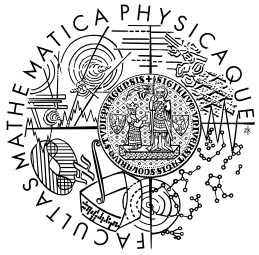

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Kristýna Neumannová

Identification and analysis of Czech equivalents of German compounds

Institute of Formal and Applied Linguistics

Supervisor of the bachelor thesis: Mgr. Magda Ševčíková, Ph.D.
Consultant of the bachelor thesis: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Study programme: Computer Science
Study branch: General Computer Science

Prague 2021

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

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

I would like to express my gratitude to my supervisor Mgr. Magda Ševčíková, Ph.D. and to my consultant doc. Ing. Zdeněk Žabokrtský, Ph.D. for supervising this thesis and giving me advice. I also would like to thank my family and friends for supporting me during my studies.

Title: Identification and analysis of Czech equivalents of German compounds

Author: Kristýna Neumannová

Institute: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Magda Ševčíková, Ph.D., Institute of Formal and Applied Linguistics

Consultant: doc. Ing. Zdeněk Žabokrtský, Ph.D., Institute of Formal and Applied Linguistics

Abstract: This bachelor thesis deals with automatic identification of Czech equivalents of German nominal compounds and their linguistic analysis. Compounding is a word formation process which is exploited in both languages, however, in German it is much more productive than in Czech, where the derivation word formation process predominates.

The first part of the thesis copes with identification of Czech counterparts of German compounds with the help of parallel corpora and tools for phrase-based statistical machine translation. After the identification, one-word, two-word and multi-word Czech equivalents were distinguished. The Czech equivalents were analysed according to their part-of-speech tags. Over a half of the German nominal compounds correspond to a sequence of two or more words in Czech, most of the sequences are made up of an adjective and a noun. Morphological structure of one-word equivalents was studied and these equivalents were distinguished into compounds and derivatives, in which the second part of the German compound corresponds to a suffix in the Czech counterpart.

Keywords: word formation, composition, derivation, morphology, syntax, natural language processing, parallel corpus, alignment

Název práce: Identifikace a analýza českých ekvivalentů německých kompozit

Autor: Kristýna Neumannová

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Magda Ševčíková, Ph.D., Ústav formální a aplikované lingvistiky

Konzultant bakalářské práce: doc. Ing. Zdeněk Žabokrtský, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Tato bakalářská práce se zabývá automatickou identifikací českých ekvivalentů německých substantivních kompozit. Skládání slov je doloženo v obou jazycích, ačkoliv v němčině je více produktivní než v češtině, kde převládá odvozování.

V první části práce jsme se věnovali identifikaci českých protějšků německých kompozit za pomoci paralelních korpusů a nástrojů na statistický strojový překlad založený na frázích. Poté jsme rozdělili české protějšky na jednoslovné, dvouslovné a víceslovné. České ekvivalenty byly analyzovány podle jejich slovnědruhového zařazení. Více než polovina německých substantivních kompozit odpovídá sekvenci dvou nebo více slov v češtině, většina sekvencí je tvořena přídavným a podstatným jménem. U jednoslovných ekvivalentů byla zkoumána jejich morfologická struktura a byla mezi nimi rozlišena kompozita a odvozená slova, kde druhá část německého kompozita odpovídá příponě v českém protějšku.

Klíčová slova: slovotvorba, skládání slov, odvozování, morfologie, syntax, zpracování přirozeného jazyka, paralelní korpus, zarovnání

Contents

1. Introduction

Compounding is a word formation process, where two or more bases or words are connected and create together a new word. Compounding is attested in many languages including German and Czech, which are in the focus of the present thesis. However, in German it is exploited much more frequently than in Czech. Compounds pose specific challenges to natural language processing – for example, to machine translation, because their equivalents in languages, where a different word formation process predominates, are mostly multi-word phrases or even dependent clauses. German compounds are therefore subject of many studies. However, there are not many papers concerning German compounds in relation to Czech word formation from linguistic point of view and the thesis is pioneering in terms of natural language processing of German compounds.

