Loanword Annotation Task

Sina Ahmadi (sina.ahmadi@uzh.ch)

June 25, 2024

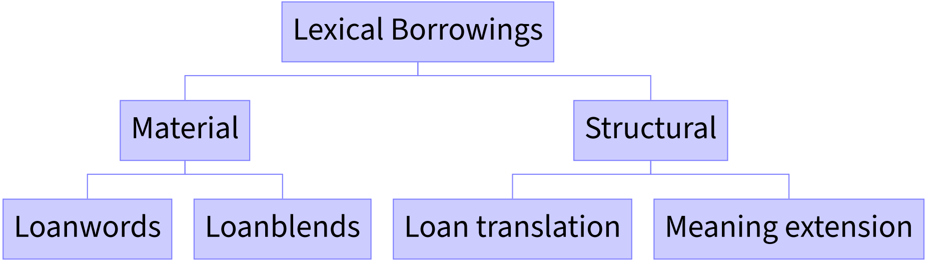
Outline

- Loanwords
- Annotation task description
- Demo

German
Greek
French
Latin
English
What the hell is this?

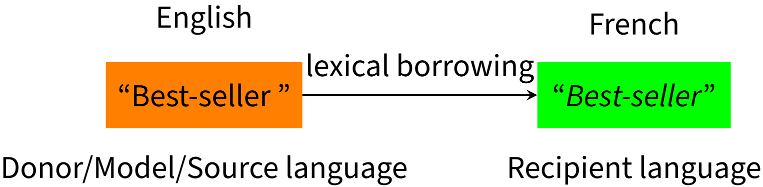
Introduction

Lexical Borrowing



Lexical Borrowing: loanwords

- **Material**: borrowing of sound-meaning pairs, i.e. lexemes
- **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]


English

French

“Best-seller ”

lexical borrowing

“*Best-seller*”

Donor/Model/Source language

Recipient language

- 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]
- “Many loanwords start out as singly occurring switches that gradually get conventionalized” [Myers-Scotton, 1997]

Lexical Borrowing: loanwords (cont.)

- But then, what is even a ***native word***?
- The status of native words is always relative to what we know about the history of a language
- Are these native words?

Word	Cognate
‘disk’	Proto-West Germanic **disk* (→ *Tisch*)
‘window’	Old Norse *vindauga*
‘bikini’	Marshallese *Pikinni*
‘mother’	PIE **mater*

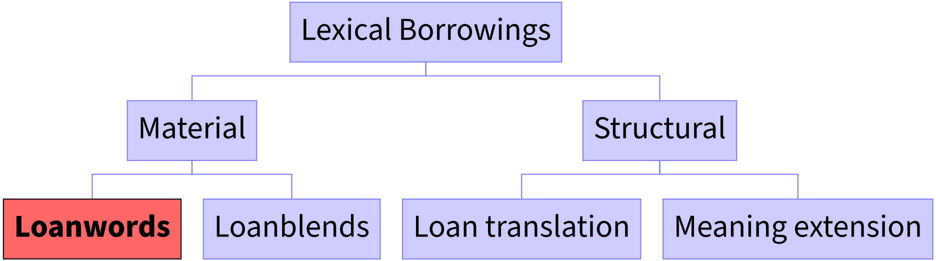
⇒ **We can identify loanwords, but we cannot identify “non-loanwords”, i.e. a word for which we have no knowledge that it was borrowed.**

Lexical Borrowing: types

- **Material**: borrowing of sound-meaning pairs, i.e. lexemes
  - **Loanwords**
  - **Loanblends**: hybrid borrowings which consist of partly borrowed material and partly native material
    - Greek ‘σουβλατζής’ where *-τζής* borrowed from Turkish *-cı*
    Not widely attested: Most hybrid-looking expressions are loan-based creations
    - English ‘desk lamp’ are loanwords etymologically but not loanwords

Lexical Borrowing: types (cont.)

