Identification of Morpheme Origin

Aleš Manuel Papáček 2025
Supervisor: Zdeněk Žabokrtský

Task

Given morphologically segmented Czech sentences, the task is to determine for each morpheme whether it is native or borrowed, and if borrowed, to identify the languages through which it entered Czech.

Thesis goals

- Formalize the problem of determining the etymological origin on morpheme level
- Create manually annotated dataset for this task
- Extract features that can be used to train classifier model that will beat given baselines

Dataset

The input sentences, taken from the SIGMORPHON 2022 Task, were already morphologically segmented. Each morpheme was then manually annotated with its etymological origin. This is the first dataset of its kind for Czech.

Subset	# Sentences	# Words	# Morphemes
Training	200	2,774	7,016
Development	50	599	1,460
Test	50	609	1,485

Baselines

- **All Native** – Predicts Czech origin for all morphemes.
- **Most Frequent Origin** – Remembers the most common origin for each morph from training. Defaults to Czech if unseen.
- **Word Lemmatization** – Lemmatizes each word, assigns the root's origin from a word-level etymological dictionary, and checks affixes against a list of known borrowed affixes.
- **Large Language Model** – OpenAI's o3 model, prompted to annotate morpheme origins.

Model

We trained several classifiers to predict the origin of each morpheme from the extracted features. Among others, we tested MLPs, SVMs, and logistic regression models.

Repository:
`https://github.com/ampapacek/MorphemeOrigin`

Features

Each morpheme is represented individually using the following features:

- **Character n-grams** – 1-grams and 2-grams extracted from the morpheme
- **Morpheme Type** – root / derivational affix / inflectional affix (one-hot encoded)
- **Morpheme Position** – prefix / root/ interfix / suffix
- **Vowels** – Binary flags for vowel at start/end.
- **Embeddings** – FastText embedding vectors

Results

Model	F1	RER	Native	Borrowed
All Native	90.0	0.0	**100.0**	0.0
Most Frequent Origin	94.4	43.4	99.2	50.9
Word Lemmatization	**94.8**	**48.1**	98.6	**60.9**
Large Language Model	94.4	43.5	99.3	50.9
SVM	95.0	50.0	**99.3**	56.1
MLP30	95.3	53.0	98.9	63.2
MLP300-embedding	95.4	54.2	99.1	62.3
MLP30-extended	**96.1**	**61.0**	98.4	**75.4**
MLP30-self-train	96.1	60.9	98.3	75.6
MLP30-extended+dev	**96.8**	**67.9**	98.9	**77.8**

RER: relative error reduction, Native/Borrowed: category-specific F1.
Extended: additional training data from etymological dictionary
+dev: trained on combined train and dev sets
self-train: trained on dataset labelled by the same model

Conclusion

- I created morpheme-level etymology dataset for Czech.
- I trained and evaluated multiple classifiers. The best model reduced error by 67.9 % over the baseline.
- Embeddings and semi-supervised training showed no substantial gain.

Future Work

- Expand the dataset with additional annotated data
- Explore ways to reduce the need for manual annotation
- Experiment with complex architectures, including fine-tuned pretrained large language models
- Apply the approach to other languages