Although Czech language has been influenced by German for centuries, Czech is a Slavic language with inflectional and word formation features different from German, which belongs to Germanic languages. Even though composition is the second most important word formation process in Czech, derivation clearly predominates. The objective of this thesis is to study how German compounds are expressed in Czech. The experiments and analysis carried out in this thesis provides results that might be useful for further work with this language pair – for example, for translation or identification of compounds or multi-word phrases.

The main goal of this thesis is to analyse Czech counterparts of German nominal compounds, and with the help of the analysis, to find out the connections and differences between these languages. As the first step, Czech equivalents of German compounds must be identified, so the first goal of this thesis is an automatic identification of Czech equivalents of German compounds, which will be taken from an existing list. The identification will be designed on the basis of available corpora and implemented with the help of natural language processing tools. Czech counterparts of German compounds will be distinguished into one-word, two-word and multi-word phrases. The equivalents made up of more than one word will be classified according to the part-of-speech categories of their constituents. One-word equivalents will be studied for their morphological structure. Additionally, syntactic structure of phrases or frequency properties of Czech equivalents will be taken into account.

In what follows, the thesis is divided into six chapters. In Chapter 2, we present a short summary on the word formation processes in German and Czech in order to provide linguistic background to the issues modelled and studied in the thesis. In Chapter 3, we present works related to ours: we focus on cross-lingual studies (both linguistic and natural language processing) and on natural language processing of German compounds. All data sources used in our experiments are presented in Chapter 4. Since the practical part of the thesis consists of two tasks, each of them is described individually in a separate chapter. In Chapter 5, we describe the process of identification of Czech counterparts of German compounds including information about all used tools and description of the whole procedure of the experiments. The subsequent linguistic analysis is provided in Chapter 6 where each of the types of Czech equivalents is described separately in detail.

2. Linguistic background

In this chapter we provide basic information on German compounding and other word formation processes and a short summary of the word formation system of Czech.

2.1 German word formation

Compounding, derivation and conversion are distinguished as main word formation processes, but only compounding and derivation are relevant for our work and therefore described below in detail. German word formation has a lot of productive models of formation, especially for nouns. Bases of the models are stems or parts of words with phrases (for example, models for derivation of nouns are based on nominal stems and noun phrases). In addition, nouns have the largest number of suffixes. Formation of adjectives is very similar to nouns (mainly in compounding and derivation), in spite of the fact that there are fewer models for them. However, the formation of verbs is conducted differently. In case of verbs, conversion and expansion through prefixes and particles prevails [Barz, 2016, p. 2388].

Not only the models but also the means of German word formation are very multiform, one form can have more different meanings or, conversely, two or more different forms can have the same meaning. A set of means is multiform because not only the native affixes (*be-, -ung, -arm*) but also many non-native affixes in combination with non-native bases (*re-, -iv, -ion*) are used [Barz, 2016, p. 2388].

Word formation overlaps with syntax in German. Identical sequences can be found in syntactic phrases and lexemes. For instance, the adjective-verb construction in example 1 has properties not only of a syntactic phrase but also of a lexeme. In this case, both interpretations are acceptable. On the other hand, in noun-verb constructions we need to differentiate between them; if the noun is syntactically and semantically dependent, then it is considered as word formation (see example 2a) otherwise as a syntactic phrase (ex. 2b) [Barz, 2016, p. 2388–2389].

(1) 'to warm the soup'
   a. die Suppe warm stellen
   b. die Suppe warmstellen

(2) a. eislaufen 'to skate'
   b. Rad fahren 'to cycle'

2.1.1 Composition

As mentioned above, composition is typical of nouns and of adjectives and it is a frequently exploited word formation process in German. German compounds are the subject of interest in our thesis, therefore we describe the compounding process in more detail.

The majority of German nominal compounds has two nominal parts which are made out of either simplex (ex. 3) or complex stems (ex. 4). A stem is the part of the word which carries the lexical meaning and remains after removing all affixes of the word. Very complex compounds can be found in the technical sphere (example 6). However, most compounds consist of two or three stems (two simplex parts or one simplex and one complex nominal part with two stems – see examples 3 and 5). The main reason for that is the need of a speaker to speak briefly and plainly. The first part of the compound tends to be more complex than the second one, the opposite case does not occur very often. The first constituent can be made up of stems from different part-of-speech (hereafter POS) categories or phrases [Barz, 2016, p. 2390].

