Lexical Borrowing in Modern NLP

How Do Models Handle Loanwords?

Sina Ahmadi

Archimedes NLP Theme Meeting

December 18, 2025

Universität Zürich[UZH]

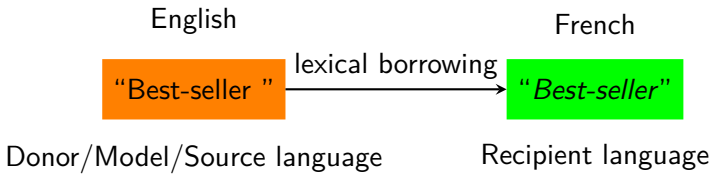

German
Greek
French
Latin
English
What the hell is this?

Introduction: Lexical Borrowing

- Borrowing: the process of adopting words from one language into another one.

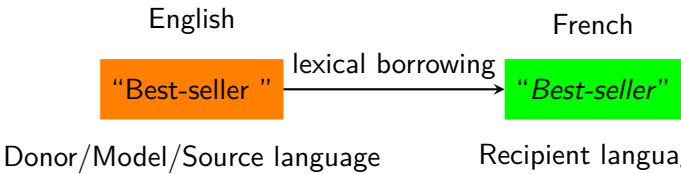
Introduction: Lexical Borrowing

- Borrowing: the process of adopting words from one language into another one.
- **Loanwords**: A word that at some point in the history of a language entered its lexicon as a result of *borrowing* (or *transfer*, or *copying*) (Haspelmath, 2009)



Introduction: Lexical Borrowing

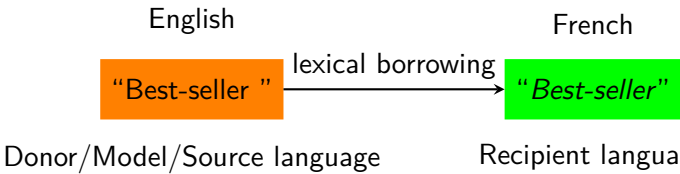
- Borrowing: the process of adopting words from one language into another one.
- **Loanwords**: A word that at some point in the history of a language entered its lexicon as a result of *borrowing* (or *transfer*, or *copying*) (Haspelmath, 2009)



- "Many loanwords start out as singly occurring switches that gradually get conventionalized" (Myers-Scotton, 1997)

Introduction: Lexical Borrowing

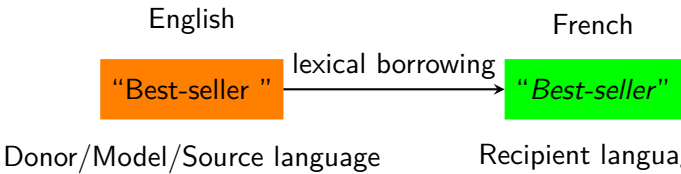
- Borrowing: the process of adopting words from one language into another one.
- **Loanwords**: A word that at some point in the history of a language entered its lexicon as a result of *borrowing* (or *transfer*, or *copying*) (Haspelmath, 2009)



- "Many loanwords start out as singly occurring switches that gradually get conventionalized" (Myers-Scotton, 1997)
- Loanwords are opposed to native words, i.e. words "which we can take back to the earliest known stages of a language" (Lehmann, 2013, p. 212)

Introduction: Lexical Borrowing

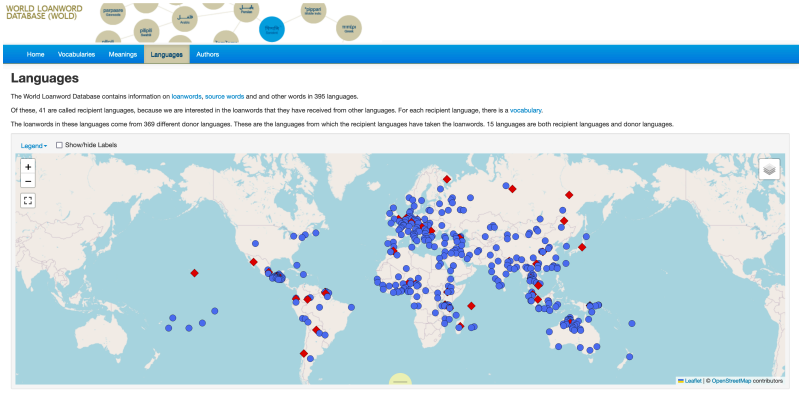
- Borrowing: the process of adopting words from one language into another one.
- **Loanwords**: A word that at some point in the history of a language entered its lexicon as a result of *borrowing* (or *transfer*, or *copying*) (Haspelmath, 2009)



