Conclusion

- Loanword replacement is complex: varies significantly by language and context
- Comprehensive multilingual resources needed for loanword identification and analysis
- ConLoan enables evaluation of cross-linguistic lexical bias in LLMs
- Current NLP systems show systematic bias toward loanwords over native alternatives

Swiss National Science Foundation