Typically, nominal compounds in German have the first element semantically dependent on the second one. The determined second element stands for the complex word as a hyperonym and the first element provides the exact specification of the second element (see example 3) [Barz, 2016, p. 2391].

(3) Haus—wand 'house wall' [1]

(4) Früh-jahrs—mündig-keit 'spring tiredness'

(5) Haus-tür—schloss 'front door lock'

(6) Lebens-mittel-farb-stoff-zulassungs—verordnung 'food coloring approval regulation'

Adjectival compounds consist mostly of constituents of the same word class (see example 7), but the first elements can also be nominal (example 8) or verbal stems (example 9) or seldom uninflectable words (ex. 10). Compounds with participle second element are often also considered as adjectival compounds (ex. 11). Superlative forms also appear in the first position of the compound (example 12). There is a limited set of adjectival elements that occur as the second compound part and combine with a high number of different first parts (*-arm* 'poor' (9), *-frei* 'free' (8), *-förmig* 'shaped') [Barz, 2016, p. 2392–2393].

(7) dunkel—blau 'dark blue'

(8) alkohol—frei 'alcohol-free'

(9) bügel—arm 'iron-free'

(10) vor—schnell 'hastily'

(11) hoch—begabt 'highly gifted'

(12) schnellst—möglich 'fastest possible'

German compounds are considered as right-headed. That means, the second element (head) determines the morphological and syntactic features of the resulting compound. The first part is morphologically and syntactically dependent on the head. In order to link both parts of the compound, semantically empty elements can be added (for nominal compounds *-e-, -en-, -es-, -ens-, -er-, -e-, -s-* and for verbal *-e-*) [Barz, 2016, p. 2390–2391].

[1] "—" is the delimiter between two constituents of German compounds
"-" marks the segmentation of the constituents

Three classes of compounds are distinguished – determinative, copulative and possessive compounds. The determinative compounds are the most common: they resemble syntactic phrases consisting of a governing item (captured by the second part of the compound) and an item depending on it (first part of the compound); the second part is semantically specified by the first part (ex. 13). Constituents of the copulative compounds are coordinated. Both concepts are attributed simultaneously to one individual (ex. 14 – *poet-composer* is someone who is both a poet and a composer). The possessive compounds have the same structure as the determinative ones, but they are used to express a possession (ex. 15) [Olsen, 2015, p. 3–4].

(13) Bierflasche 'beer bottle'

(14) Dichter-Componist 'poet-composer'

(15) Dickbauch '(person with a) fat belly'

2.1.2 Derivation

Derivation is the second most frequent word formation process after composition. Although we study German compounds in the thesis, derivation is also relevant for our work.

Barz differentiates between prefix derivation (see examples 16, 21 and 22), suffix derivation (examples 17, 18, 19 and 23) and circumfix derivation, which occurs in German only rarely (ex. 20). Prefixes are added mostly to complete words. On the other hand, suffixes combine with stems and phrases [Barz, 2016, p. 2395].

As mentioned, most affixes are used for derivation of nouns in comparison to adjectives, verbs and adverbs. Despite the fact that most of the affixes are polyfunctional, nominal derivation has the most models [Barz, 2016, p. 2395].

(16) Glück 'luck' → Un-glück 'bad luck'

(17) schön 'pretty' → Schön-heit prettiness'

The most often used affixes in derivation of adjectives are suffixes. New words are formed from verbal and nominal stems with the help of the suffixes *-bar, -ig, -isch, -lich, -mäßig* (see examples 18 and 19). Prefix derivation is done mainly with the negating prefix *un-* (ex. 20) [Barz, 2016, p. 2397].

(18) brennen 'to burn' → brenn-bar 'flammable'

(19) Öl 'oil' → öl-ig 'oily'

(20) un-aufhalt-sam 'unstoppable'

Verbal prefix derivatives can be divided into two classes. Verbs from the first class are morphologically and syntactically inseparable, formed with prefixes without homonymous particles *be-, ent-, er-, ver-, zer-* (21) and with prefixes with homonymous particles *durch-, über-, um-, unter-* (22). The second class are particle verbs where the particle carries word stress and the verbs are morphologically and syntactically separable. Suffix derivatives are formed with *-(e)l(n), -ie(en), -isier(en), -ifizier(en)* (ex. 23) [Barz, 2016, p. 2399–2400].

