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

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.
- Importance of studying the etymology of morphemes.
- 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)

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*.[1] 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*)

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 irregular plural forms, which we will discuss later.

[1] The word *respect* could be further divided into *re-* + *spect* if the segmentation is motivated etymologically.

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 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*).

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.

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 may exhibit characteristics of both. For example, *nej-* in Czech functions as an inflectional superlative marker but also resembles a derivational prefix in some contexts.

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 {Ø} 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*.

If someone unfamiliar with English saw the words *go* and *went* or *good* and *better*, they would probably not recognize them as related forms where one was derived from the other.

Another forms that break the concept of morphs being building blocks for word formation are so called *cranberry morphemes*. They are bound only to one concrete word. An example from which the name came is the English word *cranberry*, the morph *berry* is productive in many words (e.g. blueberry, blackberry, strawberry etc. in contrast *cran* is used only in the word *cranberry*. Another example of cranberry morphemes is *cobweb* where *cob* is only bound to the word cobweb.

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 history and development of words and their origins. It seeks to answer questions such as: “Where did this word come from?”, “How was it created?”, and “Is it a native word or a loanword from another language?”.

The curiosity about the origins of things resonates with people, making etymology a field of great interest to the general public. As a result, many instances of *folk etymology* exist, as people often speculate about the possible origins of words.

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 gain a deeper understanding of 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:[2]

  - **Western Romance**: Spanish, French, Portuguese, Catalan, Italian
  - **Eastern Romance**: Romanian, Dalmatian[3]

- **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:

[2] Lists are not exhaustive

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

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

- **Hellenic** Mainly Greek

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, Czech could not have borrowed words directly from Greek—by the time the Czech language was forming, the period of Greek cultural dominance had already passed.

Although Greek had little direct influence on Czech, it still left traces, mainly through Christianity. When Cyril and Methodius brought Christianity to Slavic lands in the 9th century, many Greek-origin words entered the language. Most of them came through Old Church Slavonic, which was used in religious texts and services and helped pass Greek words into Czech.

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 *virus* gained its modern meaning in the 20th century. 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 examples (and many more calques) are described in (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

- Overview of previous studies on morpheme etymology and morphological segmentation.
- Existing NLP approaches for analyzing word formation and origins.
- Datasets used in similar research.

2.1 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 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 contains 5,781 derivational trees with loanwords, enriched with etymological data from CzEtyL. Beyond word-formation structure, it provides detailed morphological segmentation and classification of morphs, making it a valuable resource for studying the relationship between derivation and etymology. DeriNet is particularly useful for this work, as it provides a structured way to study the connection between etymology and word formation. It allows tracing how words evolve in derivational networks, offering insights into the origins of lexemes and their morphological changes.

In many cases, derived words keep the etymology of their base form. If the origin of one lexeme is known, its derivational subtree will often share at least part of that origin. Borrowed words frequently become the base for new derivations with native affixes.

Even when a native affix is added to a borrowed stem, the entire word is still classified as a loanword. However, if we classify each morpheme separately, some parts would be loaned while others would be native.

2.2 Rejzek's Etymological Dictionary

The latest etymological dictionary of the Czech language by Rejzek (2019) includes over 11,000 core entries, covering both common vocabulary and newly adopted or ungrammatical words. Additionally, it features approximately 21,000 derived words, nearly 64,000 references to words from other languages, an overview of key spelling changes from the Indo-European proto-language to Czech, and a classification of cognate languages worldwide.

2.2.1 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 its 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-* (*adekvátní*), *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*)

2.2.2 SIGMORPHON Shared Task

SIGMORPHON (Special Interest Group on Computational Morphology and Phonology) regularly organizes shared tasks aimed at advancing research in morphological analysis. In my work, I use data from the *SIGMORPHON 2022 Shared Task on Morpheme Segmentation* **CITATION?**.

The dataset for this task was created by integrating multiple morphological resources, including UniMorph (for inflectional morphology), MorphyNet (for derivational morphology), Universal Dependencies, and ten editions of Wiktionary (for compounds and root words). **CITATION? From readme on github for the task**

I use the Czech portion of this dataset, which includes full sentence data with morph-level segmentation. These annotations serve as the basis for my own work.

I further manually annotated the morphs with etymological labels to create a dataset for morph-level etymology prediction.

