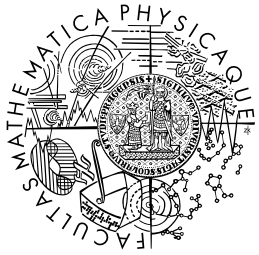

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Aleš Manuel Papáček

Identification of Morpheme Origin

Institute of Formal and Applied Linguistics

Supervisor of the bachelor thesis: prof. Ing. Zdeněk Žabokrtský, Ph.D.

Study programme: Computer Science – Artificial Intelligence

Prague 2025

I declare that I carried out this bachelor thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication

I want to thank...

Title: Identification of Morpheme Origin

Author: Aleš Manuel Papáček

Institute: Institute of Formal and Applied Linguistics

Supervisor: prof. Ing. Zdeněk Žabokrtský, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Use the most precise, shortest sentences that state what problem the thesis addresses, how it is approached, pinpoint the exact result achieved, and describe the applications and significance of the results. Highlight anything novel that was discovered or improved by the thesis. Maximum length is 200 words, but try to fit into 120. Abstracts are often used for deciding if a reviewer will be suitable for the thesis; a well-written abstract thus increases the probability of getting a reviewer who will like the thesis.

Keywords: Etymology, Morphology

Název práce: Identifikace původu morfémů

Autor: Aleš Manuel Papáček

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: prof. Ing. Zdeněk Žabokrtský, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Abstrakt práce přeložte také do češtiny.

Klíčová slova: Etymologie, Morfologie

Contents

Introduction **7**

1 Theoretical Background **8**
1.1 Morphology and Word Structure 8
1.1.1 Morphemes and Morphs 8
1.1.2 Morph types 9
1.1.3 Problematic affixes 11
1.2 Etymology 12
1.2.1 Languages of the World 12
1.2.2 Language groups 12
1.2.3 Influence of Cultures and Language Contact 16
1.3 Etymology of Morphemes 16
1.3.1 Calques 17

2 Related Work and Data Sources **19**
2.1 Deconstructor 19
2.2 DeriNet 20
2.3 Czech Etymological Dictionary 21
2.3.1 CzEtyL 21
2.3.2 SIGMORPHON Shared Task 22
2.4 Segmentation 22
2.5 Classification 23

3 Practical Part **24**
3.1 Data 24
3.1.1 Annotations 24
3.2 Baselines 26
3.2.1 Always-Czech Baseline 27
3.2.2 Memorization-Based Approach 27
3.2.3 Word Lemmatization Approach 27
3.2.4 Morph Based Baseline 27
3.3 Evaluation Methodology 28
3.3.1 F1-Score 28
3.3.2 Native vs. Borrowed 28
3.3.3 Score Grouped by Morphs 29
3.3.4 Relative error reduction 29
3.3.5 Expected Bounds of Performance 29
3.3.6 Inter-Annotator Agreement 30
3.4 Results of inter annotator agreement 31

4 Model for Morph-Level Etymology Prediction **32**
4.1 Features 32
4.1.1 Morph n-grams 32
4.1.2 Morph Types 33
4.1.3 Vowel Start and End 33

4.1.4 Embeddings 33
4.2 Extending the Training Data with Etymological Dictionary 34
4.3 Representation of Target Classes 34
4.3.1 Label Statistics from Training Data 35
4.3.2 Extended Training Data 35
4.4 Classification Models 37
4.5 Self training 37
4.6 Documentation, Code 37

5 Experiments and Results **39**
5.1 Experiments on Development Set 39
5.2 Final Results on Test Set 39

Conclusion **40**

Bibliography **41**

List of Figures **43**

List of Tables **44**

List of Abbreviations **45**

Introduction

This is an introduction for my bachelor thesis on the identification of morpheme origin. I would like to talk about these points in this chapter:

- Overview of the research problem.
- Research objectives and key questions.
- Potential applications in NLP, lexicography, and language learning.
- Challenges in identifying morpheme origins automatically.
- Structure of the thesis.

Etymological Dictionary by J. Rejzek

I will frequently refer to the Czech Etymological Dictionary by Rejzek (2019). When discussing etymologies described in this dictionary, I will use the [R] mark to indicate the source.

Languages used

The concepts covered in this thesis apply to all natural languages. However, the experimental part will focus mainly on Czech, as it offers the best available data for this type of research. In the theoretical sections, examples will be given in English, Czech, and occasionally Spanish or German to show how the same concepts appear in different languages. These languages were chosen because they are among the most widely spoken in the world and are also languages that the author knows well.

TODOs

- Use of proper quotation marks. ” vs ” vs “ (I think it is done)
- Write Introduction chapter
- Add citations where needed
- The Final experiments and chapter

1 Theoretical Background

In this chapter, we define the key terms used in the following chapters, primarily related to morphology and etymology. We establish fundamental concepts such as morpheme, morph, morphological segmentation, and types of language borrowings. Although many papers have been written on this topic, the terminology is not always unified and may vary slightly across different publications. To ensure consistency throughout this work, we provide clear definitions of these terms as they will be used in the subsequent chapters.

1.1 Morphology and Word Structure

Morphology is a branch of linguistics that studies the form and the internal structure of words. The word morphology was first used by Goethe (1817) in a biological context, where it referred to the study of the structure of organisms. The etymology of the word morphology comes from Greek *morphē*, meaning "shape" or "form," and *-logía*, meaning "study" or "science" (derived from the verb *légō*, "to say, read, or collect"). [R]

Morphology analyzes words in a language and divides them into smaller units that carry meaning; these units together form the structure of the word. For example, the complex word *disrespectfulness* can be segmented as *dis-* + *respect* + *-ful* + *-ness*.

Each part has a defined meaning, either lexical or grammatical. The smallest indivisible units are called morphemes. Some words can be just one morpheme (e.g. *pen*)

The root *respect* could be further segmented as *re-* + *spect* based on its etymology. The word *respect* comes from Latin *re-* + *specere*, meaning "to look again."[R]However, these individual parts are no longer used productively in modern English.

1.1.1 Morphemes and Morphs

The terms morpheme and morph are often interchanged, or in some cases, people use just the term morpheme when talking about any smaller part of a word.

German linguist M. Haspelmath (2020) identifies at least three uses of the term *morpheme*. For our purposes, we will adopt the following definition:

> A morpheme is a set of minimal forms with identical syntacticosemantic content.

A morph is the concrete realization of a morpheme. To better understand the distinction, we will illustrate it with an example. In English, the plural form is created by adding *-s*, *-es*, or *-ies* to the singular noun. The morphs *-s*, *-es*, and *-ies* are all elements of the same morpheme representing the plural form (denoted as {Plural} or {-s, -es, -ies}).

Similarly, in Spanish, the plural morpheme {Plural} can be realized through different morphs, such as *-s* (e.g., *cerveza* + *-s* → *cervezas* (beers)) or *-es* (e.g., *corazón* + *-es* → *corazones* (hearts)).

This is in case of regular nouns, in both English and Spanish there are irregularities, which we will discuss later.

Likewise, in Czech, the actor morpheme {Actor} can be realized through different morphs, such as *-el* (e.g., *učit* + *-el* → *učitel* (teach + -er)) or *-ář* (e.g., *kov* + *-ář* → *kovář* (blacksmith)).

The phenomenon where a morpheme has multiple variants is called *allomorphy*, and the morphs that serve as realizations of the same morpheme are called *allomorphs*.

1.1.2 Morph types

In this section, we define the root of a word, affixes, and their types. We will understand that roots and affixes are specific types of morphs, not morphemes, as previously defined. If affixes were considered a type of morpheme, we would have, for example, just one suffix representing all actor-forming suffixes such as {-el, -ář, …} in Czech. Instead, we recognize that *-el* and *-ář* are distinct suffixes, each serving the same function but realized as different morphs.

Root

The *root* morph carries the main lexical meaning and usually can function as a standalone word. Morphs that cannot stand alone and must attach to a root are called *affixes*. These affixes can modify either the lexical meaning or grammatical properties of the word.

For example, Czech is a morphologically rich language that can form complex words by adding multiple affixes to a root. The word

*nejneob**hospod**ařovávatelnějšími*

meaning "with the most impossible to continuously manage" contains multiple affixes surrounding the root *hospod*.

Although this word is grammatically correct, it is artificially created to demonstrate word length and is not used in everyday language.

Affixes

A word may have zero or more *prefixes* preceding the root, while *suffixes* appear after it. In some languages, the final suffix, which carries grammatical meaning, is further classified as an *ending*.

We illustrate this with an example from the beginning of this chapter: the word *disrespectfulness* consists of one prefix (*dis-*), the root (*respect*), and two suffixes (*-ful* and *-ness*).

