Tracing Morph Origins in Czech: A Computational Approach to Morph-Level Etymology

Aleš Manuel Papáček, Zdeněk Žabokrtský
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics (ÚFAL)
Malostranské náměstí 25, 118 00 Prague, Czech Republic
papacek@ufal.mff.cuni.cz, zabokrtsky@ufal.mff.cuni.cz

Abstract

Modern languages remain connected to ancient ones in multiple ways, including through etymology; for instance, Latin is among the most influential sources of borrowings in (modern) Czech, whether transmitted directly or mediated through other languages. This work focuses on predicting the etymological origin of individual morphs in Czech words. Given morphologically segmented Czech sentences, the task is to determine for each morph whether it is native or borrowed, and if borrowed, to identify the languages through which it entered Czech. Although some linguists have examined etymology at the level of individual morphs rather than whole words (Arkadiev et al., 2015), to our knowledge, no computational work has yet addressed this level of analysis. We created a manually annotated dataset of 300 Czech sentences comprising around 10,000 morphs with morph-level etymology labels, and trained supervised models using character-based and structural features. Our best lightweight system is a feed-forward neural network with a single hidden layer, trained on data augmented with entries from an etymological dictionary, reaching 96.2 % F1 on the test set. We also developed and tested several prompting variants for large language models; the best model `Claude-Opus-4.5`, achieved 97.8 % F1. We release the code, prompts, and dataset as open source at `https://github.com/ampapacek/MorphemeOrigin`.

Keywords: Etymology, Language contact, Borrowing, Morphological segmentation, LLM

1. Introduction

Borrowing is one of the main forces driving language change (Grant, 2020). English, for example, owes much of its vocabulary to French and, through it, to Latin. Latin itself borrowed extensively from Greek (Durkin, 2014). What gets borrowed is not always uniform at the word level: borrowed roots often combine with native morphology, and native roots can take borrowed affixes.

Modern European languages therefore contain many hybrid words whose morphs come from different sources. For instance, *antiviruses* combines the Greek-derived prefix *anti-* with the Latin root *virus* and the English inflectional ending *-es*.

Morphs are the building blocks of words, yet most existing etymological resources stop at the word level. Studying morph origins offers a much finer view of language history and contact. Linguistic research has explored morph borrowing in works such as Arkadiev et al. (2015) and Seifart (2015). However, so far, existing computational approaches have focused on word-level etymology rather than morph-level analysis.

Automatically predicting morph origins is difficult: the space of possible morphs is large, and many morph surface forms are short and ambiguous, meaning the same form can occur in different words with different etymological origins. This is also hard to model computationally because much of the relevant knowledge is implicit in linguistic expertise, and existing etymological resources are mostly word-level rather than fine-grained at the level of word subparts.

We frame morph-level etymology prediction as a supervised learning task. Given a segmented sentence, the goal is to decide for each morph whether it is native or borrowed and, if borrowed, which languages are involved in its borrowing path.

To enable this, we manually annotated almost 10,000 morphs in 300 sentences from the SIGMORPHON 2022 shared task (Batsuren et al., 2022) with morph-level etymological labels.

We compare several baselines, including a majority-label baseline, memorization of morph–origin pairs, and rule-based methods using the Czech Etymological Lexicon (Rejzek et al., 2025). For the supervised approach, we extract character-based and structural features (character n-grams, morph type, and position within the word) and train standard classifiers: logistic regression (LR), support vector machines (SVM), and multilayer perceptrons (MLP). We also test morph and word embeddings and a semi-supervised self-training setup.

In addition, we evaluate prompt-only predictions from several state-of-the-art LLMs under different in-context training settings.

Although the general ideas are language-independent, our experiments focus on Czech, where suitable etymological resources are available, primarily the Czech Etymological Dictionary (Rejzek, 2019). We believe that the proposed

Proceedings of LT4HALA 2026 @ LREC 2026, pages 7–18.

methodology should in principle be transferable to other languages, given the availability of morph-level etymological annotations.

The rest of the paper is organized as follows. Section 2 introduces key concepts from morphology and etymology that provide the necessary background. Section 3 reviews related research. Section 4 describes the data and annotation process and Section 5 outlines the design of our machine learning approaches. Section 6 explains the evaluation setup and Section 7 reports and analyses the experimental results.

In addition, the appendix A presents an etymological analysis of a diachronic Czech legal corpus covering more than 100 years. It investigates how the ratio of native to borrowed morphs evolves over time and which source languages are most strongly represented among the borrowed morphs.

2. Background Theory

One of the central tasks of morphology is to investigate the internal structure of words, often by segmenting them into the smallest meaning-bearing units, known as **morphs**.[1] For example, the English word *disrespectfulness* can be decomposed into *dis-* + *re-* + *spect* + *-ful* + *-ness*. Each morph carries lexical or grammatical meaning.

2.1. Basic Terminology

The distinction between *morpheme* and *morph* is notoriously vague (Haspelmath, 2020). Haspelmath observes that the term *morpheme* is used in several conflicting senses, ranging from concrete minimal forms to abstract grammatical units. When we work with segmented word forms, we refer to concrete realizations such as *-s* or *-es*, rather than abstract categories like *plural*. In that case, we use the term *morph* in the sense defined by Haspelmath (2020): a morph is a minimal linguistic form.