(21) grüßen ‘to greet’ → be-grüßen ‘to welcome’

(22) feucht ‘damp’ → durch-feuchten ‘to soak’

(23) Kanal ‘canal’ → kanal-isieren ‘to canalize (the river)’

2.2 Czech word formation

Derivation is far more important than composition in Czech word formation. Despite the fact, the combination of derivation and composition is not rare (ex. 24, a compound with a suffix *-ič*) [Bozděchová, 2016, p. 2875].

Derivational affixes add specific meanings to the base word, but also classify the output word into a particular inflectional class. The suffix *-tel* is used to form an agent noun which is a masculine animate noun in Czech (ex. 25). By adding the suffix *-ka*, the masculine noun turns into a feminine one (ex. 26) [Bozděchová, 2016, p. 2872].

(24) čas ‘time’ měřit ‘to measure’ → čas-o-měř-ič ‘timekeeper’

(25) učit ‘to teach’ → učitel ‘teacher’

(26) učitel ‘teacher’ → učitelka ‘teacher (fem.)’

Hundreds of suffixes can be found in nominal word formation. It provides possibilities to form verbal nouns. In Czech, there are on average ca. 31 derivatives sharing root [Bozděchová, 2016, p. 2875].

Czech is an inflectional language which implies that there exist a lot of morphonemic alternations (see examples 27, 28). There is a limitation on the influx of foreign words into Czech because of the typological profile of the language. The profile of the language prevents an excessive acceptance of ready-made naming from foreign languages. However, this property of Czech is weakened in the newest vocabulary. It is mainly caused by the extra-linguistic factors (internationalization of the vocabulary). [Bozděchová, 2016, p. 2875].

(27) vybrat ‘to choose’ → výběr ‘choice’

(28) šťastný ‘happy’ → štěstí ‘happines’

Derivation and composition can be clearly distinguished from multi-word expressions in Czech. The words in multi-word expressions are separated by spaces. Despite this, derivatives and compounds are written together (or sometimes with a hyphen). However, this does not hold for other languages, for example, for English. Constituents of the English compounds can also be separated by spaces (*car park*). Some of the multi-word expressions are used frequently as collocations. They consist mainly of two words linked through a determinative relationship. These are some commonly used combinations: A+N (ex. 29), N+N (30) or V+N (31) [Bozděchová, 2016, p. 2876].

(29) základní škola ‘elementary school’

(30) hod oštěpem ‘javelin throw’

(31) dávat pozor ‘to pay attention’

If some of the collocations are used frequently, a new one-word equivalent can be formed. This very productive process in Czech is called univerbation. Univerbation is related to language economy. The determinative constituent of the collocation together with a suffix form a new word (see example 32). The second constituent which carries the basic meaning of the collocation is omitted [Bozděchová, 2016, p. 2876].

(32) minerální voda → minerálka 'mineral water'

Lately, the productivity of composition, especially the nominal and adjectival composition, in Czech word formation has grown. There are several reasons for that, for example: the lexicon is internationalized, so some items from other languages are accepted and adopted. Speakers aspire to greater exactness of expressions (mainly in technical language) [Bozděchová, 2016, p. 2876].

2.2.1 Composition

Apart from derivation, composition is the second most frequent process in Czech word formation. It can be divided into proper (ex. 34) and synthetic composition (ex. 38). Synthetic composition combines affixes with the word stems and it is more frequent. There is often a linking vowel in both types of compounds (*-o-* or rarely *-i-, -e/ě-*, see examples 33, 34, 35 or 37). In composition, it is differentiated between determinative (see for example, 34, 35, 41 and 42) and copulative compounds (ex. 33 or 44) in Czech word formation [Bozděchová, 2016, p. 2877].

Nominal determinative compounds are formed in similar conceptual areas as derived nouns, they name people, means, actions, bearers or properties. The most common structures of determinative proper compounds are A+N (ex. 34), N+N (ex. 35), Num+N (36) and Pron+N (37). In the determinative synthetic compounds, neither the combination of the first two parts nor the second part are independent words (38). Copulative compounds occur mainly in specialized areas (ex. 33) [Bozděchová, 2016, p. 2877–2878].