2.3 Classification

In this section, we focus on the task of morphological classification. Once a word has been segmented into individual morphs, each morph can be assigned to a specific category. As discussed in Section 1.1.2, there are multiple ways to categorize morphs. For our purposes, the primary classification is the distinction between *roots* and *affixes*.

Affixes are further categorized based on their function-either *derivational* or *inflectional*-and their position relative to the root, such as *prefixes*, *suffixes*, or *interfixes*.

Morphological classification was performed on the data. Model from Vojtěch John was used.

describe it more, add citation

3 Practical Part

- Designing a model for predicting the origins of morphemes in the Czech language.

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, and its digital form, CzEtyL, 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.

Even though the lexicon does include a few examples of affixes with their etymological origins, a complete dataset with consistent morph-level annotation does not exist, so manual annotation is necessary for work at this level of detail.

In addition, many morphs are homonymous. For that reason, annotation needs to be done within actual sentence examples, not just on isolated morphs or words.

3.1.1 Annotations

Data from the SIGMORPHON 2022 Shared Task were used. This dataset consists of sentences segmented into individual morphs. The sentences were manually annotated, and each morph was assigned a sequence of language origins. The annotation was based primarily on information from the Czech Etymological Dictionary by Rejzek (2019).

Number of sentences: Train set (beginning of the sigmorfon ces.sentences.train dataset and few sentences from the end):

The annotated dataset was divided into three parts:

- **Training set**:
    - 200 sentences
    - 2,774 words
    - 7,016 morphs
- **Development set**:
    - 50 sentences
    - 599 words
    - 1,460 morphs
- **Test set**:
    - 50 sentences

  - 609 words
  - 1,485 morphs

Counting just morphs where we want to predict etymology. So excluding numbers, abbreviations, punctuation, and special characters.

The training data consists of approximately 170 sentences from the beginning and about 30 sentences from the end of the `ces.sentences.train.tsv` file.

The sentences for development and test sets were extracted from the `ces.sentences.dev.tsv` file, which contains 500 sentences in total. Every 10th sentence (1st, 11th, 21st, ...) was assigned to the test set, and every 10th sentence starting from the second (2nd, 12th, 22nd, ...) to the development set.

The remaining 400 sentences from `ces.sentences.dev.tsv` were not used in this work and are left available for potential future evaluation or testing. Additionally, the 500 sentences in `ces.sentences.test.tsv` were also left unused.

Example annotation: **Sentence:** *Faxu škodí především přetížené telefonní linky*

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

- **linky**
  - *lin* — `deu,lat`, `R`
  - *k* — `ces`, `D`
  - *y* — `ces`, `I`

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

3.2 Baselines

To determine how good the model is, we need to have some baselines so we have a bottom bound for comparison. Baselines help us evaluate whether the model actually learns something meaningful, or if its performance could be reached by a much simpler approach.

In this work, we define two main baselines: a trivial one that always predicts Czech, and a word-level etymology baseline that relies on CzEtyL, using root-based etymology along with a predefined list of borrowed affixes.

3.2.1 Baseline 0

A simple baseline is to always predict Czech as the origin for all morphs. This trivial approach achieves around 85% accuracy (depends a lot on the test set, change this number after performing measurements on the final test set XXX), which is naturally quite high, as most words and morphs in the dataset are native to the Czech language.

3.2.2 Roots approach

For a more advanced 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 list of known borrowed affixes, which is also included in CzEtyL. Additionally, all inflectional affixes are assumed to be native.

We iterate through all words in CzEtyL, extract their roots, and associate the full word-level etymology with the root. This results in a list of root morphs, each mapped to a multi-set of possible etymological origin sequences based on the words they appear in.

This simple algorithm works as follows:

- If a morph is classified as a root, it is assigned the most frequent origin sequence from its multi set of possible origins.
- If the morph is a derivational affix, we check whether it appears in the predefined list of borrowed affixes and assign its origin accordingly.
- If the morph is an inflectional affix, or if it is not found in the lexicon or the affix list, we fall back to predicting Czech, which is the most frequent class.

This results in around 90% accuracy (depends a lot on the test set - Change this number later after performing measurements on the final test set **XXX**) reducing the error rate by around 33%.