The **root** is the core morph of a word and carries its main lexical meaning. Morphs that must attach to a root are called **affixes**, which modify either lexical meaning or grammatical properties. **Prefixes** attach before the root, while **suffixes** follow it. An **ending** is a specific type of suffix that marks grammatical features such as tense, person, or number. An **interfix** appears between roots.

[1] In current theoretical debates, word-based approaches to morphology are gaining prominence (Blevins, 2016). These approaches treat words as the primary units of morphological analysis, focusing on relations between whole words, and regard morphs as secondary abstractions derived from them. However, we believe that modeling the origins of individual morphs, as pursued in this paper, is compatible with viewing morphs as secondary abstractions.

Affixes are commonly classified as either **inflectional** or **derivational**. **Inflection** modifies a word's grammatical form without changing its lexical meaning, whereas **derivation** creates a new lexeme, often with a different meaning or part of speech (Aronoff and Fudeman, 2011).

2.2. Etymology

Etymology studies how words originate and evolve over time, and whether they are inherited or borrowed from other languages. Morph-level etymology examines this at the level of individual morphs, offering finer-grained insights, especially for hybrid words whose parts come from different sources.

In much of the language contact literature, the common assumption is that affixes are usually borrowed only indirectly. That is, words containing a given affix are first borrowed as wholes from another language, and only later are these words reanalysed, and the recurring affix generalized to native stems.

For example, the suffix *-ism* (from Greek, via Latin and French) entered English through borrowed words such as *baptism*, *criticism*, and *organism*, and was subsequently extended to native bases in formations like *snobbism* or *veganism*.

There are, however, also cases where affixes have been borrowed directly, without the mediation of complex loanwords. Although such cases are rare (Seifart, 2015), examining etymology at the level of individual morphs can still provide valuable insights, even when most affixes have entered a language through indirect borrowing.

Borrowing can also occur through a chain of intermediary languages. For example, the English word *school* entered Middle English as *scole* from Old English *scol*, from Proto-West Germanic **skolu*, from Late Latin *schola*, from Ancient Greek *skhole*, and probably from Proto-Indo-European (Oxford English Dictionary, 2025).

We should note that our assumption that historical pathways of words can be captured rigorously as plain sequences of languages, is an operational simplification. The etymology of a word is not always straightforward; due to various reasons, it can be difficult – or even impossible – to determine the exact trajectory by which it entered the language or through which intermediate languages it passed.

3. Related Work

To our knowledge, no computational linguistics research specifically addresses identifying the etymological origin of individual morphs on a larger scale. While there is some research on affix borrowing and a few tools exist for determining the etymology of entire words, no approach combines these

perspectives at the morph-level.

3.1. Linguistic Research on Morph Borrowing

Linguistic research has examined how morphology spreads across languages through language contact and how such processes can be distinguished from ordinary inheritance. Johanson and Robbeets (2013) draw a key distinction between shared affixes that are *copies*—replicated through contact—and those that are *cognates* inherited from a common ancestor, framing borrowing as an active process of *code-copying*. Seifart (2015) further distinguishes between *direct* borrowing, based on speakers' knowledge of the donor language, and *indirect* borrowing, where affixes spread through complex loanwords that later become productive in the recipient language.

Gardani (2008) shows that even inflectional morphs—traditionally viewed as resistant to borrowing—can transfer when language contact is intense and structural compatibility is high. Broader perspectives are provided by Arkadiev et al. (2015), which focuses on the borrowing of affixes and shows that, although less common than lexical borrowing, derivational affixes and some basic grammatical endings can also be transferred between languages. Together, these works form a theoretical basis for our approach to studying morph-level etymology computationally.

3.2. Computational Etymology

Building on this linguistic background, recent computational studies have begun to model etymological relationships automatically, though still mostly at the word level. Wu and Yarowsky (2020) predict both the type of etymological relationship (e.g., borrowing or inheritance) and the parent language from which a word originated. Their approach shows that computational modelling of etymology is both feasible and informative, though it does not address the internal morphological structure of words.

The work by Kovácsová (2025) explores methods for training models to automatically identify loanwords across multiple languages. Our research extends this line of inquiry to the morph-level, providing a finer-grained view of how native and borrowed elements combine within words. It also advances beyond binary or single-language classification by reconstructing the complete sequence of etymological origins.

3.3. Resources with Morph-Level Etymological Information

Despite the growing interest in computational etymology, few resources provide morph-level annotation. Wiktionary (2025) includes occasional etymologies for affixes and bound morphs, but coverage is sparse and inconsistent. Similarly, the *Czech Etymological Dictionary* (Rejzek, 2019) contains etymological information for several dozen affixes but focuses primarily on whole words.

Another relevant source is the open dataset by Goldberg (2019), which lists morphs with basic linguistic information. Around 1,500 entries include etymological notes—mostly for technical affixes such as *acet-*, *agri-*, *exo-*, or *-ant*—derived mainly from Greek or Latin. However, the dataset lacks detailed source references and provides only minimal etymological depth.