Less common affixes

In addition to prefixes and suffixes, there are other types of affixes classified based on their relative position to the root.

A *circumfix* is an affix that consists of two parts, which attach to a root from both sides. Unlike a simple combination of a prefix and a suffix, circumfixes always function as a fixed pair, meaning their individual parts cannot be used separately. A common example is in German, where the past participle is often formed with the circumfix *ge-...-t*, as in *gesagt* ("said") from *sagen* ("to say").

An *infix* is an affix inserted within the root. Infixes are relatively rare in Indo-European languages but occur in other language families, such as Austronesian. For example, in Tagalog, the infix *-um-* is used in verb formation, as in *liwanag* ("clearness") becoming *lumiwanag* ("to become clear"). (Schachter; Otanes, 1983)

A *compound* is a word that contains more than one root. The roots in a compound can be connected by *interfixes*. For example, in English, *blackboard* is a compound made up of two roots: *black* and *board*, without an interfix. In German, the word *Liebesbrief* ("love letter") is formed from the roots *Liebe* ("love") and *Brief* ("letter"), with the interfix *-s-* appearing between them.

Derivation and Inflection

A key criterion for categorizing affixes is whether they contribute to *derivation* or *inflection*. Inflection changes the grammatical form of a base word, while derivation often alters its meaning or part of speech.

The term *inflection* comes from the Latin *flectere*,[R]meaning "to bend." It applies to a base word and "bends" its shape to express the desired *morphosyntactic information*, including tense, aspect, number, and case. Inflection modifies a word's form without changing its core meaning, typically by adding an inflectional affix. However, it can also involve internal modification rather than affixation. For example, the English verb *talk* inflects to *talked* with the suffix *-ed* (past tense), while *sing* changes to *sang* (past tense) through vowel alternation, a process known as *ablaut*.

Words formed through inflection belong to the same *lexeme*, meaning they share a base meaning and only differ in grammatical form. For instance, *talk*, *talks*, *talked*, and *talking* all belong to the lexeme *talk*. Dictionaries typically list lexemes rather than their inflected forms, so a person searching for *talking* would look under *talk*.

The distinction between derivation and inflection is not always clear-cut, as some affixes can show characteristics of both. For instance, the Czech prefix *nej-*, which forms the superlative degree of adjectives, does change the meaning to some extent, but since adjective degree is a grammatical feature, it is typically considered an inflectional affix.

According to Aronoff and Fudeman (2011), the main distinction is that derivation creates new lexemes, whereas inflection produces different forms of the same lexeme, with the specific form determined by the syntactic context. For example, *walk* and *walks* (where *-s* is an inflectional affix) share the same base meaning, and the addition of *-s* depends on subject-verb agreement. In contrast, the difference between *run* and *runner* (where *-er* is a derivational affix) is greater, as the first is a verb and the second is a noun. Derivational affixes often change the part of speech, as seen in Czech verbs *bít* ("to beat") and *zabít* (*za-* + *bít*, "to kill"), where the derivational prefix *za-* significantly alters the meaning.

Another distinction recognized by Aronoff and Fudeman (2011) is that inflectional affixes are generally positioned further from the root than derivational

affixes. For example, in the English word *rationalizations*, the derivational suffixes *-al*, *-iz*, and *-ation* appear closer to the root *ration*, while the inflectional suffix *-s* is positioned at the very end.

Stem

The *stem* of a word is defined as its base form without inflectional affixes (endings). For instance, in English, the verb *walking* has the stem *walk*, which remains unchanged in different conjugations such as *walks* and *walked*, where the inflectional endings *-s* and *-ed* indicate grammatical distinctions.

The stem can sometimes be identical to the root, as in the previous example, or it can be more complex. For instance, in the word *rebuilding*, the stem is *rebuild*, while the root is simply *build*.

1.1.3 Problematic affixes

Up to this point, everything seemed "pretty" and regular. The segmentation into morphemes was clear and non-overlapping, but this is not always the case.

For example, in irregular plural forms, we cannot directly segment out the affix {plural}. The word *fish* has the same form in both singular and plural, so *fish* + {plural} results in the same word *fish*. Morphs that change the meaning but do not have an explicit form are usually called *zero morphs* (often denoted as {∅} or {0}).

Also falling under this category are morphs that cause a change in the root itself, such as *woman* + {plural} → *women*. These are called *simulfixes*.

When the entire root is replaced by another form while adding a derivational morpheme, it is called a *suppletive morph*. A well-known example is *good* + {comparative} → *better* or *go* + {past tense} → *went*.

Another type of morpheme that breaks the idea of morphs being reusable building blocks are the so-called *cranberry morphemes*. These are morphemes that appear only in one specific word and are not used anywhere else.

The name comes from the English word *cranberry*. The part *berry* is productive and appears in many words like *blueberry*, *strawberry*, etc. But the part *cran* is only used in *cranberry* and has no meaning on its own.

Another example is *cobweb*, where *web* is meaningful, but *cob* is not used in any other word in modern English.

Sometimes, when attempting to break words down into morphemes to derive new words, the morphological segmentation can be incorrect, resulting in new words with morphs that originally had no meaning. This occurs most often when analyzing loanwords from different languages.

A widely known example is the word *hamburger*, originally derived from the German city of *Hamburg*. In English, it was reanalyzed as *ham* + *burger*, which led to the creation of new words such as *cheeseburger*, *chickenburger*, and others.

Another example is *alcoholic*, which should be segmented as *alcohol* + *-ic*. The word *alcohol* comes from Arabic *al-kuḥl* (where *al-* is just an article). Through Latin, it entered French and later English.[R]

However, the segment *-holic* was misinterpreted as an independent morpheme and was later used productively to create new words such as *workaholic*, *chocoholic*, and *shopaholic*.

1.2 Etymology

Etymology is the study of the origin and historical development of words. It investigates how words were formed, how their meanings and forms have changed over time, and whether they are native or borrowed from other languages.

The word *etymology* comes from Latin *etymologia*, which itself originates from Greek *etymología*, derived from *étymos* (meaning "correct" or "truthful") and *-logia* ("study" or "science"). The term was already used by ancient Greek philosophers in discussions about whether words truthfully describe the meaning of the things they denote.[R]

In the following subsections, we will take a closer look at the languages of the world and their evolution, which will help us better understand how languages influence one another.

1.2.1 Languages of the World

The source of information for this subsection is primarily from the chapter on this topic in the *Czech Etymological Dictionary* by Rejzek (2019).

There are an estimated 3,000 to 7,000 languages spoken worldwide. The wide range in this estimate exists because the distinction between a language and a dialect is often unclear.

Some languages are so similar that speakers of these different languages can understand each other, even though they are classified as separate languages. This is often due to political, historical, or cultural reasons. For example, Serbian, Croatian, and Bosnian are very similar, but they are considered separate languages mainly for political reasons.

On the other hand, some dialects of a single language can be so different that speakers have difficulty communicating with each other. A well-known example is Mandarin and Cantonese, which are both considered dialects of Chinese but differ significantly in spoken form.

Additionally, the exact number of languages cannot be precisely determined, as there are still remote regions of the world—such as the Amazon rainforest, parts of Africa, and isolated Pacific islands—that contain undocumented or barely studied languages.

To better understand the relationships and similarities between languages, linguists classify them into language families—groups of languages that evolved from a common ancestor. By studying these families, we can better understand how modern languages have evolved over time.

1.2.2 Language groups

There are around 146 recognized language families, this number depends on the granularity with which we divide languages into family groups. The largest among them are (Eberhard et al., 2025):

- **Niger-Congo** (1,537 languages, ~612 million speakers)
- **Austronesian** (1,225 languages, ~328 million speakers)
- **Trans-New Guinea** (476 languages, ~3.8 million speakers)

- **Sino-Tibetan** (457 languages, ~1.4 billion speakers)
- **Indo-European** (446 languages, ~3.3 billion speakers)
- **Afro-Asiatic** (377 languages, ~633 million speakers)
- **Other language groups** (2,646 languages, ~1.1 billion speakers)

We will focus more on the Indo-European language family group.

Indo-European

The Indo-European language family is a group of languages that are believed to have evolved from a common ancestor, *Proto-Indo-European*. The earliest speakers of this language probably lived between approximately 4000 and 3000 BCE in what is now Ukraine and neighboring regions.

This language group further branches into many subgroups, including Italic, Germanic, Slavic, Hellenic, Anatolian, Baltic, Celtic, Tocharian, Indo-Iranian, and others. (Olander, 2022)

I will explore the branches that have had the greatest impact on modern European languages and have most influenced the Czech and English languages.