- "Many loanwords start out as singly occurring switches that gradually get conventionalized" (Myers-Scotton, 1997)
- Loanwords are opposed to native words, i.e. words "which we can take back to the earliest known stages of a language" (Lehmann, 2013, p. 212)
- But then, what is even a ***native word***? ('disk', 'bikini', 'mother')

Motivation

- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (https://wold.clld.org)



Motivation

- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (https://wold.clld.org)
- In NLP/CL, there is a large body of research focusing on **loanword identification** (Mi et al., 2020; Nath et al., 2022)

Motivation

- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (https://wold.clld.org)
- In NLP/CL, there is a large body of research focusing on **loanword identification** (Mi et al., 2020; Nath et al., 2022)
- Many gaps still exist, including:

Motivation

- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (https://wold.clld.org)
- In NLP/CL, there is a large body of research focusing on **loanword identification** (Mi et al., 2020; Nath et al., 2022)
- Many gaps still exist, including:
  - Many under-explored fields: constrained decoding in NMT, language education, low-resourced NLP, code-switching

Motivation

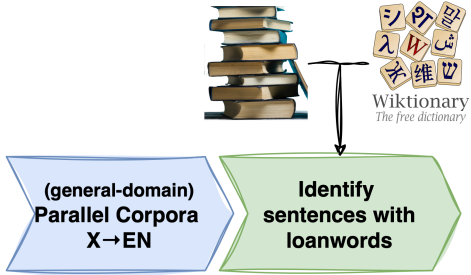
- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (`https://wold.clld.org`)
- In NLP/CL, there is a large body of research focusing on **loanword identification** (Mi et al., 2020; Nath et al., 2022)
- Many gaps still exist, including:
  - Many under-explored fields: constrained decoding in NMT, language education, low-resourced NLP, code-switching
  - $\Rightarrow$ **Loanwords in context and across languages esp. for machine translation**

Motivation

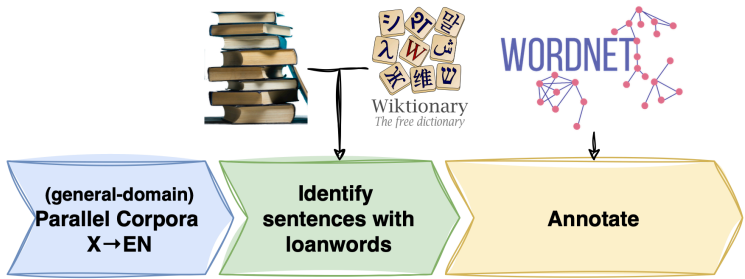
- Loanwords have been studied for decades in historical and comparative linguistics
  WOLD – the World Loanword Database containing loanwords in 395 languages (`https://wold.clld.org`)
- In NLP/CL, there is a large body of research focusing on **loanword identification** (Mi et al., 2020; Nath et al., 2022)
- Many gaps still exist, including:
  - Many under-explored fields: constrained decoding in NMT, language education, low-resourced NLP, code-switching
  - $\Rightarrow$ **Loanwords in context and across languages esp. for machine translation**
  - (Ahmadi et al., 2025, ACL 2025) & (Silva and Ahmadi, 2025, LREC 2026?)