Overall, these resources offer isolated examples of morph-level etymology but lack the consistency and structure required for systematic computational modelling. Our work aims to fill this gap by developing a comprehensive, linguistically grounded dataset of Czech morphs with etymological labels.

4. Data

This section describes the linguistic resources, annotation process, and datasets used for training and evaluating our models.

4.1. Czech Etymological Dictionary

The latest Czech Etymological Dictionary (Rejzek, 2019) contains over 11,000 main entries and around 21,000 derived forms.

Its digital version, the Czech Etymological Lexicon (**CzEtyL**) (Rejzek et al., 2025) reformats this information into a structured, machine-readable format. Each entry specifies whether a word is native or borrowed and, for borrowed words, lists the sequence of languages through which it passed before entering Czech.

Example entry:

```
architekt   deu,lat,ell Loan
```

The word *architekt* ("architect") comes from Greek, entering Czech via Latin and German.

We extended CzEtyL using derivational relations from DeriNet (Olbrich et al., 2025), a large Czech lexical network linking derived words with their bases. For each CzEtyL entry, we added all lexemes from the same DeriNet tree and assigned them the same etymological label, expanding the coverage from about 10,500 to over 500,000 annotated words.

4.2. Our Annotation

We used the Czech part of the morpheme segmentation dataset from (SIGMORPHON, 2022), which

provides sentences segmented into morphs. Each morph was automatically classified as *root*, *derivational*, or *inflectional affix* using the model of John (2024), and then manually annotated with its sequence of language origins based mainly on Rejzek (2019) and Wiktionary (2025).

Example: *Šetřete peníze netelefonujte faxujte.*
Translation: *Save money, don't call, fax instead.*

- **Šetřete** (*save*)
  - *Šetř* — `R` — `ces`
  - *e* — `D` — `ces`
  - *te* — `I` — `ces`
- **peníze** (*money*)
  - *peníz* — `R` — `deu,lat`
  - *e* — `I` — `ces`
- **natelefonujte** (*don't call*)
  - *ne* — `D` — `ces`
  - *tele* — `R` — `ell`
  - *fon* — `R` — `ell`
  - *uj* — `D` — `ces`
  - *te* — `I` — `ces`
- **faxujte** (*fax*)
  - *fax* — `R` — `eng,lat`
  - *uj* — `D` — `ces`
  - *te* — `I` — `ces`

The annotation distinguishes three morph types: roots (`R`), derivational (`D`), and inflectional affixes (`I`). Language origins use ISO 639-3 codes such as `ces` (Czech), `deu` (German), `lat` (Latin), and `ell` (Greek).

The dataset was annotated by two native Czech speakers: a university student with a Bachelor's background in NLP and etymology, and an IT university student without prior formal linguistic training. Both annotators received detailed guidelines and a short calibration on example sentences, as well as access to resources such as the Czech Etymological Dictionary and Wiktionary (2025). Although neither is a professional linguist in this field, their work provides a useful reference point for human performance on this task.

Cohen's kappa was 0.82, and the exact match rate reached 95.9 %, indicating strong agreement. Remaining differences mostly stemmed from varying annotation granularity (e.g., `gmh` vs. `deu`) or inconsistencies in etymological sources.

4.3. Training and Evaluation Data

The dataset was then split into training, development, and test sets. Table 1 shows their sizes in terms of sentences, words, and morphs.

Dataset	Sentences	Words	Morphs
Train	200	2,774	7,016
Dev	50	599	1,460
Test	50	609	1,485

Table 1: Size of the annotated dataset used for training, development, and testing

We drew the training set from the Czech SIGMORPHON (2022) Shared Task training data (200 sentences). We derived the development and test sets from the 500-sentence development file by selecting every 10th sentence for testing and an every 10th sentence for development (offset by 1), leaving the remaining sentences unused for possible future work. Dataset size is limited by the cost of manual annotation, which requires linguistic expertise or substantial background research.

The training data contains about 7,200 annotated morphs, but most fall into just a few common etymological classes (Table 2). The etymology column lists language codes using the ISO 639-3 standard. As expected, the majority of morphs are of Czech (native) origin. Since the training set covers only 200 sentences, it should be seen as a small sample and not a fully representative picture of Czech in general.

Rank	Etymology	Count	% of total
1	ces	6,229	87.7
2	lat	280	3.9
3	ell	100	1.4
4	deu,lat	84	1.2
5	eng,lat	41	0.6
6	lat,ell	38	0.5
7	ita,deu	38	0.5
8	eng	36	0.5
9	deu,lat,ell	29	0.4
10	fra,lat	26	0.4
–	**Rest**	299	4.2

Table 2: Top 10 most frequent etymological sequences in the training data.

5. Proposed approaches to Morph-Level Etymology Prediction

5.1. Baselines

To evaluate performance, we define several baselines to establish a lower bound on expected results and assess how much the model improves beyond simple heuristics. These include always predicting Czech, a memorization approach, methods using *CzEtyL*, and a large language model baseline.

Always-native This baseline predicts all morphs as native (ces). It serves as a simple reference point and performs well when native morphs dominate the dataset.

Memorization We store all training morphs and predict their most frequent etymology label. Unseen morphs default to the majority class (ces). This baseline fails on ambiguous forms whose origin depends on context. For example, the suffix *-um* in *muzeum* ('museum') is of Latin origin, while the *um* in *rozum* ('reason, mind') is native Czech, derived from the verb *umět* ('to be able to').