- **Italic**

  From the Italic language group, Latin had the greatest impact on other languages. The Romance languages evolved from Latin and are traditionally divided into two main groups:[1]

  - **Western Romance**: Spanish, French, Portuguese, Catalan, Italian
  - **Eastern Romance**: Romanian, Dalmatian[2]

- **Slavic**

  The Slavic branch is further divided into three groups:

  - **West Slavic**: Czech, Slovak, and Polish.
  - **East Slavic**: Russian, Ukrainian, and Belarusian.
  - **South Slavic**: Serbian, Croatian, Bulgarian, Slovenian, and Macedonian.

- **Germanic**

  The Germanic branch is further divided into three groups:

  - **West Germanic**: German, English, Dutch, Afrikaans, Frisian, Yiddish
  - **North Germanic**: Swedish, Danish, Norwegian, Icelandic
  - **East Germanic**: Gothic (extinct)

- **Hellenic** Mainly Greek

[1] Lists are not exhaustive

[2] Dalmatian became extinct in the 19th century. It was spoken along the coast of present-day Croatia.

Italic language group

The most influential language from this group is Latin. Originally, Latin was spoken in a small region around Rome. The oldest known inscriptions date back to the seventh - fifth centuries BCE.(Rejzek, 2019)

Alongside Classical Latin, a spoken simpler variety known as Vulgar Latin developed, which later became the foundation for the Romance languages.

Latin is no longer a natively spoken language, but it is still widely used in fields such as medicine, law, and science. Additionally, Latin remains one of the formal languages of the Vatican and is still used in the Roman Catholic Church.

Latin had a profound impact on most European languages due to the expansion of the Roman Empire and the spread of Christianity across the continent. Although many languages were not influenced by Latin directly, in many cases, this influence came through an intermediate language. Examples include the influence on English through French and on Czech through German.

Evolution of Czech language

Czech belongs to the *West Slavic branch*. Over time, it gradually began to separate and develop distinct characteristics. Significant linguistic changes occurred in the 10th century, and by the beginning of the second millennium, we can begin to refer to the emerging language as *Proto-Czech*. (Kosek, 2017)

- **11th–12th century: Proto-Czech** - From this period there are not any written documents.
- **12th–15th century: Old Czech** - The first complex Czech texts appear in this period, with significant literary expansion in the 14th century.
- **16th–18th century: Middle Czech - 16th–early 17th century: Humanist Era** – A period of language refinement, as Czech scholars tried to make the language more elegant, following the model of Latin.

  - The form of Czech from this time was later used by J. Dobrovský as the basis for written Czech during the National Revival, which created lasting differences between spoken and written Czech.(Rejzek, 2019)

  - **Mid-17th–18th century: Baroque Era** – A time of decline for the Czech language. Due to political and historical events, German became the dominant language in administration and education, while Czech was spoken mostly informally and on the countryside.

- **Late 18th century: New Czech** - Developed as a reaction to the decline of Czech in previous centuries. - During the *Czech National Revival*, scholars and writers worked to standardize and revive the language.

Germanic Language Group

This section is primarily based on the chapter *Germanic Languages* from the Czech Etymological Dictionary. Rejzek (2019) Only relevant parts were selected, translated from Czech, and slightly rephrased.

The oldest Germanic written records are runic inscriptions dating from the second to the sixth century.

This section focuses on the West Germanic branch, which includes German and English, among other languages. These two are particularly relevant for this discussion.

German developed between the fifth and eleventh centuries from various dialects. The oldest written records in German date back to the 8th century. German dialects are traditionally divided into:

- *High German* – spoken in the south, it became the foundation of modern written German.
- *Low German* – spoken in the north, it evolved from Saxon and shares a common origin with Dutch and English.
- *Franconian dialects* – had a significant influence on the development of German.

Old English, also called Anglo-Saxon, evolved from the same base as Low German. This was due to the migration of the Saxons and Angles to the British Isles between the 5th and 7th centuries.

Later, English was influenced by Scandinavian languages due to Viking presence in England. However, the most significant external influence came with the Norman Conquest in 1066, which introduced a large number of French words and also affected the grammar of English.

Modern English

Nowadays, English has two main varieties: American and British. They differ slightly in spelling, with American English using simpler forms (e.g., *color* vs. *colour*), as well as in vocabulary and pronunciation.

In recent decades, the global influence of English has been growing, establishing itself as the modern *Lingua franca*. Among its two main varieties, American English appears to be more widespread than British English, though measuring this precisely is difficult. The distinction between them is not always clear, as many non-native speakers mix elements of both in their usage.

Historically, British English had a stronger global presence. Even today, in Europe and many other parts of the world, British English is the standard taught in schools. However, as people engage more with international media, technology, and online content, they are often exposed more to American English later in life.

The internet is strongly influenced by American English, shaped by multiple factors. It is widely used in global entertainment, including movies, TV, and music, and is also more common on social media platforms like Instagram, Facebook, and X (formerly Twitter), where a major proportion of content is in American English rather than British. Additionally, many major tech companies are based in the USA, which contributes to the widespread use of American English in technology and digital communication.

In recent years, the presence of American English has increased even further due to large language models (LLMs), which are primarily trained on internet sources, most of which are written in American English. This trend is likely to

continue as the use of AI models grows, exposing more people to American English rather than British English.

1.2.3 Influence of Cultures and Language Contact

When two nations interact over a long period, especially as neighbors, their languages often influence each other. This is particularly true when there are strong ties through trade, science, religion, or politics. Throughout history, larger and more influential nations have shaped the languages of smaller surrounding ones, often leaving a lasting impact.

A well-known example is the expansion of the Roman Empire, which spread Latin across much of Europe. Over time, Latin gave rise to the Romance languages and influenced many other linguistic groups, including Germanic and Slavic languages.

The Czech language has also been shaped by historical contact with German, especially during its long association with the Holy Roman Empire and later the Austro-Hungarian monarchy. Many German words entered Czech, particularly in areas like administration, trade, and urban life, and some of these loanwords are still used today.

Language contact is not just a historical phenomenon—it continues to shape languages today. Globalization, migration, and the dominance of English in international communication have led to the borrowing of many English words into other languages, including Czech. This process reflects how languages evolve based on cultural and technological influences over time.

Indirect borrowings

For borrowing to occur, there must have been some form of contact between the languages. This is why, for example, English could not have borrowed words directly from Greek—by the time the English language was forming, the period of Greek cultural dominance had already passed.

Loanwords from languages that never had direct contact can still appear, usually through an intermediate language. This phenomenon is called *indirect borrowing*. A word may pass through multiple languages before reaching its final form in Czech, with each stage potentially altering its pronunciation or meaning.

To illustrate indirect borrowing, the word *admiral* (Czech: *admirál*) originally comes from Arabic. The Arabic word *amīr* (commander) is followed by the definite article *al* when used in compounds, forming expressions like *amīr al-mā* (commander of the fleet) or *amīr al-baḥr* (commander of the waters). The term entered Latin, then passed into French as *amiral*, from where it was borrowed into English and German, eventually making its way into Czech.[R]

1.3 Etymology of Morphemes

Sometimes, only part of a word is borrowed from another language. For example, the Greek prefix *anti-* is used productively in many languages with native roots.

In the word *antivirus*, the prefix *anti-* comes from Greek (*anti-*, meaning "against"), while *virus* originates from Latin (*vīrus*, meaning "poison, slime, venom").

The word *antivirus* is a hybrid, with the prefix borrowed from Greek and the root from Latin. To determine the etymological origin of the word, we need to break it down into its smaller components. This is the core idea behind the *etymology of morphemes*—it allows for finer granularity in linguistic analysis. While some words are entirely borrowed or entirely native, many contain morphemes of different origins.

Other examples of words with mixed etymology include *television*, *sociology*, and *hyperactive*. The word *television* combines the Greek prefix *tele-* with the Latin root *vision*. Similarly, *sociology* is formed from the Latin *socius* and the Greek *-logy*. The word *hyperactive* follows the same pattern, with the Greek prefix *hyper-* and the Latin root *active*.

One Czech example is *kopírovat* ("to copy"), which originates from the German *kopieren*, itself derived from the Latin *copiare*. The root *kop-* is borrowed from Latin, *-ír-* reflects the German verb-forming element *-ieren*, and *-ovat* is a native Czech verb-forming suffix.

1.3.1 Calques

Sometimes a word is not borrowed in its original form, but instead its structure or meaning is translated with the use of native morphs. This process is called a *calquing*, or loan translation. A *calque* is typically a morpheme-by-morpheme translation of a word from another language, transferring meaning without borrowing actual morphemes (Thomason, 2001).