(33) jih 'south' východ 'east' → jih-o-východ 'South-East'

(34) černý zem 'black ground' → čern-o-zem 'black soil'

(35) led 'ice' bořit 'to break' → led-o-borec 'ice breaker'

(36) dva 'two' hlas 'voice' → dvoj-hlas 'two-part singing'

(37) sám 'self' hláska 'phone' → sam-o-hláska 'vowel'

(38) román 'roman' psát 'to write' → roman-o-pisec 'novelist'

Adjectival compounds are more frequent than the nominal ones in contemporary Czech especially due to the formation of relational adjectives from multiword expressions (ex. 39). Most of the compounds are synthetic. Although there are more copulative compounds among adjectival compounds than in nominal composition, determinative compounds still prevail. Determinative adjectival compounds consist mostly of compounds which are formed from nominal and verbal stems, with structures A+N (see example 39), Num+N (ex. 40), N+V (ex. 41), Adv+V (42) or Pron+V (43). Copulative adjectival compounds

are mostly formed from stems of nouns and they can be written together or with a hyphen (ex. 44) [Bozděchová, 2016, p. 2879–2880].

(39) velký pán 'noble man' → velk-o-panský 'aristocratic'

(40) několik tisíc 'several thousand' → několika-tisícový 'several thousand'

(41) čas 'time' měřit 'meassure' → čas-o-měrný 'chronometric'

(42) nově 'newly' rodit 'to give birth' → nov-o-rozený 'newly-born'

(43) sám 'self' činit 'to act' → sam-o-činný 'self-acting'

(44) Labe 'Elbe' Odra 'Oder' → labsk-o-oderský (written labsko-oderský) 'Elbe-Oder'

2.2.2 Derivation

As already mentioned, derivation is the most productive process in Czech word formation. It can start from three different points – from complete words (example 45), stems of the basic words (ex. 46) or from morphological forms of the words (example 56). Some sound alternations of the word base often accompany the process (see examples 51 and 57) [Bozděchová, 2016, p. 2881].

The main procedure is suffixation. Some suffixes carry practically a constant function (*-ství* – abstract nouns, see ex. 47, *-dlo* – instruments), whilst others serve a lot of different functions [Bozděchová, 2016, p. 2882].

Denominal nouns can be formed with so called true prefixes (*ne-* – negation, see example 45, *pře-* – emphasis etc.). The prefixal-suffixal word formation is also very frequent for nouns (ex. 48). Suffixes can be used to derive personal (*-ař, -ák, -an, -ec*) and place nouns (*-iště, -na*, see ex. 50) as well as diminutives (*-ka, -ek*, see ex. 51) and collective nouns (*-stvo, -ctvo*, see ex. 52) [Bozděchová, 2016, p. 2882–2883].

Abstract nouns (*-ost, -ství, -ina*, see example 47) or attributive nouns (*-ec, -ík, -ka*, see ex. 54) can be formed from adjectives by suffixation. Deverbal nouns can be formed with the help of suffixes (*-ot, -ek, -el, -áč, -ent* etc., see ex. 55) as well [Bozděchová, 2016, p. 2883].

(45) pořádek 'order' → ne-pořádek 'disorder'

(46) půjčit 'to loan' → půjč-ka 'loan'

(47) bohatý 'rich' → bohat-ství 'fortune'

(48) hrdlo 'neck' → ná-hrdel-ník 'necklace'

(49) voda 'water'→ vod-ák 'paddler'

(50) oheň 'fire' → ohn-iště 'fireplace'

(51) lžíce 'spoon' → lžič-ka 'teaspoon'

(52) námořnický 'naval' → námořni-ctvo 'navy'

(53) moudrý 'wise' → moudr-ost 'wisdom'

(54) polední 'midday' → poledn-ík 'median'

(55) spát 'to sleep' → spán-ek 'sleeping'

Suffixation is the main means for adjectival derivation. Relational (ex. 57) and qualitative (ex. 58) adjectives are two semantic types of adjectival derivatives. Relational denominal adjectives are in relation to animate beings, inanimate objects, concepts and to materials. Qualitative adjectives describe features of a noun. Prefixation can be found in deadjectival derivatives (*ne-* – negative meaning, see ex. 58, *pre-* – intensification). Suffixes in adjectival derivation mostly form comparatives (see example 59) or express intensification and approximation [Bozděchová, 2016, p. 2884–2885].