Word lemmatization Each word is lemmatized with *MorphoDiTa* (Straková et al., 2014) and the lemma is looked up in the extended *CzEtyL* (4.1). The retrieved etymology is assigned to the root morph(s). Non-root morphs are matched against a *CzEtyL*-derived list of borrowed affixes; unseen affixes and roots default to ces, and inflectional endings are always predicted as ces.

5.2. Feature-based Classification

To predict the etymological origin of each morph, we extract features from annotated data that capture the characteristics of the morph and its context. These features serve as input to the classification model. We use character n-grams, morphological classification, position within the word, and embeddings, among others. They serve as input to the classification models used in our experiments.

We evaluate the following standard classifiers,[2] Logistic Regression (LR), Support Vector Machines (SVM), and Multi-Layer Perceptrons (MLP).

Character n-grams Character n-grams capture language-specific letter patterns and help the model associate certain combinations with particular etymological origins. We use 1-grams and 2-grams, which are appropriate given the average morph length of about 2.2 characters.

Morph Types Each morph is categorized as *Root*, *Derivational affix*, or *Inflectional affix*, encoded as a one-hot vector. Positional classification is also used, labelling morphs as *Root*, *Prefix*, *Suffix*, or *Interfix* depending on their position relative to the root(s) in a word. This classification is obtained using a model from John (2024).

Vowel Start and End This feature encodes whether a morph starts or ends with a vowel, using two binary values. Some languages favour vowel-initial or vowel-final patterns, so even such a simple feature can capture useful language-specific tendencies.

It becomes especially informative when combined with morph type. Many Proto-Indo-European roots, for example, follow a consonant-vowel-consonant (CVC) structure (Gamkrelidze and Ivanov, 1995). Affixes, by contrast, often attach more smoothly when starting or ending with a vowel. As a result, prefixes frequently end in vowels, while suffixes often begin with them.

Impact of Embeddings We use FastText embeddings (Bojanowski et al., 2017), which represent each word as the sum of its character n-gram embeddings. This method enables generating vectors even for morphs that are not stand-alone words. We use Czech FastText embeddings trained on Wikipedia and Common Crawl data (Grave et al., 2018).

We experiment with two variants: using embeddings directly for individual morphs, and using embeddings of the entire words in which the morphs occur.

5.2.1. Representation of Target Classes

We model target classes using two strategies:

- **Whole-sequence classification:** Treats each language sequence as a single label, predicting only sequences seen in training.
- **Multi-label classification:** Treats each language independently, allowing prediction of unseen combinations by training a binary classifier per language.

5.2.2. Data Augmentation

Beyond the original training set, we experiment with expanding the training material using two additional sources: data from the CzEtyL lexicon and automatically self-labeled data produced by our own model.

Use of Data from CzEtyL We extracted a dictionary of roots and affixes from *CzEtyL* with their probable etymological origins to serve as additional training data for the supervised models. The extracted root and affix dictionaries together contain roughly 12,000 morphs, covering about 400 distinct language sequences and 67 individual languages. These dictionaries are not perfect, as the root dictionary was constructed by assigning each root the most frequent etymology found among the words containing it, which can occasionally introduce errors.

[2] We use the implementations from https://scikit-learn.org.

Self-Training Self-training is a semi-supervised approach where a model is first trained on a manually annotated dataset, then used to label a larger unlabelled corpus. The newly labelled examples are added to the training set to improve the model. This is useful when additional annotated data is costly or difficult to obtain.

In this work, we applied self-training using a portion of the Prague Dependency Treebank (PDT) (Hajič et al., 2020). The selected subset contains about 88,000 sentences and approximately one million running words. Before processing, the text was normalized by lowercasing and removing punctuation, numbers, and special characters.

5.3. Using a Large Language Model

We evaluated several LLMs in a prompt-only setup for morph-level etymology annotation, using the same instruction format across models. The models were asked to label Czech morphs with ISO 639-3 origin codes, including any intermediary donor languages, while preserving the input structure and morph-type tags. We tested `GPT-5.2`, `GPT-5-mini`, `GPT-4o-mini`, `Claude-Opus-4.5`, `Gemini-3-Pro`, and `Deepseek-Chat-v3.1`.

We tested multiple prompt variants and selected the one that produced the most consistent, automatically parseable outputs. The final instruction prompt defines the required output format, annotation rules, and the meaning of ISO 639-3 language codes and morph-type tags.

All models were queried via the OpenRouter API[3] in batches of roughly 350 morphs (rounded to full sentences, typically $\sim$10 sentences). Each batch was processed in a fresh chat/session, i.e., no context beyond the instruction prompt and the current batch was present. We varied the amount of in-context training material from zero-shot up to the full training set (0, 50, 200, 2,000, and $\sim$7,000 training morphs), and report results under these settings.

The full prompt, API scripts, and data are released with the code (link hidden for anonymous review). We used temperature $= 0$ and otherwise default API settings to support reproducibility.

