Nunc profana tractemus. Detecting Code-Switching in a Large Corpus of 16th Century Letters

Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Schroffenegger,
Raphael Schwitter, Phillip Ströbel, Benjamin Suter

The Bullinger Letter Collection

This paper introduces a corpus of 12'000 letters in **Latin and Early New High (ENH) German** for studies in history, linguistics, and theology. The corpus also contains a few letters in French, Greek and Italian. We focused on the detection of code-switching between Latin and ENH German.

Latin to **German**	Crastino comitia erunt Domus tantum Dei propter dissidium Zuziensium et Samadensium **von stok und galgen wegen**.
German to **Latin**	Dann **Galli nostri** treüwend unnd erschreckend mengem das hertz, das er hinschlichen last, **ne privetur stipendio**.

Example sentences from our corpus with intra-sentential code-switching.

Motivations for Code-Switching Detection

1. Linguistic Studies
2. Language-specific Search
3. Training Language-specific Hand-written Text Recognition (HTR)
4. Machine Translation from Latin to Modern German
5. Normalisation of ENH-German

Sentence-based Language Identification

We trained our language identifier FurL on

- 150 sentences of (16th century) Latin and
- 150 sentences of ENH-German

We tested FurL on Caesar's "Bello Gallico" (314 sentences cut down to 20 characters): 100% correctly classified as Latin.

In our corpus, FurL classified around

- 165,500 sentences as Latin (with 2.7 million tokens)
- 39,600 sentences as ENH-German (with 0.8 million tokens)

If the number of characters for a letter exceeded 3% for either language, OR if it has at least two sentences with at least 30 characters in the other language, then we counted it as code-switching letter.

Code-sw ENHG	Code-sw Latin	ENHG	Latin
688	1330	920	5309
2018		6229	

2018 of 8247 letters (24%) contain code-switching on the sentence level.

Word-based Language Identification

Examples of the overlapping vocabularies:

token	freq(DE)	freq(LA)	vocab
Albrecht	41	1	German
Alexander	9	18	undec
Africa	2	10	undec
Augustinus	5	147	Latin
in	9298	50,340	undec
bis	259 *up to*	145 *twice*	undec
breve	9 *letter*	67 *short*	undec
briefen	22 *letters*	1 –	German
dies	17 *this*	1236 *day*	Latin

Word-based Language Identification

Step 1: **Collecting Latin and ENH-German vocabularies**

- Accept all sentences classified as Latin and collect the tokens as LA-vocabulary: ∼ 158,600 types
- Accept all sentences classified as ENH-German and collect the tokens as DE-vocabulary: ∼ 72,000 types

Both vocabularies contain word form types of the other language because of code-switching in the sentences (and because of few language misclassifications).
Step 2: **Filtering the Latin and ENH-German vocabularies**

- If a type in the LA vocabulary is also in the DE vocabulary, then keep it if it is **10 times** more frequent in LA: ∼ 152,300 types
- If a type in the DE vocabulary is also in the LA vocabulary, then keep it if it is **5 times** more frequent in DE: ∼ 64,600 types

Steps 3 and 4: **Classification of each token in each sentence**

- If a token is in either vocabulary, then classify it accordingly.
- If a token is in **neither** vocabulary, then classify it based on the surrounding words.

Result: If at least 2 subsequent words in a sentence are classified with a language that is different from the sentence language, then we count the sentence as code-switching.
We obtain 1505 sentences with intra-sentential code switching.
Evaluation: we manually checked 50 sentences with a total of 1075 tokens:
99% of the tokens get the correct language label

Conclusion

- The pretrained model of the language identifier `langid` did not reliably distinguish between Latin and German.
- Our special-purpose language identifier FurL which we trained on only 150 sentences worked well for binary German vs. Latin sentence-level language labelling.
- Based on this sentence classification, we bootstrapped a word-based language identifier which works with high accuracy and reliably identifies sentence-internal code-switches.

Our method is easily applicable and guarantees high lexical coverage which is important for languages like ENH German with many spelling variants. We will make both the corpus and the digital edition available online. The corpus will size up to roughly 3.5 million tokens in Latin and 1.2 million in ENH German.

LREC, Marseille, June 2022
We gratefully acknowledge project funding provided by various sponsors through the UZH Foundation (see `www.bullinger-digital.ch/about`).

Contact

University of Zurich[UZH]

Department of Computational Linguistics
`http://www.cl.uzh.ch`

Martin Volk
Andreasstrasse 15, CH-8050 Zurich
volk@cl.uzh.ch