(56) vyspěl 'he matured' → vyspěl-ý 'mature'

(57) sestra 'sister' → sestř-in 'sister's'

(58) hezký 'pretty'→ ne-hezký 'non-pretty'

(59) hloupý 'dull' → hloup-ější 'duller'

Different semantic classes of nouns, adjectives or verbs can be turned into verbs through derivation. Prefixes are mainly used for derivation from verbal stems (for example, *roz-, vy-, do-, o-, při-*, see ex. 60). The are two typical suffixes (*-ovat, -it*, see example 61) in verbal derivation. The first one is added mostly to foreign stems [Bozděchová, 2016, p. 2886].

(60) dělat 'to do' → do-dělat 'to finish'

(61) bilance 'balance' → bilanc-ovat 'to make up the balance'

3. Related work

In this chapter, we describe cross-linguistic studies (both linguistic and natural language processing) based on parallel data which compare Czech and German, eventually Czech and English, or more European languages. Furthermore, we provide a brief overview of papers and theses that deal with German compounds processing from the perspective of natural language processing (hereafter, NLP).

3.1 Cross-linguistic studies

We considered comparative or contrastive linguistic papers and theses focused mainly on word formation and compounds as well as translation studies about compounds. We also went over cross-linguistic NLP works including mainly machine translation.

3.1.1 Linguistic literature

Word formation – comparison

We found several linguistics works dealing with word formation or specifically with compounds which used parallel data and are related to our work. First, we concerned Czech-German studies. Each of the studies has a different point of view on parallel data and word formation.

Šemelík [2014] presents observations about word formation in German-Czech dictionaries based on his work on the Academic German-Czech dictionary (*Das große akademische Wörterbuch Deutsch-Tschechisch*). He focuses on external texts in the dictionary, on macro-structure forms (ordering of the lemmas of the dictionary according to certain principles) and on word formation elements and text segments in the dictionary where morphologically related vocabulary units are listed in different German-Czech dictionaries. He mentions German compounds as words that are difficult to translate into Czech and put Czech and German word formation in contrast – mainly because of German preference of polymorphematic compounds. He also describes a tendency of keeping transparent, lexicalized and idiomatized compounds apart in the dictionary.

Koprdová [2013] deals with anglicisms in German and Czech and she analyses journalistic texts about events in the EU for that purpose. She focuses on number of anglicisms in texts and compares the numbers in conclusion. She chose five articles written in Czech translated to German and five articles written in German translated to Czech as her dataset. Her attitude is in some way similar to ours, she works with parallel data, nevertheless she does not examine phrases which have the same meaning in both languages but specific words and their occurrences in texts.

The perspective of these theses is different from ours, although they study German and Czech word formation and put them into contrast, so we looked for more similar works about Czech-English word formation. English is just like German a Germanic language and therefore such works are also relevant to us.

We found three theses on similar topic that compare Czech and English word

formation. Levová [2012] describes processes that both languages share and also the specific ones for English and Czech and she examines whether they are based on the same principles. She selects two short articles from online blogs related to newspapers. She analyses manually the words in the articles and describes to which word formation process they belong to. She concentrates on derived, converted and compound words as well as on minor means of word formation. Differently to our analysis, she analyses whole texts and does it manually.

Ficenecová [2011] carries out a similar analysis to Levová [2012], but she focuses on nouns. Her dataset consists of 100 English nouns from a novel and their Czech equivalents from a Czech translation. Beside word formation processes, she also compares the number of nouns. The described word formation processes include affixation, conversion, compounding and minor word formation processes. Her work is in some way closer to ours because she analyzes nouns and focuses more on word distribution in both texts as well as POS of Czech equivalents of English compounds. On the other hand, she does not use big corpus data, only 100 words from a novel.

Šimková [2011] studies most frequent word formation processes in both languages and examines some specific features which occur either in only one of the languages or in both. She examines the productivity of the particular processes. Her approach is more comparative and less based on data-analysis.