6. Evaluation Methodology

The task is to predict, for each morph, the sequence of languages through which it entered Czech, including any intermediate stages of borrowing. Although this sequence represents a historical progression, we treat and evaluate it as an unordered set for modelling purposes. The chronological order can later be reconstructed using relatively reliable rules based on known contact periods between the respective languages.

We use the F1-score to evaluate prediction quality. It is computed for each morph occurrence and then averaged across all morphs. To account for class imbalance between native and borrowed morphs, we also compute separate F1-scores for these two groups (**Bor.**, **Nat.**). It provides a clearer view of model performance. Borrowed morphs are generally harder to classify due to diverse sources and complex borrowing paths.

Many morphs appear multiple times in the dataset, especially inflectional endings. In the training set, there are 7,205 total morph instances, with 972 distinct morphs. The top 10 morphs account for 1,965 occurrences (27 %); the top 20 cover 2,864 (40 %); and the top 50 make up 3,997 occurrences—over half the dataset.

To reduce bias from frequent morphs, we also report an additional metric (**Dst.**): for each distinct morph, we first compute its average F1-score across all its occurrences and then take the mean of these per-morph scores to obtain the final value.

We also report the relative error reduction (**RER**) of the F1 score over the all-native baseline to better illustrate each model's improvement.

7. Results

We use the development set for model selection and report final numbers on the held-out test set. For the supervised models, we tuned hyperparameters with a grid search over feature and classifier settings. For prompt-only LLMs, we compared a small set of prompting configurations, varying the prompt itself, in-context training size and batch size.

7.1. Learning models: base results

Table 3 reports development-set performance of the supervised classifiers. All models use *single-label* prediction, treating each full etymological sequence as one class. The SVM uses an RBF kernel, and in the MLP model names the number indicates the hidden-layer size (number of neurons).

Both logistic regression and SVM struggled on borrowed morphs, although SVM achieved the best score on native ones. All MLP variants performed better overall, and results were very similar across hidden-layer sizes. In our experiments, adding extra hidden layers did not yield a consistent improvement.

7.2. Embeddings and Multi-label Setup

We evaluated several ways of integrating embeddings into the model, using representations of the

[3]We enabled OpenRouter's "do not train on my data" setting for all requests.

Model	F1	RER	Nat.	Bor.	Dst.
LogReg	94.0	39.7	98.7	51.4	86.2
SVM	94.7	46.5	**99.5**	50.7	88.1
MLP30	95.5	54.5	98.9	64.1	90.1
MLP100	**95.6**	**55.5**	99.0	**64.4**	**90.4**
MLP300	95.5	54.3	99.3	60.5	90.2

Table 3: Performance on the development set. RER = relative error reduction over the All-native baseline; Nat. / Bor. = F1 on native / borrowed morphs; Dst. = average F1 over distinct morphs.

morph itself, the word in which it appears, and their combination. The experiments were run across multiple model configurations. Overall, results were very similar across these settings, suggesting that embeddings did not lead to a substantial improvement. Table 4 reports representative results for MLP30 on the development set.

We also tested a multi-label setup, where each language is predicted independently. Instead of selecting a single etymological sequence, the model predicts which languages belong to the etymological path of a given morph. This increases flexibility and allows combinations not seen in training. In practice, the setup trains one small classifier per language, effectively increasing total capacity, which is particularly useful when additional training data is available.

In Table 4, we compare multi-label models trained on the base data with models extended by roots and affixes from *CzEtyL*. In the model names, **M** denotes the multi-label setup and **C** marks the inclusion of *CzEtyL* data. Multi-label prediction alone did not improve performance on the base training set, but it brought clear gains when combined with *CzEtyL*-derived data. Increasing the hidden-layer size provided no additional benefit.

7.3. Prompt-only LLM results on the development set

Table 5 summarizes prompt-only development-set results with the full in-context training set. A consistent pattern across models is that native-morph performance is near ceiling, so most differences come from how well the models handle borrowed morphs.

`GPT-4o-mini` stays close to the all-native baseline because it rarely identifies borrowed morphs correctly, even when given many in-context examples. A likely reason is that this smaller general-purpose model has limited capacity for the fine-grained decisions needed for short, often ambiguous morph strings, so it tends to fall back to the dominant native label.

Most models benefit from additional in-context training material. For example, `GPT-5.2` improves from 94.7 % F1 in the zero-shot setting to 95.5 % (50), 96.0 % (200), 96.3 % (2,000), and 96.8 % with the full training set (∼7,000 morphs).

In contrast, `Gemini-3-Pro` is already strong in the zero-shot setting (95.9 % F1) and shows no clear improvement with added context, reaching 95.8 % F1 with the full training set.

Setting	F1	RER	Nat.	Bor.	Dst.
Effect of embeddings - MLP30 model					
No Embedd.	95.5	54.5	**98.9**	64.1	90.1
Word	95.0	49.5	98.7	61.2	89.2
Morph	95.1	50.8	**98.9**	60.4	89.4
Word+Morph	95.3	52.5	98.7	64.2	89.4
Multi-label MLPs					
Without CzEtyL data					
MLP30 M	95.3	52.8	98.8	64.1	89.7
MLP100 M	95.1	51.1	98.8	61.6	89.3
MLP300 M	95.0	49.8	**98.9**	59.6	89.1
With CzEtyL data					
MLP30 MC	**95.8**	**58.2**	98.8	69.5	**90.9**
MLP100 MC	95.6	56.1	**98.9**	66.0	90.7
MLP300 MC	**95.8**	57.3	98.4	**71.6**	90.6

Table 4: Development-set results: (i) effect of embeddings for MLP30, and (ii) multi-label MLP variants with and without additional CzEtyL-derived training data.