An example from Czech is the word *předseda* ("chairman"), which is a calque of the German *Vorsitzer*, itself based on the Latin *praesidēns*, from the verb *praesidēre*, composed of *prae-* ("before") and *sedēre* ("to sit")—literally meaning "the one who sits in front."

The Czech word follows the same structure, combining *před-* ("before") and *seda*, derived from the verb *sedět* ("to sit").

Another example is *časopis* ("magazine"), a calque of the German word *Zeitschrift*, which comes from *Zeit* ("time") and *Schrift* ("writing"). The Czech equivalent mirrors this by combining *čas* ("time") and *pis* ("writing").

An example borrowed from English is *mrakodrap* ("skyscraper"). It is a calque of the English compound *skyscraper*, made up of *sky* and *scraper*. The Czech version uses *mrak* ("cloud") and *drap* (from *drápat*, "to scrape") — literally something that "scrapes the clouds."

These and many more examples are described in the Czech Etymological dictionary by (Rejzek, 2019).

Calques and Etymological Ambiguity

With calques, it becomes difficult to clearly determine whether a word should be considered a loan. While the structure and meaning are borrowed, the actual morphological material (the individual morphs) remains native. On the word level, it is reasonable to consider such words as borrowed since they would not exist without the influence of the source language.

However, on the morph level, the situation is more complex. For instance, in the Czech word *časopis* ("magazine"), both morphemes—*čas* ("time") and *pis* ("writing")—are native Czech, both tracing back to Proto-Slavic. Even though the word is a calque of the German *Zeitschrift*, the individual components are not borrowed in form, only in conceptual structure.

In such cases, we do not take the borrowed structure into account when determining the etymology of individual morphs. Instead, we evaluate each morph on its own. If the morphemes are of native origin, we classify them as native, regardless of the fact that the overall word may be a calque or structurally borrowed.

2 Related Work and Data Sources

To the best of our knowledge, there is no existing research that focuses on identifying the etymological origin of individual morphs. There is currently no established or widely accepted methodology for this task.

Most work on etymological classification has been done at the word level. Similarly, there is a lot of research on morphological segmentation and classification. However, we have not found any scientific paper that clearly combines these two areas to study etymology at the level of individual morphs.

2.1 Deconstructor

One related tool that attempts a similar task is the online application *Deconstructor*[1]. It performs morphological segmentation and tries to reconstruct the etymology of individual morphs within a word. It also visualizes how the word may have been formed from its components. The tool supports multiple languages and uses a large language model (LLM) to generate its output.

The interface presents the analysis in a clear, graphical format, showing a step-by-step construction of the word. However, it outputs only a single origin language for each morph, without including any intermediate borrowing stages or full etymological chains.

The tool performs quite well for English. For Czech, however, the results are often less reliable, mainly due to inaccurate morphological segmentation.

It is not possible to directly compare this system to the approaches used in this thesis. Deconstructor is a web-based tool that processes one word at a time and uses its own segmentation and output format, which differs from the annotated dataset and evaluation methods used in this work.

In the following picture is an example of output to entry for word *telefonní*.

[1] `https://deconstructor.ayush.digital`

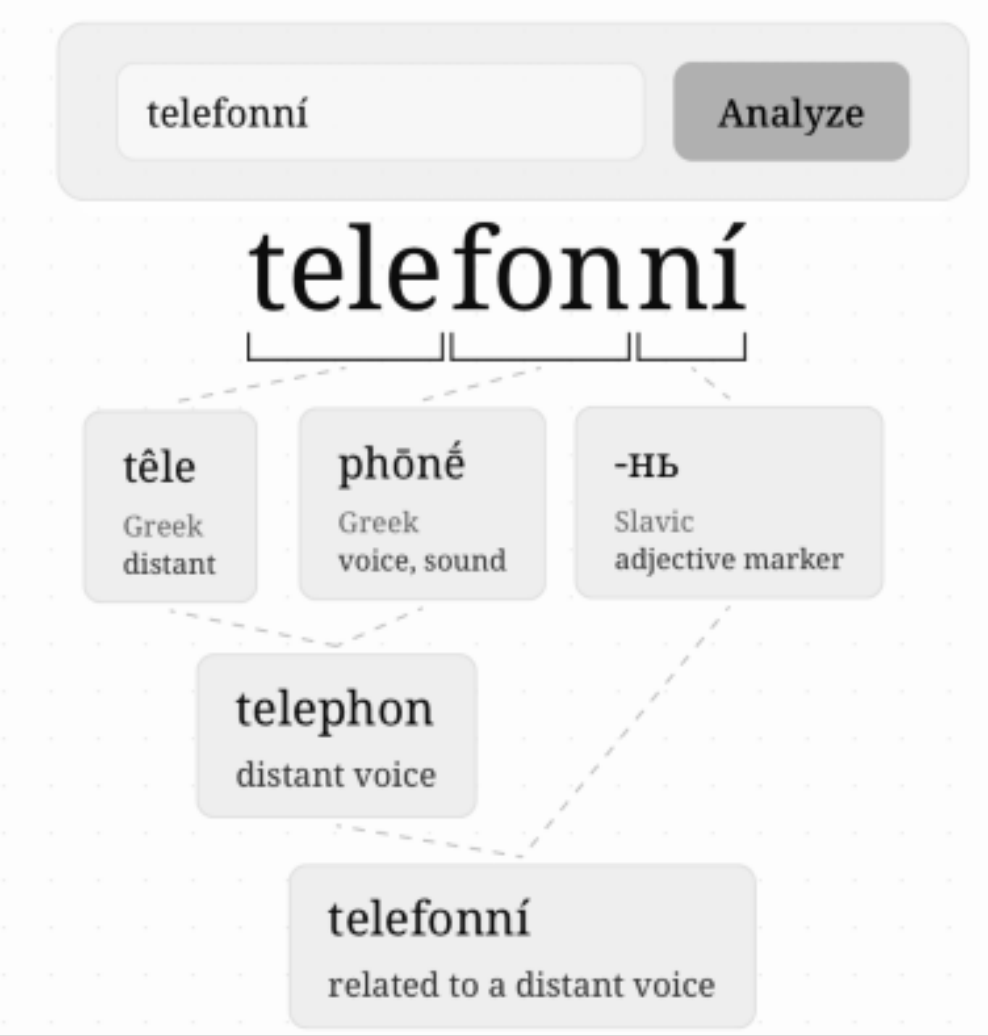

Figure 2.1 Example of the Deconstructor web interface output

Further, I talk more about the tools and datasets I have used.

2.2 DeriNet

DeriNet is a large-scale lexical network that models derivational and compositional relations in Czech. Each node represents a lexeme, while edges capture word-formation links, either connecting derived words to their base forms or linking compounds with their components. The dataset is based on the MorfFlex CZ dictionary (Hajič et al., 2024) and includes linguistic annotations such as part-of-speech tags, segmentation, morphological classification, corpus frequency, and etymological information sourced from the Czech Etymological Lexicon (CzEtyL).

The latest version, DeriNet 2.3, developed by multiple authors from ÚFAL, MFF UK (Olbrich et al., 2025), comprises 1,040,126 lexemes and 791,771 derivational and 7,598 compound relations. It also includes 5,781 derivational trees containing loanwords, enriched with etymological data from CzEtyL.

In addition, DeriNet provides detailed morphological segmentation and classification of morphs, making it a valuable resource for this work.

Borrowed words often become bases for new derivations that include native

affixes. Even when a native affix is added to a borrowed stem, the resulting word is still considered a loanword. However, if we classify each morpheme individually, some will be identified as borrowed and others as native.

2.3 Czech Etymological Dictionary

The most recent Czech etymological dictionary by Rejzek (2019) contains over 11,000 main entries, covering both standard vocabulary and more recent or marginally grammatical words. In addition, it includes around 21,000 derived forms and nearly 64,000 references to words in other languages.

The dictionary also features an introductory chapter on general etymology, the development of world languages, and the evolution of Czech from its Indo-European roots to its modern form. It includes a classification of related languages across the world.

2.3.1 CzEtyL

- Should I mention I am one of the co-authors of this lexicon?

The Czech Etymological Dictionary by Rejzek (2019) is a great resource, but it is not designed for computational use. The categories of information provided for each entry are not always consistent, as it is written primarily for human readers.

The Czech Etymological Lexicon by Rejzek et al. (2025) attempts to extract the essential etymological information for each entry. It focuses only on identifying the source languages from which a given word was borrowed, omitting additional details such as the original form in the source language, the time of borrowing, and references to similar words.

Version 1.0 of the dataset includes approximately 10,500 Czech words, each annotated with a sequence of ISO 639-3 language codes that indicate their etymological origin.