- **Structural**: copying of syntactic, morphological or semantic pattern
  - **Loan translations, aka calques**: item-by-item translation of a complex lexical unit
    - English: 'loanword' calqued from German '*Lehn-wort*'
    - Kurdish: *da-bezandin* calqued from English 'download'
    - French: '*presqu'île*' calqued from Latin *paen-insula* 'almost-island'
  - **Meaning extension**: polysemy pattern of a donor language is copied, e.g. word order patterns, case-marking patterns
    - German '*Kopf*' from English 'head' in a syntactic phrase

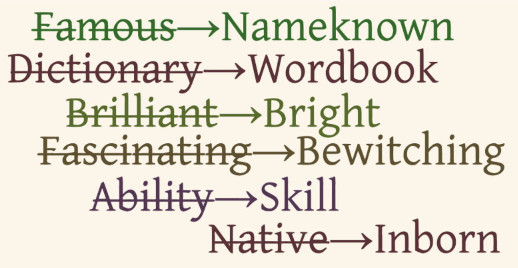


This is what we want!

Lexical Borrowing: language purity

Studying borrowing is sometimes stigmatized due to political ideologies:

- Political Ideologies
  - Totalitarian Regimes
  - Colonialism
  - Nationalism
- Enforcement methods
  - Education policies
  - Media regulations
  - Legal requirements
- Stigmatization
  - Associated with discrimination
  - Obstacle to linguistic diversity



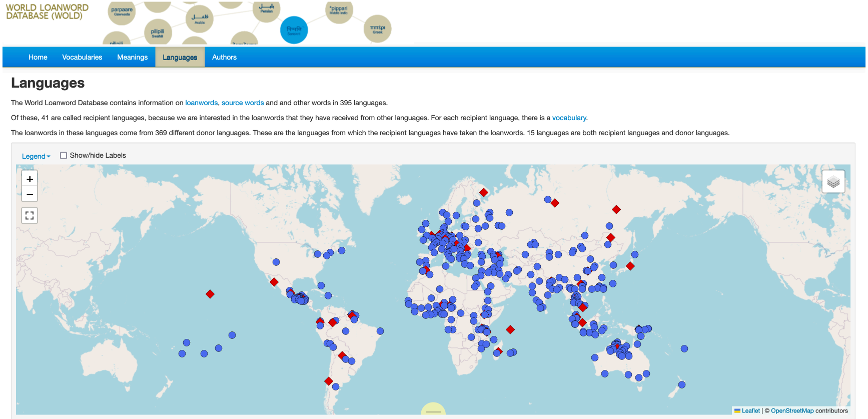
~~Famous~~→Nameknown
~~Dictionary~~→Wordbook
~~Brilliant~~→Bright
~~Fascinating~~→Bewitching
~~Ability~~→Skill
~~Native~~→Inborn

Anglish language (https://anglish.org/wiki/Anglish)

NLP and Loanwords

Lexical Borrowing: Linguistics

- Loanwords have been studied for decades in the context of historical and comparative linguistics
- WOLD – the World Loanword Database contains information on loanwords, source words and other words in 395 languages (`https://wold.clld.org`).



Languages

The World Loanword Database contains information on loanwords, source words and and other words in 395 languages.

Of these, 41 are called recipient languages, because we are interested in the loanwords that they have received from other languages. For each recipient language, there is a vocabulary.

The loanwords in these languages come from 369 different donor languages. These are the languages from which the recipient languages have taken the loanwords. 15 languages are both recipient languages and donor languages.