Model	F1	RER	Nat.	Bor.	Dst.
GPT-4o-mini	90.4	3.2	**99.9**	3.9	80.1
GPT-5-mini	96.4	63.6	99.6	67.0	92.3
GPT-5.2	96.8	67.4	99.1	**75.7**	93.5
Cld.-Opus4.5	**97.0**	**69.4**	99.4	74.9	**93.6**
Gemini-3pro	95.8	58.2	99.1	66.2	91.1
DeepSk-v3.1	95.6	55.8	99.5	60.6	90.8

Table 5: Development-set prompt-only LLM results with the full in-context training set (∼7,000 morphs).

7.4. Results on Test Set

We now evaluate all models on the test set, including the baselines, learning-based approaches, and the prompt-only LLM setup with the full training set provided in-context. Table 6 summarizes the results.

7.5. Discussion of Results

Prompt-only LLMs With the full training set provided in-context, the prompt-only LLMs reach strong performance overall (Table 6). `Claude-Opus-4.5` is the top performer in this setting, closely followed by `GPT-5.2`. `GPT-5-mini` and the other non-OpenAI models (`Gemini-3-Pro`, `Deepseek-Chat-v3.1`) form a second tier with

Model	F1	RER	Nat.	Bor.	Dst.
Baselines					
All native	89.9	–	**100.0**	0.0	77.1
Memor.	94.4	44.9	99.3	51.5	86.3
Lemmatiz.	94.8	48.1	98.6	60.7	90.0
Learning models					
MLP30	95.2	51.6	98.6	65.2	88.3
MLP100 M	95.8	57.4	99.1	66.2	90.1
MLP300 MC	96.2	62.0	98.9	72.7	90.8
MLP30MC	96.5	65.6	99.6	69.6	91.5
Self-train	96.1	60.8	98.6	72.8	91.0
Prompt-only LLMs (full train set in-context)					
Cld.-Opus4.5	**97.8**	**77.7**	99.7	**80.4**	**94.7**
DeepSk-v3.1	96.5	65.6	99.6	69.0	91.9
Gemini-3pro	96.5	65.8	99.7	68.4	91.8
GPT-4o-mini	90.9	9.4	99.9	10.1	79.4
GPT-5-mini	96.4	64.0	99.8	66.0	92.5
GPT-5.2	97.6	76.6	99.8	78.6	94.4

Table 6: Final results on the test set. Prompt-only LLMs use the full training set in-context.

similar results. In contrast, `GPT-4o-mini` lags far behind, and the gap is driven mainly by weak performance on borrowed morphs rather than on native ones, which are near ceiling for all models.

Baselines The trivial *All native* model achieved an F1 score of 89.9 %, which also reflects the proportion of native morphs in the dataset. The lemmatization approach, which draws on etymological data from CzEtyL, cut this error by nearly half. It outperformed the simpler memorization baseline.

Learning models The trends from the development set carried over to the test set. Using *CzEtyL*-derived training data and the multi-label formulation both helped, with the best supervised variants reaching around 96.5 % F1 (e.g., MLP30MC: 96.5).

In the self-training setup (5.2.2), we used the best development-set model (MLP30MC) to label the large corpus and then retrained on these pseudo-labels; the retrained model achieved similar results (96.1 % F1), correcting some morphs but introducing new errors, so overall it brought no clear improvement over the fully supervised approach.

Performance by Morph Type Since borrowing affects morph types differently, we also break down performance by *root*, *derivational affix*, and *inflectional affix*. This follows the standard intuition behind borrowability scales: lexical material and derivational morphology tend to transfer more readily than inflectional endings, which are often more resistant in ordinary contact situations. In Czech in particular, inflectional endings are overwhelmingly native, so strong scores on inflectional morphs are expected.

Model	F1	Root	Deriv.	Infl.
All native	89.9	79.6	93.1	**100.0**
Memor.	94.4	87.7	97.0	**100.0**
Lemmatiz.	94.8	91.9	94.3	**100.0**
MLP30MC	96.2	92.1	97.6	**100.0**
GPT-4o-mini	90.7	81.3	93.5	**100.0**
GPT-5-mini	95.0	89.9	96.7	99.7
GPT-5.2	97.5	94.6	98.5	**100.0**
C-Opus4.5	**97.8**	**95.1**	**98.8**	**100.0**

Table 7: F1 by morph type on the test set. For LLMs we used the full training set in-context.

8. Conclusion and Future Work

We addressed the task of predicting the etymological origin of morphs in Czech, determining whether each morph is native or borrowed and, if borrowed, which languages are involved in its borrowing path. Because most resources focus on word-level etymology, there are currently no widely used datasets providing morph-level etymological annotation for segmented words. To fill this gap, we manually annotated ∼10,000 morphs (9,961) in 300 Czech segmented sentences (3,982 words) from the Shared Task on Morpheme Segmentation by SIGMORPHON (2022). The resulting dataset is publicly available at `https://github.com/ampapacek/MorphemeOrigin`.