The data is organized in a tab-separated format with three columns:

- **First column:** Lists the lemma.
- **Second column:** Provides the corresponding language codes, separated by commas.
- **Third column:** Specifies whether the word is classified as a loanword ("loan") or a native word ("native"). In this classification, "native" refers to words that have naturally evolved in the language rather than being borrowed from another.

Example entry:

```
architekt   deu,lat,ell   loan
```

The word *architekt* originated from Greek and entered Czech through Latin and German.

The morphemes are derived from Greek:

- **Archi-** – meaning "main, leader," from *árkhō* ("I command")
- **-téktōn** – meaning "craftsman, artist"

Affixes entries

Although the lexicon is word-based, it does include a few examples of affixes with annotated origins. For instance:

- Prefixes:
  - Greek prefixes: *aero-* (*aerodynamika*), *anti-* (*antivirus*), *astro-* (*astronomie*), *elektro-* (*elektromagnetismus*)
  - Latin prefixes: *ab-* (*abdikace*), *ad-* (*administrace*), *dis-* (*disfunkce*), *per-* (*perforace*), *re-* (*rekonstrukce*)
- Suffixes:
  - English: *-bal* (*fotbal*, *handbal*)
  - Latin: *-ace* (*rekreace*, *prezentace*), *-iz* (*organizace*, *realizace*)
  - Czech: *-náct*, used in numerals from eleven to nineteen (e.g., *jedenáct*, *devatenáct*)

2.3.2 SIGMORPHON Shared Task

SIGMORPHON (Special Interest Group on Computational Morphology and Phonology) regularly organizes shared tasks to support progress in morphological analysis. This thesis uses data from the *SIGMORPHON 2022 Shared Task on Morpheme Segmentation* (SIGMORPHON, 2022).

The dataset was constructed by integrating several major morphological resources, including UniMorph for inflectional morphology (McCarthy et al., 2020), MorphyNet for derivational morphology (Batsuren et al., 2021), Universal Dependencies (Nivre et al., 2017), and multiple editions of Wiktionary for compounds and root words (Wiktionary, [n.d.]).

This work focuses on the Czech portion of the dataset, which contains full sentences segmented into individual morphs. These annotations provide the morphological basis for the etymological task addressed in this thesis.

2.4 Segmentation

Morphological segmentation is the task of splitting words in a sentence into individual morphs. This thesis uses data provided by the *SIGMORPHON 2022 Shared Task on Morpheme Segmentation*, which focuses specifically on this task. In this work, the manually segmented sentences from the shared task are used.

To enable morph-level etymology prediction, the morphs in the selected Czech sentences were manually annotated with etymological labels, forming a dataset for evaluation and supervised learning.

If other datasets were to be used, a morphological segmentation step would be required beforehand, as morph-level etymology prediction depends on having the words already segmented into morphs.

2.5 Classification

Classification of morphs serves as a valuable feature for morph-level etymology prediction. As discussed in Section 1.1.2, there are several ways to categorize morphs. In this work, the primary distinction is between *roots* and *affixes*.

Affixes are further categorized based on two criteria: their function—either *derivational* or *inflectional*—and their position relative to the root within the word. According to position, affixes are labeled as *prefixes*, *suffixes*, or *interfixes*.

The morphological classification used in this thesis was generated automatically using a model developed by V. John (2024).

3 Practical Part

In this chapter, we describe the data used and created for this work. We introduce several baseline methods for etymology prediction to establish a lower bound for model performance. To estimate an upper bound, we consider inter-annotator agreement as a reference for human-level consistency. Finally, we define the evaluation methods used to assess the quality of model predictions.

3.1 Data

Currently, there is no publicly available resource that provides words annotated with etymological information at the morph level. The Czech Etymological Dictionary by Rejzek (2019) is a valuable source at the word level, and its digital form, CzEtyL (Rejzek et al., 2025), is suitable for automated processing.

Using DeriNet, we extracted morphological segmentation and classification for all words present in CzEtyL. This resulted in a dataset of approximately 10,500 words, each annotated with segmentation, morph classification, and word-level etymology.

Assuming that the etymology provided in CzEtyL corresponds to the root of the word, we can construct a dictionary of roots and their etymological origins by assigning the etymology of the whole word to its root morph. While this approach is a simplification, it offers a rough but practical approximation of morph-level etymology for a large number of root morphs.

CzEtyL also includes a list of approximately *250* affixes with known etymologies, covering both borrowed and native affixes. Together with the root dictionary constructed from CzEtyL and DeriNet, this serves as a useful basis for simple baseline predictions or as an extension to the training data.

3.1.1 Annotations

To create a dataset of sentences where words are segmented and annotated with etymological information at the morph level, data from the SIGMORPHON 2022 Shared Task were used. This dataset contains sentences that are already segmented into individual morphs. Each morph was then manually annotated with a sequence of language origins, based primarily on information from the Czech Etymological Dictionary by Rejzek (2019).

The annotated dataset was divided into three parts: training, development, and test sets. The table below shows the number of sentences, words, and morphs in each part of the dataset.

Dataset	Sentences	Words	Morphs
Training set	200	2,774	7,016
Development set	50	599	1,460
Test set	50	609	1,485

Table 3.1 Size of the annotated dataset used for training, development, and testing

Only morphs that are candidates for etymological classification are counted. This excludes punctuation, numerals, abbreviations, and special symbols.

The training sentences were selected from the Sigmorphon shared task train set for Czech, while the development and test sets were derived from the original development set.

The training data consists of approximately 170 sentences from the beginning and about 30 sentences from the end of the original training file.

The development and test sets were selected from the development file, which contains 500 sentences in total. Every 10^{th} sentence (1st, 11th, 21st, ...) was assigned to the test set, and every 10^{th} sentence starting from the second (2nd, 12th, 22nd, ...) was assigned to the development set.

The remaining 400 sentences from the development file were not used in this work and are kept for possible future evaluation or additional training data. Similarly, the 500 sentences from the official test file were also left unused.

It would be beneficial to have more annotated sentences available for both training and evaluation. However, the annotation process is time-consuming and requires either a certain level of expertise or considerable effort spent searching for etymological information in various sources. This naturally limits how much annotated data can be realistically produced.

Example annotation:

Sentence: *Faxu škodí především přetížené telefonní linky*

- **Faxu**
  - *Fax* — `R` — `eng,lat`
  - *u* — `I` — `ces`
- **škodí**
  - *škod* — `R` — `gmh`
  - *í* — `I` — `ces`
- **především**
  - *přede* — `D` — `ces`
  - *vš* — `R` — `ces`
  - *í* — `I` — `ces`
  - *m* — `I` — `ces`
- **přetížené**
  - *pře* — `D` — `ces`
  - *tíž* — `R` — `ces`
  - *en* — `D` — `ces`
  - *é* — `I` — `ces`
- **telefonní**

  - *tele* — `R` — `ell`
  - *fon* — `R` — `ell`
  - *n* — `D` — `ces`
  - *í* — `I` — `ces`
- **linky**
  - *lin* — `R` — `deu,lat`
  - *k* — `D` — `ces`
  - *y* — `I` — `ces`

The annotation uses the following abbreviations:

- `R` – Root
- `D` – Derivational affix
- `I` – Inflectional affix

  Language codes follow ISO 639-3:

  - `ces` – Czech
  - `deu` – German
  - `ell` – Greek
  - `lat` – Latin
  - `eng` – English
  - `fra` – French
  - `gmh` – Middle High German
  - ...

The morph types (root, derivational affix, inflectional affix) are included in the dataset and were obtained by automatic prediction. The language origin sequences were manually annotated.

3.2 Baselines

To evaluate how well the model performs, we first need to define a few baselines. These serve as reference points, helping us understand whether the model actually learns something useful or if similar results could be achieved using much simpler approaches.

In this work, we define several baselines: a trivial one that always predicts Czech, a memorization-based baseline, a word-level lemmatization approach using CzEtyL, and a root-based approach using CzEtyL.

3.2.1 Always-Czech Baseline

A very simple baseline is to always predict the most frequent target (Czech) as the etymological origin for all morphs. Despite its simplicity, this approach gets a high score, which is mainly due to the fact that the majority of morphs in the dataset are of Czech origin.

3.2.2 Memorization-Based Approach

This approach simply memorizes all morphs seen during training and assigns them their most frequent etymology. If a morph appears in the test data and was seen in training, its stored etymology is used. In cases where a morph was annotated with multiple different etymologies in training, the most frequent one is selected.

This method can perform surprisingly well, especially when the training data covers a large portion of the vocabulary in the test set.

However, its main limitation is handling unseen morphs. If a morph is not present in the training data, the system defaults to predicting the most frequent class—typically the native language - `ces` (Czech).

