Resource-Lean Lexicon Induction for German Dialects

Robert Litschko, Barbara Plank, Diego Frassinelli
robert.litschko@lmu.de

LREC 20 Palma 26

Motivation

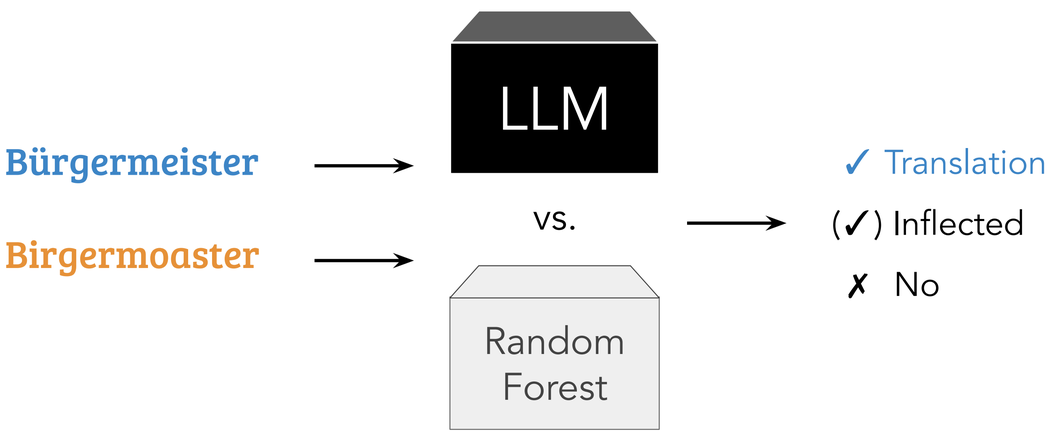
- Performance of NLP tools and LLMs depends on how well language is represented in training data.
- High-resource languages benefit, e.g., from robust translation and retrieval systems.
- Performance degrades under dialect spelling variation:

Bürgermeister (mayor) ↔ Birgermoaster Bürgermaischter Birgermóaster Burgermoaster Byrgermoaster Bürgermoaster Birgermeister Birgermaster

Bavarian spelling variations found in Wikipedia.

Word-Pair Classification

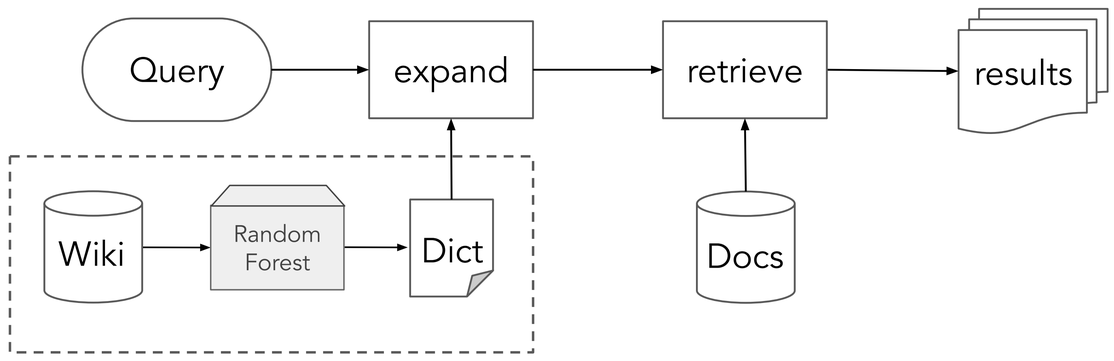
- Classify if pairs of German lemmas and dialect words are (direct or inflected) translations.
- LLMs perform poorly at identifying correct translations.[1]
- We evaluate a statistical model with string similarity features.



[1] *Make Every Letter Count: Building Dialect Variation Dictionaries from Monolingual Corpora (Litschko et al., 2025)*

Cross-Dialect Information Retrieval

- German-to-dialect retrieval, spelling variations in documents.[1]
- Dialects: Ripuarian (ksh), Low German (nds), Alemannic (als), Rhine Franconian (pfl), Bavarian (bar).
- Query expansion: augment queries with spelling variations.



[1] *Cross-Dialect Information Retrieval: Information Access in Low-Resource and High-Variance Languages (Litschko et al., 2025)*

String Similarity Features

Our string similarity features are adopted from prior work[1] and were originally developed for classifying words as cognates or false friends.

$$\text{DICE}(x, y) = \frac{2 \times |\text{ngrams}(x) \cap \text{ngrams}(y)|}{|\text{ngrams}(x)| + |\text{ngrams}(y)|}$$

rg, ge, er, rm,st

Bü, ür, rg, ge, er, rm, me, ei, is, st, te

Bi, ir, rg, ge, er, rm, mo, oa, as, st

$$\text{LCSR}(x, y) = \frac{|\text{LCS}(x, y)|}{\max(|x|, |y|)}$$

B ü r g e r m e i s t e r
B i r g e r m o a s t e r

Set-based String Similarity

- DICE coefficient between ngrams(x) and ngrams(y):
  - Shared character bigrams, trigrams,[2] and "extended trigrams" (XDICE).[3]
- XXDICE incorporates positional information.[3]

Sequence-based String Similarity

- Length of longest common prefix (PREFIX)
- Longest common subsequence ratio (LCSR;[4] uni-, bi-, trigram sequences)
- Length-normalized edit distance (NED;[5,6] uni-, bi-, trigram sequences)
- Edit distance between phonetic codes (cologne phonetics)[7]

Code and data:

[1] Automatic identification of cognates and false friends in French and English (Inkpen et al., 2005)
[2] *The use of an association measure based on character structure to identify semantically related pairs of words and document titles (Adamson and Boreham, 1974)*
[3] *Word-pair extraction for lexicography (Brew et al., 1996)*
[4] *Bitext maps and alignment via pattern recognition (Milamed, 1999)*
[5] *Binary codes capable of correcting deletions, insertions and reversals (Levenshtein, 1966)*
[6] *Identification of confusable drug names: A new approach and evaluation methodology (Kondrak and Dorr, 2004)*
[7] *Die Kölner Phonetik - Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse (Postel, 1969)*

Results