3.2.3 Word lemmatization approach

Using morphological analyzer. Get lemma for each word and look it up in the dictionary. Assign the retrieved etymology to the root. List of affixes etymologies. Endings are Czech. If a word is not in the Etymological dictionary, fall back to Czech.

3.2.4 Remembering approach

Just remember each morph from train data and predict the same etymology if appears in test (the most frequent if more different etymologies were for the same morph in train data). For types where train is big enough and the testing sentences have many words which appeared in train works good.

Problems for unknown words. (It predicts the most frequent class ("ces") in that case.

3.3 Evaluation

The target prediction for each morph is a sequence of languages from which (or through which) the morph was borrowed into Czech.

To evaluate the quality of predictions, we use the F1-score, which is the harmonic mean of precision and recall. Precision measures how many of the predicted languages are actually present in the target sequence, while recall measures how many languages from the true sequence were correctly predicted by the model.

Since the order of languages in the sequence is generally fixed or obvious, we treat the prediction as an unordered set. For example, if the correct languages are Latin and German, it is almost always the case that the morph came from Latin through German—not the other way around—so the order is not important for evaluation.

3.3.1 Relative error reduction

Because the baseline for this task is so high, absolute accuracy (or F1 scores) values 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 $1 - \text{F1-score}$.

3.3.2 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.

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.

Moreover, assigning etymological labels to individual morphs is not something traditional linguistics usually focuses on. It is often unclear whether to include only the original source language or also intermediate languages through which the borrowing occurred. In many cases, we may know the original language, but not the exact path the word took before entering the language.

Some words have been borrowed more than once or simultaneously from different sources, which makes the annotation even more difficult. These ambiguities show that the task is not strictly defined, and even the idea of a single “correct” answer can be sometimes questionable.

3.3.3 Inter-Annotator Agreement

Inter-annotator agreement measures how consistently different annotators label the same data. This kind of evaluation could be useful for assessing both the subjectivity and the reliability of the annotation process. Given the ambiguity and complexity involved in etymological annotation at the morph level, such a measure would help refine annotation guidelines, highlight unclear or borderline cases, and improve the overall quality and consistency of the dataset.

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.

Results of inter annotator agreement Or should it be in Results chapter?

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.

The evaluation included the calculation of the F-score and the percentage of exactly matched cases between the two annotators. Additionally, Cohen's kappa was computed to account for agreement occurring by chance. The results are as follows:

- F-score: 97.060%
- Exact match: 96.370%
- Cohen's kappa: 0.834

For comparison, we evaluated the agreement between the gold annotations (Annotator 1) and a baseline system producing dummy predictions (always predicting Czech for each morph). The results are:

- F-score: 89.714%
- Exact match: 89.714%
- Cohen's kappa: 0.288

While the F-score appears relatively high, this is largely due to the dominance of Czech morphs in the dataset. The low Cohen's kappa (0.288) reveals that the actual agreement beyond chance is weak, confirming the limitations of this baseline.

— **Model predictions vs gold data**

This won't be here; it will be in the results chapter or deleted.

Does it make sense to calculate Cohen's kappa for gold data and model predictions?

MLP Model, no embeddings, **Recalculate this in the end**

- F-score: 94.843 %
- Exact match: 93.392 %
- Cohen's kappa: 0.690

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

4 Model for Etymology Prediction

- Designing a model for predicting the origins of morphemes in the Czech language.

Here I will explain what approaches I tried in order to get the best possible model to predict the etymology of given morphs.

4.1 Features

What features did I extract when predicting the morph etymology?

4.1.1 Morph n-grams

From the text of the morph, we extract unigrams and bigrams. And use CountVectorizer to translate it to a sparse vector.

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 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 subword 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 morphologically rich languages or 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. The subword-based nature of FastText makes them especially suitable for representing individual morphs, including affixes and roots that do not appear as standalone words.

To reduce dimensionality and improve how effectively the classification model can learn from these features, the embeddings can optionally be compressed 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. The word embedding provides broader contextual information, which can help distinguish between allomorphs—morphs with the same form but different behavior depending on the word they appear in.

4.2 Classifier

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, specifically Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP).

5 Results

- How good was it?

5.1 Did we beat the baselines?

Yes! I hope so. On dev yes.

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.

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)*.

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

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.

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.

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

List of Figures

List of Tables

List of Abbreviations

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

A Attachments

A.1 First Attachment