3.2.3 Word Lemmatization Approach

This baseline uses a morphological analyzer to obtain the lemma for each word, which is then looked up in an expanded version of CzEtyL. The retrieved etymology is then assigned to the root morph.

Affixes are matched against a list of known borrowed affixes; otherwise, they are assumed to be native. Inflectional endings are always treated as Czech. If the lemma is missing from CzEtyL, the whole word is considered Czech by default.

The expanded version of CzEtyL builds on the original lexicon by incorporating derivational relations from DeriNet. For each entry, all lexemes belonging to the same derivational tree are included and assigned the same etymological label. This significantly improves coverage—expanding the number of annotated words from roughly 10,500 to around 511,000.

3.2.4 Morph Based Baseline

For this baseline, we assume that the etymology provided in CzEtyL corresponds to the root of the word. Affixes are considered native unless they appear in a predefined list of known borrowed affixes, which is also part of CzEtyL. All inflectional affixes are assumed to be native by default.

To build this baseline, we iterate through all words in CzEtyL, extract their roots, and assign the full word-level etymology to those root morphs. This produces a mapping from root morphs to a multi-set of possible etymological origin sequences based on their occurrences.

The algorithm then applies the following logic:

- If the morph is classified as a root, it is assigned the most frequent origin sequence from its associated multi-set.

- If the morph is a derivational affix, we check whether it appears in the list of known borrowed affixes and assign its origin accordingly.
- For inflectional affixes, or in cases where the morph is not found in the root or affix dictionaries, we default to Czech (`ces`), the most frequent class.

3.3 Evaluation Methodology

The goal is to predict, for each morph, a sequence of languages starting from the original source language, through any intermediate languages, before getting to Czech. Since the order of languages in the sequence is usually fixed or clear from context, the prediction is evaluated as an unordered set. For example, if the correct languages are Latin and German, it is almost always the case that the borrowing path was from Latin through German, not the other way around, so the order is not considered in the evaluation.

3.3.1 F1-Score

To evaluate the quality of predictions, we use the F1-score, which provides a more balanced evaluation than simple accuracy. We don't want to just check whether the prediction is exactly the same as the target or not—we also want to measure how close it is, even if it's only partly correct. It's important that the model predicts as many correct languages from the sequence as possible (recall), but also avoids adding incorrect ones (precision). Since both aspects matter equally in this task, we combine them using the F1-score.

We calculate the F1-score for each individual morph occurrence and then take the average across all morphs in the dataset. This means that morphs which appear more frequently have a bigger influence on the final score.

Alternatively, we could compute micro F1-score by summing up the total number of correct predictions, total predicted languages, and total gold labels across all morphs—and calculating precision, recall, and F1 from those aggregate counts.

As another approach, we can split the dataset into two subsets—native and borrowed morphs—and compute separate F1 scores for each group. This provides a more detailed view of how the model performs across different categories of morphs, especially since borrowed ones are generally more challenging to classify due to their lower frequency and more complex etymological paths.

3.3.2 Native vs. Borrowed

Because the majority of morphs in the dataset are native Czech, computing the F1-score across all morph instances can lead to results biased toward this dominant class. To address this imbalance and gain deeper insight into model behavior, we additionally evaluate performance separately for two categories of morphs: native and borrowed.

We define native morphs as those whose target etymology is `ces`. All other morphs are considered borrowed. By reporting F1-scores for each group independently, we can better observe how well the model performs on native morphs

and how well on borrowed morphs. Borrowed morphs typically present a greater challenge for the model, as they can originate from a wide variety of languages and may follow complex borrowing paths involving multiple intermediate languages.

3.3.3 Score Grouped by Morphs

In the dataset, many morphs appear multiple times—on average about 7 times—but some affixes, especially inflectional endings, occur dozens of times. In the training set, there are 7,205 total morphs, of which only 972 are unique. The 10 most frequent morphs alone account for 1,965 occurrences; the top 20 cover 2,864, and the top 50 together make up 3,997 occurrences—more than half of the dataset.

To reduce the bias introduced by these highly frequent morphs, we also report an additional metric where morphs are grouped by their surface form. For each unique morph, we compute the average F1-score across all of its occurrences. These per-morph scores are then averaged to obtain the final result.

3.3.4 Relative error reduction

Because the baseline for this task is so high, absolute values of the F1 score or accuracy do not fully reflect the improvements made by better models. A model might outperform the baseline by only a few percentage points, even though it significantly reduces the number of actual errors. To better highlight these improvements, we report the *relative error reduction* compared to the baseline.

The formula for computing the relative error reduction is:

$$\text{Error Reduction} = \frac{\text{Error}_{\text{baseline}} - \text{Error}_{\text{model}}}{\text{Error}_{\text{baseline}}}$$

where error is defined as $1 - \text{F1-score}$.

3.3.5 Expected Bounds of Performance

To properly interpret the results of this task, it is important to establish both a lower and an upper bound—defining an interval within which realistic performance can be expected. The lower bound is represented by simple baselines such as always predicting the most frequent class, memorizing morphs from the training data, or applying basic rule-based methods using available etymological sources.

The upper bound, on the other hand, is more difficult to define. A model with 100% error reduction would be perfect; however, such performance is not achievable. Even human annotators can struggle to consistently determine the correct etymology of morphs. Without access to reference materials like etymological dictionaries, most people would not perform better than the simpler baselines.

Problems with annotation

Even when such resources are available, disagreement is common, especially in complex or ambiguous cases. Etymological dictionaries themselves may contain inconsistencies, outdated interpretations, or lack coverage of many words.

Etymology is often uncertain, especially when it comes to the path a word or morph took through multiple languages before entering the target language. While the goal of this work is to annotate the full borrowing chain—including intermediate languages, not just the ultimate source—such detailed information is not always available. Many etymological resources focus only on the original source, and in many cases, we simply do not know which languages served as intermediaries.

Furthermore, some words have been borrowed multiple times or simultaneously from different sources, making it even harder to determine a single, well-defined etymological path.

Identifying the etymological sequence at the morph level is even more challenging. Assigning etymological labels to individual morphs is not a standard practice—most etymological dictionaries and linguistic studies focus on whole words. As a result, there are no established guidelines for morph-level annotation, and it often relies on individual interpretation. This makes the annotation process inherently difficult and frequently unclear.

3.3.6 Inter-Annotator Agreement

Inter-annotator agreement measures how consistently different annotators label the same data. It provides a useful way to assess both the subjectivity of the task and the reliability of the annotation process.

This is particularly relevant in the context of etymological annotation for individual morphs, which can be unclear. Measuring agreement can help identify borderline or problematic cases and ensure greater consistency across the dataset.

Moreover, inter-annotator agreement serves as a practical upper bound for the performance of automatic models. If human annotators cannot consistently agree on the correct etymology of a morph, it is unrealistic to expect a model to achieve significantly better accuracy.

Cohen's kappa

One commonly used metric for measuring inter-annotator agreement is *Cohen's kappa*, introduced by Cohen (1960). It evaluates the agreement between two annotators assigning categorical labels to a dataset.

Unlike simple percentage agreement, Cohen's kappa also accounts for the agreement that might occur purely by chance, which is especially important when one category is much more frequent than the others.

The formula for computing Cohen's kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

where:

- p_o is the observed agreement between the annotators,
- p_e is the expected agreement by chance.

The expected agreement p_e is calculated based on the marginal probabilities for each category assigned by the annotators:

$$p_e = \sum_{i=1}^{k} p_i^{(1)} \cdot p_i^{(2)} \tag{3.2}$$

where:

- k is the number of categories,
- $p_i^{(1)}$ is the proportion of items that annotator 1 assigns to category i,
- $p_i^{(2)}$ is the proportion of items that annotator 2 assigns to category i.

The value of Cohen's kappa ranges from -1 to 1. A higher value indicates stronger agreement between the two annotators, with 1 representing perfect agreement and 0 corresponding to agreement equal to chance. Negative values suggest less agreement than would be expected by chance. However, there is no universally accepted interpretation of specific kappa values, and their evaluation often depends on the context and nature of the task.

3.4 Results of inter annotator agreement

Annotation on the development set was also performed by a second annotator, a final-year high school student with no prior experience in linguistic annotation. Before beginning, he was introduced to the task and provided with a thorough explanation of the annotation scheme. He was given access to the Czech Etymological Dictionary by Rejzek (2019) and directed to additional resources such as Wiktionary[1]. Although not a trained linguist, his annotation offers a valuable reference point for estimating human performance on this task.

We compute Cohen's kappa and the percentage of exactly matched annotations.