We compared three families of approaches: simple rule-based baselines, supervised classifiers, and prompt-only LLM predictions. The supervised models use character n-grams together with structural cues (morph type/position and simple vowel-shape features). Among supervised systems, the strongest learning-based model was a multi-layer perceptron with a single hidden layer of 30 neurons, trained in a multi-label setup and augmented with *CzEtyL*-derived entries. It substantially outperformed the rule-based baselines. While it does not match the very best prompt-only LLM setting, it remains attractive as a lightweight, deterministic, and inexpensive alternative that can be run locally without external API access or heavy computational resources. Semi-supervised self-training on the PDT corpus yielded no clear gains.

Overall, the best results come from prompt-only LLMs when the full training set is provided in-context. In this setting, `Claude-Opus-4.5` reaches 97.8 % F1. Across LLMs, adding more in-context training morphs generally improves performance.

Future work A straightforward next step is to enlarge the annotated dataset, since more examples expose models to a broader range of morphs and borrowing patterns. We also observed that scaling supervision helps: increasing in-context training improves LLM performance, and adding more labeled data benefits the supervised models as well. The annotation could be made more fine-grained, for example by distinguishing historical stages of Latin, German, or Greek, or by marking whether a morph is pan-Slavic or specific to Czech.

For learning models, richer shape-based features may help. For example, the vowel-start/end feature was useful, and encoding the full consonant–vowel (CV) pattern of each morph could capture additional regularities. On the model side, more structured architectures that predict language sequences directly, as well as ensembles combining complementary systems, could improve robustness. Another direction is to revisit semi-supervised methods such as self-training, which did not yield clear improvements in our current setup. Finally, it would be interesting to adapt pretrained models to this task (e.g., by fine-tuning) and to more systematically explore training and inference settings, including hyperparameters and prompting choices.

Although this study focused on Czech, the methodology is not language-specific and could be applied to other morphologically rich languages.

9. Acknowledgements

This work has been supported by the project TQ12000040 (CZDEMOS4AI) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Program, by the Charles University Research Centre program No. 24/SSH/009, and by the Czech Science Foundation project No. 26-21822S. This work has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062). We also thank anonymous reviewers for their useful comments.

10. Bibliographical References

Peter Arkadiev, Francesco Gardani, and Nino Amiridze. 2015. *Borrowed Morphology: An Overview*, pages 1–23. De Gruyter Mouton.

Mark Aronoff and Kirsten Fudeman. 2011. *What is Morphology?* Fundamentals of Linguistics. Wiley.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Philip Durkin. 2014. *Borrowed Words: A History of Loanwords in English*. Oxford University Press.

Thomas V. Gamkrelidze and Vjaceslav V. Ivanov. 1995. Chapter Four: The Structure of the Indo-European Root. In *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture*, pages 185–187. De Gruyter Mouton, Berlin, Boston.

Francesco Gardani. 2008. *Borrowing of Inflectional Morphemes in Language Contact*. Phd thesis, University of Vienna.

Anthony P. Grant. 2020. *The Oxford Handbook of Language Contact*. Oxford University Press.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134.

Lars Johanson and Martine Robbeets. 2013. Copies versus cognates in bound morphology. *STUF – Language Typology and Universals*, 66(2):130–135.

Vojtěch John. 2024. Morph classifier. Master's thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL), Prague, Czech Republic.

Viktória Kovácsová. 2025. Automatic detection of lexical borrowings. Master's thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL).

Oxford English Dictionary. 2025. school, n. https://www.oed.com/dictionary/school_n1. Oxford University Press, accessed 13 Oct 2025.

Jiří Rejzek. 2019. *Český etymologický slovník*. 3rd edition. LEDA.

Frank Seifart. 2015. Direct and indirect affix borrowing. *Language*, 91(3):511–532.

Wiktionary. 2025. Wiktionary: The free dictionary. https://www.wiktionary.org/. Accessed: 2025-10-04.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

11. Language Resource References

Colin Goldberg. 2019. Morphemes dataset. https://github.com/colingoldberg/morphemes. Accessed: 2025-10-13.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank – Consolidated 1.0. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France.

Michal Olbrich, Viktória Brezinová, Šárka Dohnalová, Vojtěch John, Lukáš Kyjánek, Aleš Papáček, Emil Svoboda, Magda Ševčíková, Jonáš Vidra, and Zdeněk Žabokrtský. 2025. DeriNet 2.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jiří Rejzek, Aleš Papáček, Viktória Brezinová, and Zdeněk Žabokrtský. 2025. Czech Etymological Lexicon 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

SIGMORPHON. 2022. SIGMORPHON 2022 Shared Task on Morpheme Segmentation. https://github.com/sigmorphon/2022SegmentationST. Accessed: 2025-10-04.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

A. Etymological Analysis of a Diachronic Czech Legal Corpus

In this appendix, we apply the best lightweight supervised system described in the main paper to a diachronic corpus of Czech legal texts and examine how the predicted proportions of native and borrowed morphs change over time. The goal of this analysis is not to provide a definitive historical account, but rather to illustrate one possible downstream use of the proposed morph-level etymology framework on a long real-world time series.