Not only these theses from University of West Bohemia compare word formation processes in English and Czech. We found another article studying English and Czech word formation on sports terminologies. Cocca et al. [2015] deals with English influence on Czech in the specific sphere of sport. They distinguish three types of influence on the sport terminology – borrowed words from English, nativised and semantically modified terms under influence of English and translated terms that are in compliance with lexical standard of English. For the analysis of similarities between both languages and of the English influence on Czech terminology, they use data mainly from a modern English-Czech dictionary of sports terms which includes more than 13 thousand terms from 68 different sports from years 1927 to 2002. They also compare compounds in English and Czech in this sphere. Most of the compounds in both languages consist of two nouns where the first one is a modifier. English compounds are translated into Czech either as one-word terms, two-word phrasal terms or multi-word phrasal terms. They show that Czech language uses more periphrastic expressions (multi-word phrases or dependent clauses) than English.

Equivalents of German and English compounds

After works comparing all word formation processes in German and Czech or English and Czech, we also considered works specially about equivalents of German, eventually English compounds in Czech (or other European languages). These works are mainly theoretical with examples from parallel data, NLP works about compounds processing are described in Section 3.2.

Trachtová [2012] presents possibilities of Czech translation of German compounds. She uses data from a German-Czech dictionary including financial and economical vocabulary. She extracted German compounds starting with letter "A" and analyzed the types of Czech equivalents (adjectival attribute + noun, noun + genitive attribute, one-word equivalent, noun + prepositional phrase, pe-

riphrasis or other possibilities – for example, compound). Distribution of types presented in her work is very similar to ours in the analysis of Czech equivalents of German compounds. However, she carries out the analysis manually and only on small data and takes all German compounds into account – not only nominal compounds.

Hegerová [2009] analyses neologisms from the English novel Lolita by V. Nabokov and their equivalents in Czech and German translations with a focus on compounds. She presents examples of English compounds and their Czech or German equivalents. The most frequent type of compounds used by Nabokov are nominal compounds with pattern "sb + sb" (that means two independent nouns bounded together). Other types presented in the thesis are "sb/adj + -ed" (noun/adjective + word with suffix -ed), "sb + dvb sb-ing" (noun + deverbal substantive with suffix -ing) and compounds representing colours. Examples with different equivalents obtained from the translation of the novel are presented for each type. There are not unambiguous equivalents in Czech for English neologisms, because the neologisms are problematic to translate. She concludes that in Czech compounding is less productive than in English, so most of the English compounds are translated as phrases or dependent clauses in Czech.

Smutný et al. [2008] describe differences between lexical systems of English and Czech based on compound substantives. The study was done on 4,500 items from Czech translations of English literature and English translations of Czech literature. They analyzed English compounds and their Czech counterparts in order to determine whether each part of the compound corresponds to an element of its Czech equivalent. The compounds were divided into nine groups according to that. They concentrated more on differences between language communities on compounds data and not on their word formation processes.

We found works not only concerning two languages, but also a study concerning all European languages. Finkbeiner and Schlücker [2019] describe compounds and multi-word expressions from the morphological and lexical point of view and study the problem of their correspondence among languages. Beside that, they made a contrastive overview comparing German with other West Germanic languages, North Germanic languages, Romance languages, Slavic languages, Greek languages and Finno-Urgic languages. German is a West Germanic language as well as English or Dutch, therefore there are more similarities than differences described. The most interesting section for our work is the comparison between German and Slavic language genera which also includes Czech. However, the section includes examples from Russian and Polish. The predominant type of compounds discussed in the section is N+N.

3.1.2 Machine translation

Our work is not specialized in translation, however, we used several machine translation (MT) methods for automatic identification of Czech equivalents of German compound and machine translation between German and Czech is therefore also related to our work. We considered only machine translation between those languages in general in this section, further translation works specifying on compounds are in Section 3.2.2.

Kvapilíková [2020] concentrates on unsupervised machine translation that are MT methods trained on monolingual data applied to low-resource language pairs. She used data from German and Czech and demonstrated how monolingual models gain cross-lingual knowledge. Due to the fact that German-Czech is not a low-resource language pair, she had an opportunity to compare supervised and unsupervised methods. In order to train unsupervised model, she used tools for alignment of phrases as we did (her implementation relied on Moses – see Section 5.2.2). She also worked with parallel corpora from OPUS (see Section 4.2.1) for training the supervised benchmark model.