- Cohen's kappa: 0.82
- Exact match: 95.96%

These results show a high level of agreement between the two annotators. The differences were sometimes caused by clear mistakes made by one of the annotators, which, once corrected, helped improve the overall quality of the dataset. In other cases, the disagreement arose from differences in the granularity of annotation. For example, one annotator might have labeled a morph as `gmh` (Middle High German), while the other used the more general `deu` (German).

There were also cases where inconsistencies between different etymological sources led to disagreement. This highlights the difficulty of the task—even professional linguists do not always agree on the etymology of certain words. In this work, the challenge is even greater, as the goal is to determine etymology at the level of individual morphs rather than whole words.

[1] `https://www.wiktionary.org`

4 Model for Morph-Level Etymology Prediction

This chapter describes the approaches explored to develop a model capable of predicting the etymological origin of individual morphs in Czech. The goal is to assign each morph a sequence of languages from which it was borrowed into Czech.

4.1 Features

To successfully predict the etymological origin of each morph, it is essential to extract useful features from the annotated data. These features serve as input to the classification model and help it learn patterns associated with language origin.

Several types of features were used in this work, each capturing different aspects of the morph and its context. These include character-level n-grams, morphological classification, position within the word, and others. Additionally, abstract representations such as morph and word embeddings were also tested.

The following subsections describe each feature type in more detail.

4.1.1 Morph n-grams

From the text of each morph, we extract character n-grams and convert them into sparse vectors. This is done by building a vocabulary of all n-gram combinations seen during training. Then, for each morph, a vector is created where each dimension counts how many times a specific n-gram from the vocabulary appears in the morph's text.

The idea behind using character n-grams is that different languages tend to favor specific letter combinations, and the model can learn to associate these patterns with particular language origins. In this work, we focus on 1-grams and 2-grams, which are both effective and computationally feasible.

Using 3-grams would significantly increase the dimensionality of the feature space due to the large number of possible combinations. Since the average morph length is only about 2.2 characters, most morphs do not even contain three characters, making 3-grams less useful in practice.

Additionally, including 3-grams increases the risk of overfitting, as the model may start memorizing individual morphs and their corresponding etymologies instead of learning general patterns.

One drawback of this method is that morphs containing character combinations unseen during training will result in a zero vector. In such cases, the model lacks meaningful input for prediction.

However, this is less of a problem with enough training data. The model may also learn to default to the most frequent class when encountering unknown n-grams or rely more heavily on other features.

4.1.2 Morph Types

Each morph is classified into one of three basic categories: *Root*, *Derivational affix*, or *Inflectional affix*. This classification is encoded using a one-hot representation.

In addition, a positional classification is also extracted. Each morph is categorized as either *Root*, *Prefix*, *Suffix*, or *Interfix*, depending on its position relative to the root(s) in the word. Typically, a word contains a single root, but in the case of compounds, multiple roots may be present. Affixes that appear before the first root are labeled as prefixes, those that come after the last root as suffixes, and those occurring between two roots are labeled as interfixes.

In rare cases, there are words that do not contain any identifiable root. This situation often occurs with certain prepositions or conjunctions, which may consist of only a single morph, making it unclear how to classify them.

There are also examples of multi-morph words that lack a root entirely. For instance, the words přední and zadní ("front" and "rear") are formed from two affixes: před- / zad- and the adjectival suffix -ní without a clear root morpheme present.

In these cases, the first (often the only) morph is classified as a *root* with respect to the position type, while the remaining morphs are classified as *suffixes*.

Just approximately 3 % of words in the dataset do not contain a root and consist of more than one morph.

4.1.3 Vowel Start and End

This feature captures whether a morph starts and/or ends with a vowel. Two binary values are used: one indicating if the first character is a vowel, and the other if the last character is a vowel.

Some languages tend to favor specific phonological patterns, such as vowel-final or vowel-initial affixes. Even though this is a very simple feature, it can still help the model notice language-specific patterns.

This feature is especially helpful when combined with information about whether the morph is a root or an affix. For example, many Proto-Indo-European roots follow a consonant-vowel-consonant (CVC) structure. On the other hand, affixes often connect to roots by starting or ending with a vowel for better pronunciation. So, prefixes more often end with vowels, and suffixes often begin with them. **Add Citation**

4.1.4 Embeddings

Using embeddings is a way to represent words or sub-words as numerical vectors that capture aspects of their meaning. The core idea is that words with similar meanings are mapped to vectors that are close to each other in the embedding space.

The first widely adopted word embeddings were Word2Vec embeddings. An extension of this approach is FastText by Bojanowski et al. (2017), which works at the sub-word level rather than treating words as atomic units. In FastText, each word is represented as a bag of character n-grams, and its final embedding is

computed as the sum of the embeddings of its n-grams. This allows FastText to generate embeddings even for words that were not seen during training.

This property is particularly useful in tasks like morpheme-level analysis, as it enables the model to generate meaningful representations for individual morphs. This makes FastText a good fit for the task of predicting morpheme etymology, where we need embeddings for affixes and roots that are not standalone words.

In this work, we use FastText embeddings trained on the Czech language, provided by Grave et al. (2018). These embeddings were trained on Czech Wikipedia and Common Crawl data and have 300 dimensions. To reduce dimensionality and improve how effectively the classification model can learn from these features, the embeddings can optionally be compressed to a lower dimension using Principal Component Analysis (PCA).

When predicting the etymology of a morph, embeddings can be computed both for the entire word and for the morph itself.

4.2 Extending the Training Data with Etymological Dictionary

The root and affix dictionaries previously extracted for use in baseline models can also serve as a valuable source of additional training data for the learning model. By adding these entries into the training set, the model has more examples of morphs with known etymology.

The original training set contains 7,205 morphs, of which 972 are unique. After extending the data with dictionary entries, the training set increases to 19,430 morphs, covering 12,211 unique morphs. The dictionary entries consist of roots and affixes, which typically do not repeat, so the added morphs are almost all unique.

4.3 Representation of Target Classes

There are two main strategies for modeling the target classes in morph-level etymology prediction:

- **Whole-sequence classification:** The entire sequence of languages is treated as a single label. This reduces the task to a standard classification problem where each unique language sequence is a distinct class. The model predicts exactly one of the predefined sequences (sequence that appeared in the train set) for each morph.
- **Multi-label classification:** Each language is treated as an independent label. The model predicts a subset of languages (from languages which appeared in the train set), determining for each one whether it should be included in the sequence. This allows the model to predict language sequences which were not seen in training.

The first approach is simple but limited to the set of language sequences present in the training data. The second one is more flexible and can produce

combinations of languages not explicitly seen during training, but it also increases the complexity of the classification task. We have to train one classifier for each language. This can be viewed as a binary classification for each language, determining whether or not it should be included in the output sequence.

4.3.1 Label Statistics from Training Data

To help choose an appropriate strategy for representing the target labels, it is useful to examine the number of unique language sequences and individual languages present in the training data. In the manually annotated Czech sentences used for training, the following statistics were observed:

- **Unique language sequences:** 35
- **Unique languages:** 11

Top Etymology Sequences

The following table shows the 10 most common etymological origin sequences found in the training data. The dataset contains approximately 7,100 annotated morphs. As shown, the majority of morphs are of Czech origin.

Language Sequence	Count
`ces` (Czech)	6,229
`lat` (Latin)	280
`ell` (Greek)	100
`deu,lat` (German, Latin)	84
`eng,lat` (English, Latin)	41
`lat,ell` (Latin, Greek)	38
`ita,deu` (Italian, German)	38
`eng` (English)	36
`deu,lat,ell` (German, Latin, Greek)	29
`fra,lat` (French, Latin)	26

Table 4.1 Top 10 most frequent etymological sequences in the training data

4.3.2 Extended Training Data

To provide more data for the model, we can extend the dataset with entries extracted from root and affix etymology dictionaries. This significantly increases label diversity:

- **Total morphs in extended dataset:** 19 430
- **Total unique morphs in extended dataset:** 12 211
- **Unique language sequences:** 406
- **Unique languages:** 67

To reduce noise from very rare classes and simplify the model—especially in the multilabel setting—we consider frequency cutoffs to filter out infrequent classes.

Minimum Frequency	Unique Sequences	Unique Languages
All	406	67
> 1 occurrence	216	54
> 2 occurrences	141	43
> 3 occurrences	111	35
> 4 occurrences	95	33
> 5 occurrences	81	27

Table 4.2 Number of unique language sequences and languages in the training data based on minimum frequency thresholds.

If we limit the classes to only those that appear with a frequency greater than 1 per 1,000 morphs (i.e., more than 19 occurrences), we are left with:

- **13 most frequent languages**
- **40 most frequent language sequences**

This shows that although the overall label space is large, the majority of the data is concentrated in a relatively small number of classes.