NLP and Loanwords

- Borrowings often account for a common source of out-of-vocabulary words
- Automatically detecting lexical borrowings from text has proven to be relevant for NLP tasks:
  - Parsing [Alex, 2008], ASR [Leidig et al., 2014], SMT [Tsvetkov and Dyer, 2016]
- A large body of research focuses on **loanword identification** [Mi et al., 2020, Nath et al., 2022]
- A couple of initiatives like ADoBo – automatic detection of borrowings [Mellado et al., 2021]
- Many under-explored applications: constrained decoding in NMT, language education, low-resourced NLP
- **Existing gaps: loanwords in context and across languages in machine translation**

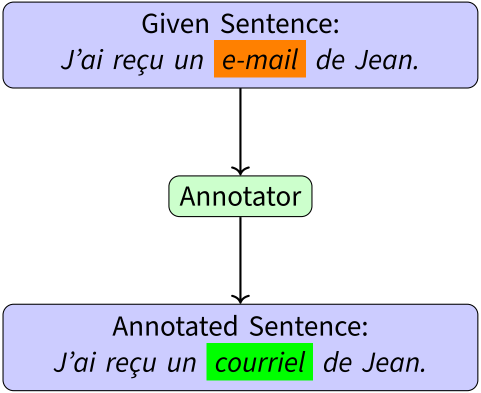
Objectives of the Annotation Task

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

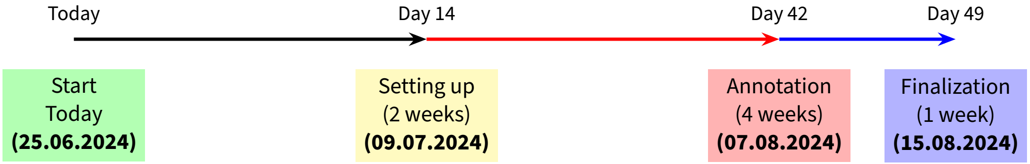
- You will be given some suggestions
- Annotation in two directions
- The tool saves your annotations

  `https://github.com/chamisshe/CLoAn`

  (Thanks to Micha Hess!)


Given Sentence:
J'ai reçu un e-mail de Jean.
Annotator
Annotated Sentence:
J'ai reçu un courriel de Jean.

Task Workflow



Today | Day 14 | Day 42 | Day 49

Start Today **(25.06.2024)**

Setting up (2 weeks) **(09.07.2024)**

Annotation (4 weeks) **(07.08.2024)**

Finalization (1 week) **(15.08.2024)**

1. Setting up:
   - Provided a parallel corpus, can it be useful for this task?
   - Loanword list: Is there a list of loanwords and their native alternatives for your language? (e.g. `https://www.academie-francaise.fr`)
2. Loanword annotation:
   - Using the annotation tool, replace loanwords in the sentences (or the other way)
3. Documentation:
   - Facing any challenges or interesting cases? Make sure to write them done.
   - Make sure to backup your work everyday
   - Calculate the time spent on the task

Potential Challenges

- The annotation tool doesn't suggest relevant replacements → rely on external sources, e.g. online dictionaries
- A few criteria to recognize loanwords:
  1. **Foreignisms**: "*il a stalké ses voisins*"
  2. **Morphology**: if the word is morphologically analyzable in one language but unanalyzable in another one, then it must come from the first language.
     - German 'Grenze' (border) ← Polish '*granica*'
  3. **Phonology**: if a word shows signs of phonological integration in language A but not in language B, it must come from language B.
     - English 'facade' ← French '*façade*'
  4. **Related languages**: if the word is attested in a sister language of language B that cannot have been under the influence of language A, it must come from language B.
  5. **Meaning**: French 'baskets' for 'sneakers'

  ⇒ Not always easy: rely on your hunch!

Join our channel on Slack: `#loanword-annotation`

References

Alex, B. (2008).
Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection.
In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May-1 June 2008, Marrakech, Morocco*, pages 2693–2697. European Language Resources Association (ELRA).

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.

Leidig, S., Schlippe, T., and Schultz, T. (2014).
Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus.
In *SLTU*, pages 207–214.

Mellado, E. Á., Anke, L. E., Arroyo, J. G., Lignos, C., and Zamorano, J. P. (2021).
Overview of adobo 2021: Automatic detection of unassimilated borrowings in the Spanish press.
arXiv preprint arXiv:2110.15682.

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.

Tsvetkov, Y. and Dyer, C. (2016).
Cross-lingual bridges with models of lexical borrowing.
Journal of Artificial Intelligence Research, 55:63–93.