ConLoan: Data Collection

ConLoan: Data Collection

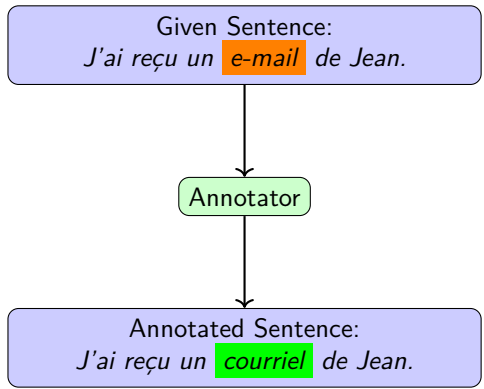


ConLoan: Data Collection



ConLoan: Annotation

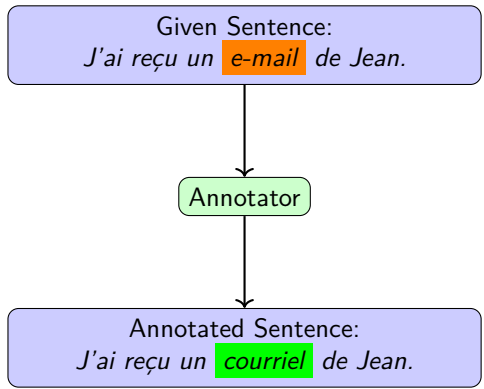
Create a contrastive dataset where in a given sentence loanwords are replaced by native alternatives



ConLoan: Annotation

Create a contrastive dataset where in a given sentence loanwords are replaced by native alternatives

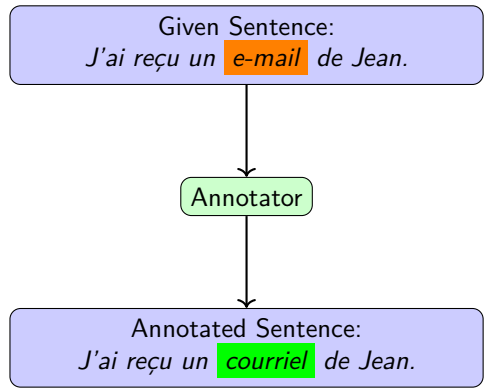
- Is there a loanword in the sentence?
- If yes, replace it by with a native alternative.



ConLoan: Annotation

Create a contrastive dataset where in a given sentence loanwords are replaced by native alternatives

- Is there a loanword in the sentence?
- If yes, replace it by with a native alternative.
- Retention of loanwords: If no native alternative is known, the original loanword is kept in the sentence.

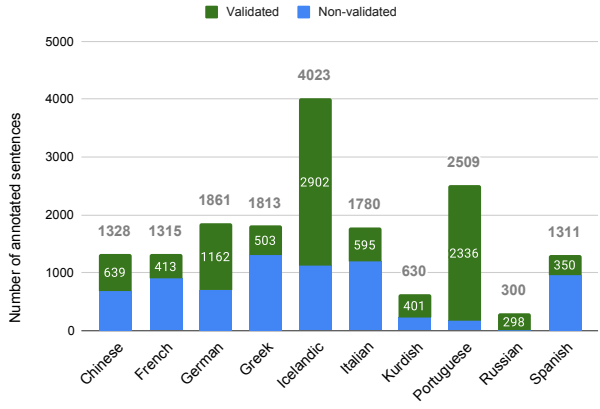


ConLoan: Annotation

Source	Target	Replacement Suggestions	Comments?
Όταν έρθει η ώρα να κοιμηθούν, τοποθετήστε τις τσάντες τους στο καρότσι αποσκευών και κυλήστε το *στο παιδικό* **<L1>ασανσέρ</L1>** *που θα* τους μεταφέρει στα δωμάτιά τους.	When it's time to go to sleep, place their bags in the luggage cart and roll it onto the kid-powered elevator to bring them to their rooms.	**ανελκυστήρας** **ασανσέρ**	
Όταν έρθει η ώρα να κοιμηθούν, τοποθετήστε τις τσάντες τους στο καρότσι αποσκευών και κυλήστε το *στο παιδικό* **<N1>ανελκυστήρα</N1>** *που θα* τους μεταφέρει στα δωμάτιά τους.	When it's time to go to sleep, place their bags in the luggage cart and roll it onto the kid-powered elevator to bring them to their rooms.		
☑			
Φαίνεται ότι τους διαφεύγει ο *δραματικός σχεδόν* **<L1>συμβολισμός</L1>** της ενέργειας αυτής.	It seems that they fail to grasp the almost dramatic symbolism of this action.	**συμβολική αναπαράσταση** **συμβολισμός**	
Φαίνεται ότι τους διαφεύγει ο *δραματικός σχεδόν* **<N1></N1>** της ενέργειας αυτής.	It seems that they fail to grasp the almost dramatic symbolism of this action.		
☐			