The corpus was constructed from publicly available Czech legislative texts obtained through the e-Sbírka API. We downloaded texts year by year and constructed one sampled file for each year by randomly sampling sentences up to approximately 1,000 words. The resulting dataset contains 109,257 words. The corpus was created within the CZDEMOS4AI project. Because it uses a fixed yearly word budget and includes only the texts available at the time of collection, it should be interpreted as a sample of legal language in each year rather than as a complete picture of legal language over the full period.

We then automatically segmented the corpus and applied the classifier from the main paper to the resulting morphs. For this analysis, we used the same MLP configuration as in our best lightweight system: a single hidden layer of size 30, multi-label prediction, and extended training data derived from the etymological dictionary. On the original test set in the main paper, this model achieves 96.2% F1.

Several caveats should be kept in mind. First, the legal corpus belongs to a different domain than the data used for evaluation in the main paper, so the model's accuracy on legal texts may be lower. Second, both the morphological segmentation and classification were produced automatically, which means that some errors may already arise in the preprocessing stage. The figures and observations presented below should therefore be interpreted as exploratory rather than definitive.

In Table 8, we summarize the main predicted origin labels, including the most frequent borrowed source languages. The values show what percentage of morphs is assigned to each label. Native is the dominant label overall, while Latin is the dominant borrowed source in every available year of

Origin	Mean	Median	Max	Std. dev.
Native	88.12	88.22	93.03	2.69
Latin	5.58	5.29	9.97	1.49
English	1.67	1.62	3.21	0.41
German	1.41	1.31	2.54	0.47
French	1.14	1.08	2.79	0.44
Greek	1.10	1.00	3.73	0.55
Polish	0.33	0.30	1.61	0.19
Italian	0.22	0.17	0.88	0.18

Table 8: Summary of the main predicted origin labels across the years 1918–2026. Shares are computed with respect to all predicted labels within each year, and standard deviation is computed across yearly shares. Labels are sorted by mean yearly share. All values are in (%).

the sample. Other prominent borrowed sources include English, German, French, and Greek, but all of them remain clearly below Latin in the aggregate picture.

Across all sampled years, the mean predicted share of borrowed morphs is 8.96%. In the earliest available year, 1918, it is 6.03%, while the maximum in the current sample is observed in 2015, where it reaches 15.25%. Figure 1 shows the five-year weighted smoothing[4] of the overall borrowed share together with fitted linear trend lines for the overall borrowed share and for the three most prominent borrowed source languages. Overall, the fitted trend line suggests a slight upward trend in the proportion of borrowed morphs over time, corresponding to roughly 2.5 percentage points per century.

The code, data, and generated outputs for this appendix are available at: https://github.com/ampapacek/MorphemeOrigin/tree/main/legal_corpus_analysis.

[4] Each year's value is replaced by a weighted average over a five-year window centered on that year, using weights 0.1, 0.2, 0.4, 0.2, and 0.1.

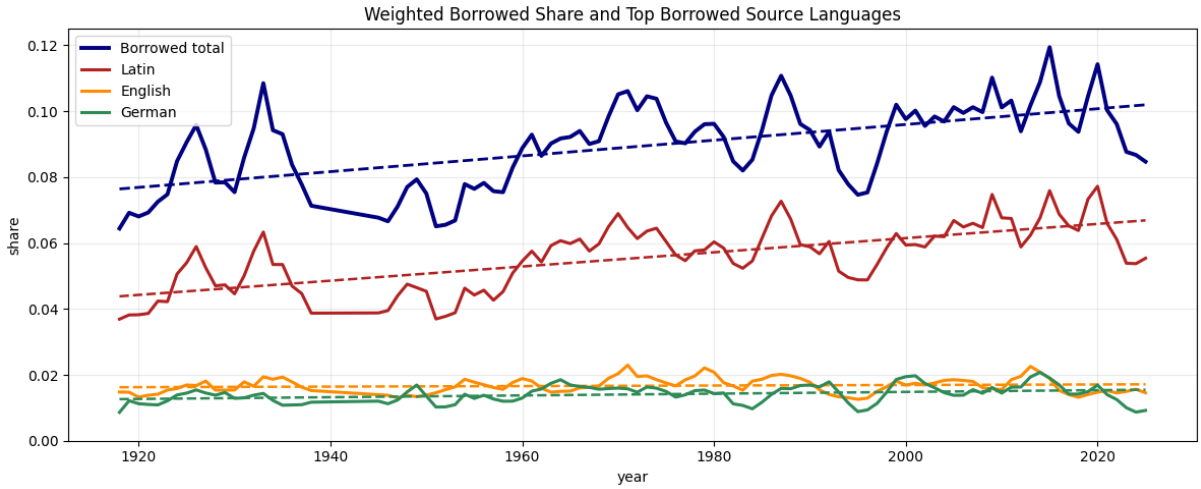

Figure 1: Five-year weighted smoothing of the overall predicted borrowed-morph share together with fitted linear trend lines for the overall borrowed share and for the three most prominent borrowed source languages: Latin, English, and German.