While such filtering helps reduce noise and simplify the label space, we avoid applying it too aggressively. The model itself should learn which labels are rare or unlikely based on the training data, rather than having these decisions hard-coded during preprocessing.

The high number of low-frequency language sequences may also be partly due to incorrectly parsed etymological information from the Czech Etymological Dictionary. In some cases, the dictionary lists multiple languages not as part of the actual borrowing chain, but rather to indicate similar borrowings in other languages or to illustrate certain phenomena. This can introduce noise into the dataset.

Most Frequent Language Origins

The extended training set contains 19,430 annotated morphs. While there is a wide variety of etymological origin sequences, a large portion of the data is concentrated in just a few frequent classes. The table below lists the most common language origin sequences, showing that the majority of morphs come from a relatively small set of etymological paths.

Language Sequence	Count
`ces` (Czech)	11,736
`lat` (Latin)	1,776
`ell` (Greek)	638
`deu` (German)	616
`lat,ell` (Latin, Greek)	515
`deu,lat` (German, Latin)	388
`eng` (English)	380
`fra,lat` (French, Latin)	297
`deu,fra,lat` (German, French, Latin)	196
`fra` (French)	195
`gmh` (Middle High German)	166
`deu,lat,ell` (German, Latin, Greek)	138

Table 4.3 Most common etymological origin sequences in the extended training set (∼19,000 morphs)

4.4 Classification Models

For the final classification step, which predicts the etymological origin of each morph based on extracted features, we use standard machine learning models implemented with the `scikit-learn` library. Since the amount of annotated training data is relatively small, larger and more complex models would likely not perform so well.

For this reason, we focus on simpler models, especially Logistic Regression (LR), Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP).

4.5 Self training

Using some model to predict targets on a big amount of data and then use that as training data. Did not work much on the rest of train and dev data from the Sigmorphon shared task. That is just 1200 sentences (800 from train and 400 from development). On other data, we would need segmentation and classification.

4.6 Documentation, Code

Is it sufficient to describe it just here briefly?

All code and data used in this thesis are publicly available on GitHub: `https://github.com/ampapacek/MorphemeOrigin`

The repository includes:

- Python scripts implementing the baseline models, feature extraction, and training pipeline for the machine learning classifier.
- Annotated datasets used for training, development, and evaluation.
- Affix and root dictionaries extracted from etymological sources.

- Scripts for computing evaluation metrics and inter-annotator agreement.

The code is written in Python and relies primarily on the `scikit-learn` library. Experiments can be reproduced by running the main script with the desired parameters. The repository also includes a `Makefile` for convenient setup.

The seed used for all experiments is 34867991 (university identification number of the author).

5 Experiments and Results

- First results on the development set and why we decided on some model parameter options.

In the next section on the test set.

5.1 Experiments on Development Set

Just quickly added some results. Will provide more and structure it better.

Model	F1	RER	Native	Borrowed	Unique
– Baselines –					
Dummy Baseline	90.1	0.0	100.0	0.0	79.4
Word Dictionary	94.0	39.2	98.5	53.0	88.9
Most Frequent Origin	94.2	41.8	99.3	48.0	86.8
Morph Dictionary	94.6	45.4	98.9	55.7	89.5

Table 5.1 Performance of baseline models. RER = Relative Error Reduction over the Dummy Baseline. Unique = Score grouped by morph surface forms.

Model	F1	RER	Native	Borrowed	Unique
– Learning models –					
Logistic Regression	94.0	39.7	98.7	51.4	86.2
SVM-base	94.7	46.5	99.5	50.7	88.1
SVM-embeddings	94.6	45.5	99.5	50.3	87.8
SVM-extend-multi	95.2	51.9	99.6	55.4	89.7
SVM-emb-extend-multi	95.1	51.1	99.0	60.5	89.4
MLP-base	95.5	54.5	99.2	62.0	90.3
MLP-extend-multi	96.1	60.4	99.0	69.8	91.5
MLP-emb-extend-multi	95.8	57.5	99.0	66.5	91.0

Table 5.2 Evaluation results of different models on the development set. All metrics are F1 scores in %, The first column is F1 averaged per each morph instance. RER is relative error reduction over the dummy baseline in %. See chapter 3.3 on evaluation methodology for more details.

MLP-extend-multi refers to a multi-layer perceptron classifier with a hidden layer of size 30, trained without embeddings in a multi-label setting using a one-vs-rest strategy and on extended data from CzEtyL.
emb - use of embeddings
MLP-base refers to a MLP with 30 neurons without embeddings, which predicts the whole sequence at once, only trained on default data.

5.2 Final Results on Test Set

Conclusion

- Summary of findings.

Bibliography

Aronoff, M.; Fudeman, K., 2011. *What is Morphology?* Wiley. Fundamentals of Linguistics. ISBN 9781444351767. Available also from: https://books.google.cz/books?id=bolGMMyZVjMC.

Batsuren, K. et al., 2021. MorphyNet: A Large Multilingual Database of Derivational and Inflectional Morphology. In: *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 39–48.

Bojanowski, P. et al., 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. Vol. 5, pp. 135–146. ISSN 2307-387X. Available from DOI: 10.1162/tacl_a_00051.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. Vol. 20, no. 1, pp. 37–46.

Eberhard, D. M. et al., 2025. *Ethnologue: Languages of the World*. 28th. SIL International. Available also from: https://www.ethnologue.com/insights/largest-families/. Online version, accessed on 21.2. 2025.

Goethe, J. von, 1817. *Zur Morphologie*. J.G. Cotta. Goethes Werke, no. sv. 1. Available also from: https://books.google.cz/books?id=jk5StAEACAAJ.

Grave, E. et al., 2018. Learning Word Vectors for 157 Languages. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hajič, J. et al., 2024. *MorfFlex CZ 2.1*. Available also from: https://hdl.handle.net/11234/1-5833. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Haspelmath, M., 2020. *The morph as a minimal linguistic form*. De Gruyter Mouton.

John, V., 2024. *Morph Classifier*. MA thesis. Charles University.

Kosek, P., 2017. Periodizace vývoje češtiny. In: Karlík, P. et al. (eds.). *CzechEncy - Nový encyklopedický slovník češtiny*. Masarykova univerzita. Available also from: https://www.czechency.org/slovnik/PERIODIZACE%20V%C3%9DVOJE%20%C4%8CE%C5%A0TINY. Last accessed: 23. 2. 2025.

McCarthy, A. D. et al., 2020. UniMorph 3.0: Universal Morphology. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. European Language Resources Association (ELRA).

Nivre, J. et al., 2017. *Universal Dependencies 2.1* [https://universaldependencies.org/]. Accessed: 2025-04-10.

Olander, T., 2022. *The Indo-European Language Family*. Cambridge University Press.

Olbrich, M. et al., 2025. *DeriNet 2.3*. Available also from: http://hdl.handle.net/11234/1-5846. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rejzek, J., 2019. *Český etymologický slovník*. LEDA. No. 3.th edition. ISBN 978-80-7335-393-3. Available also from: https://leda.cz/Titul-detailni-info.php?i=623.

Rejzek, J. et al., 2025. *Czech Etymological Lexicon 1.0*. Available also from: http://hdl.handle.net/11234/1-5845. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Schachter, P.; Otanes, F., 1983. *Tagalog Reference Grammar*. University of California Press. California library reprint series. ISBN 9780520049437. Available also from: https://books.google.cz/books?id=E8tApLUNy94C.

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

Thomason, S., 2001. *Language Contact: An Introduction*. Edinburgh University Press. ISBN 9781474473125. Available also from: https://books.google.cz/books?id=XLZJzgEACAAJ.

Wiktionary, [n.d.]. *Wiktionary: The free dictionary* [https://www.wiktionary.org/]. Accessed: 2025-04-10.

List of Figures

2.1 Example of the Deconstructor web interface output 20

List of Tables

3.1 Size of the annotated dataset used for training, development, and testing 24

4.1 Top 10 most frequent etymological sequences in the training data 35
4.2 Number of unique language sequences and languages in the training data based on minimum frequency thresholds. 36
4.3 Most common etymological origin sequences in the extended training set (∼19,000 morphs) 37

5.1 Performance of baseline models. RER = Relative Error Reduction over the Dummy Baseline. Unique = Score grouped by morph surface forms. 39
5.2 Evaluation results of different models on the development set. All metrics are F1 scores in %, The first column is F1 averaged per each morph instance. RER is relative error reduction over the dummy baseline in %. See chapter 3.3 on evaluation methodology for more details. 39

List of Abbreviations

e.g. for example (from Latin *exempli gratia*)