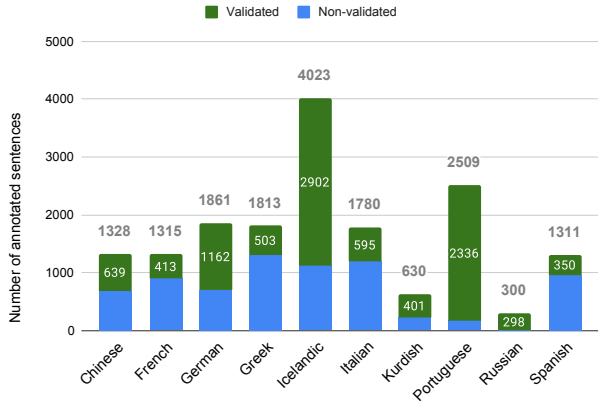
ConLoan: Statistics

- Out of the 16,870 sentences, 56.9% were checked as containing loanwords.



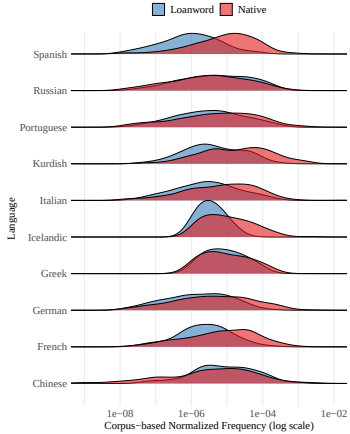
ConLoan: Statistics

- Out of the 16,870 sentences, 56.9% were checked as containing loanwords.
- 55.78% of the loanwords are replaced by native non-identical words.



ConLoan: Statistics

- Out of the 16,870 sentences, 56.9% were checked as containing loanwords.
- 55.78% of the loanwords are replaced by native non-identical words.
- Native words are more frequent than their loanword counterparts (except for Chinese)

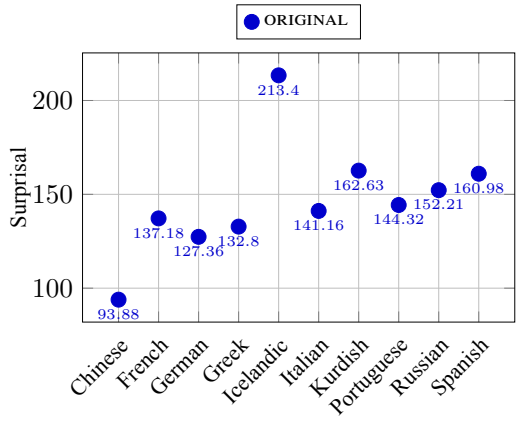


Experiments: Surprisal

- Surprisal measures sentence unpredictability using language models (Llama 2.7 & 3.1, EuroLLM)

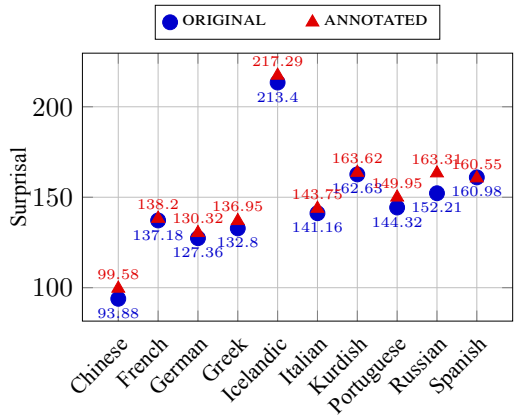
Experiments: Surprisal

- Surprisal measures sentence unpredictability using language models (Llama 2.7 & 3.1, EuroLLM)
- Higher surprisal = less probable sentences or model limitations



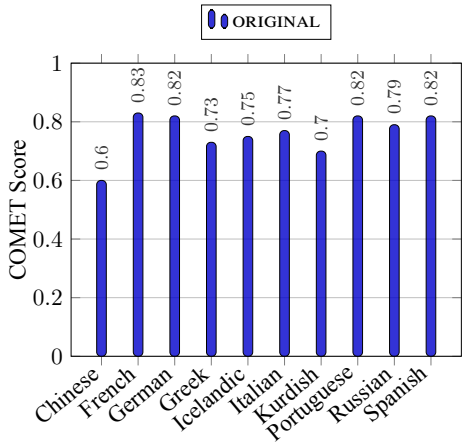
Experiments: Surprisal

- Surprisal measures sentence unpredictability using language models (Llama 2.7 & 3.1, EuroLLM)
- Higher surprisal = less probable sentences or model limitations
- LLMs show lower surprisal for sentences with loanwords



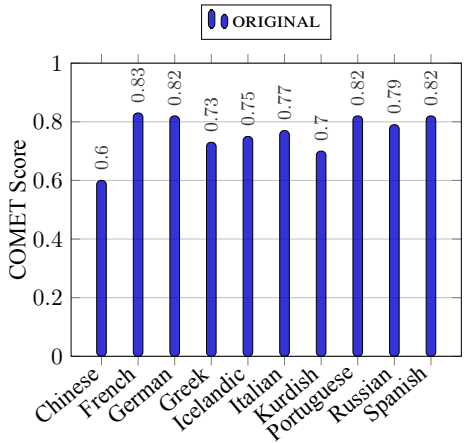
Experiments: Neural Machine Translation X→EN

- How does NMT perform on sentences with vs. without loanwords?



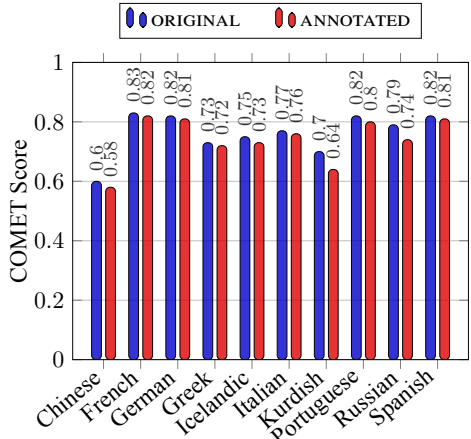
Experiments: Neural Machine Translation X→EN

- How does NMT perform on sentences with vs. without loanwords?
- Robust NMT should perform equally on both



Experiments: Neural Machine Translation X→EN

- How does NMT perform on sentences with vs. without loanwords?
- Robust NMT should perform equally on both
- NLLB translates loanword sentences more efficiently than native alternatives

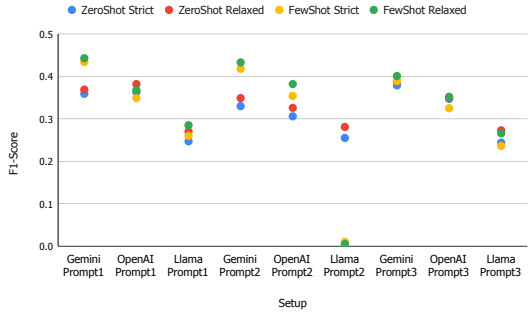


Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?

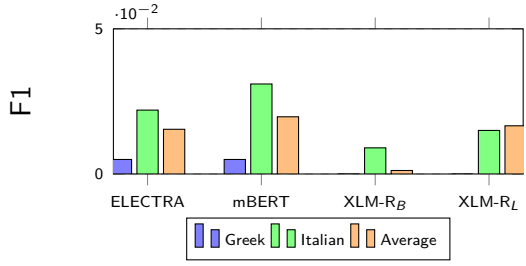
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts



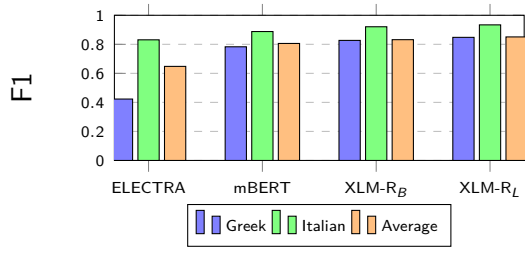
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)



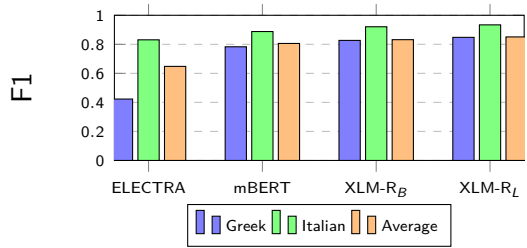
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)
- Fine-tuned XLM-R$_L$: F1 $= 0.85$



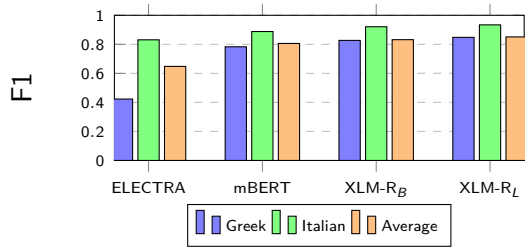
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)
- Fine-tuned XLM-R$_L$: F1 $= 0.85$
- Models struggle with:



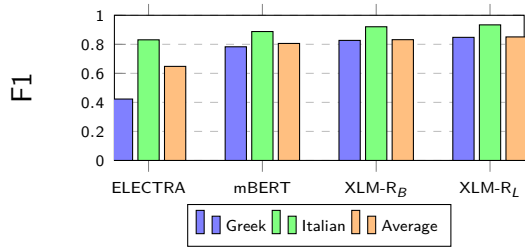
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)
- Fine-tuned XLM-R$_L$: F1 $= 0.85$
- Models struggle with:
  - Code-switching vs. loanwords



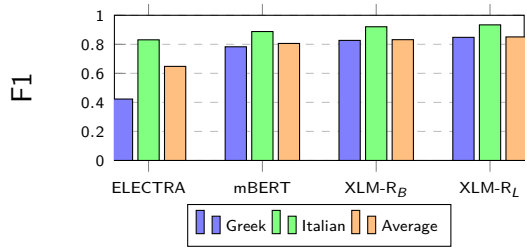
Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)
- Fine-tuned XLM-R$_L$: F1 $= 0.85$
- Models struggle with:
  - Code-switching vs. loanwords
  - Named entities



Experiments: Loanword Identification

- Can models identify loanwords when explicitly asked?
- LLMs (Gemini, GPT-4.1, Llama-3): F1 < 0.50 across all prompts
- Zero-shot encoders: F1 ≈ 0 (near random)
- Fine-tuned XLM-R$_L$: F1 $= 0.85$
- Models struggle with:
  - Code-switching vs. loanwords
  - Named entities
  - Greco-Latin terminology



Key Findings

1. This study highlights the need for more comprehensive multilingual resources for loanword identification and analysis

Key Findings

1. This study highlights the need for more comprehensive multilingual resources for loanword identification and analysis
2. Identifying loanwords and finding appropriate native replacements is complex, varying by language and context

Key Findings

1. This study highlights the need for more comprehensive multilingual resources for loanword identification and analysis
2. Identifying loanwords and finding appropriate native replacements is complex, varying by language and context
3. Current NMT neural models and LLMs are biased towards processing loanwords more efficiently than native alternatives in some contexts

Thank you!

Contact: sina.ahmadi@uzh.ch

Resources: https://github.com/ZurichNLP/ConLoan

Questions?

References

Ahmadi, S., Hess, M. D., Álvarez-Mellado, E., Battisti, A., Ding, C., Göhring, A., Gao, Y., Jiang, Z., Michail, A., Morad, P., Niklaus, J., Panagiotopoulou, M. C., Perrella, S., Opitz, J., Shaitarova, A., and Sennrich, R. (2025). ConLoan: A contrastive multilingual dataset for evaluating loanwords. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30070–30090, Vienna, Austria. Association for Computational Linguistics.

Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. *Loanwords in the world's languages: A comparative handbook*, 35:54.

Lehmann, W. P. (2013). *Historical linguistics: An introduction*. Routledge.

Mi, C., Xie, L., and Zhang, Y. (2020). Loanword identification in low-resource languages with minimal supervision. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):1–22.

Myers-Scotton, C. (1997). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Nath, A., Saravani, S. M., Khebour, I., Mannan, S., Li, Z., and Krishnaswamy, N. (2022). A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013.

Silva, M. S. and Ahmadi, S. (2025). Language models are borrowing-blind: A multilingual evaluation of loanword identification across 10 languages.