Bojar and Zeman [2014] present achievements within the project CzechMATE in statistical MT from English, German, Spanish and French into Czech. They discuss phrase-based translation methods on very large corpora and discuss their errors which is a part related to our work. Next to that, different evaluation methods are described. They compare English-Czech translation with translations from German, Spanish and French into Czech. Their findings are based on experiments with various MT systems (mostly provided by the Moses decoder – see Section 5.2.2). They used parallel data (also Europarl and News Commentary datasets – see Section 4.2.1) for translation models and monolingual data for target language models. In German-Czech translation, they mention German compounds which are difficult to translate. So, they decided to split them into individual stems during the data preprocessing which increased the performance of the translation methods.

3.2 Processing of German compounds

Apart from cross-linguistic (both linguistic-theoretical and NLP) approaches, also mono-lingual NLP of German compounds is related to our work. We describe papers and thesis about compound splitting – segmentation of compound parts, different works about compounds classification and also about translation and aligning of compounds.

3.2.1 Splitting and classification

Splitting is a subtask of compounds processing in NLP. The output of the task are separated constituents of the compounds – mostly a head and a modifying constituent. For further work with the compounds or analysis of them, it is essential to know their morphological structure – their constituents, therefore we describe several works about compound splitting that are relevant to us. We also present several classification tasks that are related to compound splitting. Each classification method determines its own classes and the process of classification is done in each task differently.

Henrich and Hinrichs [2011] present a compound splitter, which determines the immediate constituents of compounds. Their splitter was used to identify the compounds in German network and determine their constituents (see Section 4.1.1). In order to achieve results, they combined three different classifier systems. One splitter uses pattern matching for gathering of all potential modifiers and heads of the compounds (considering linking elements), the second classifier reverses the denominalization of the head constituents or splitting of all

the affixes and the last one is the modified ASV Toolbox compound splitter, whose results are further processed by interpolating GermnaNet's graph structure. After that, the combined hybrid compound splitter was created which takes also information about German derivation morphology, beside compounding knowledge, into account. Their method is based on resources with the exception of a small set of hand-crafted rules.

Krotova et al. [2020] developed a deep learning based approach of noun compound splitting and idiomatic compound detection. They used a dataset of 82 thousand nominal compounds from German WordNet (v.14.0 see Section 4.1.1) for training. For compound splitting they chose a set of recurrent neural network (RNN) models where each of them is a binary classifier which determines for each sub-word whether it is a split-position or not. From their perspective, a compound definition is sum of the meanings of its constituents only if the compound is non-idiomatic. The non-idiomatic compounds can not be literally translated using its parts after splitting. For the detection of the idiomatic compounds, they present a dataset of idiomatic and literal uses of German compounds nouns based on the GermaNet data. Only the most frequent compounds from GermaNet were selected and provided with definitions from Duden dictionary. The data were automatically annotated (and manually post-corrected) according to their constituents and their Duden definitions. According to the annotation scheme, compounds were classified into four classes, dependently on the position of the non-idiomatic constituents (both constituents, first or second constitient and none of the constituents were non-idiomatic). After that, they trained machine learning models to classify the compounds into two classes – idiomatic and non-idiomatic compounds (compounds with one non-idiomatic part were considered on the border).

Callow [2019] concentrates on linking elements and ways how to recognize which linking element should be used in connection with input words. He used corpus data from Tiger Corpus including 700 000 tokens. Only compounds were selected and split into parts by a dictionary-based algorithm. After that, only N+N compounds without hyphen were extracted – the resulting data contained 24 819 words. He used two RNN models and Naïve Bayes classification for linking element prediction. His task is related to compound splitting, however, it is not a typical classification task. He classifies the pairs of nouns into classes according to their linking element which will be used for creating a compound from them (none, *-e-, -en-, -er-, -n-, -s-*).

Hätty and Schulte im Walde [2018] focus on automatic identification of German compound terms and their understandability. The recognition of the terms can be a basis for further NLP tasks such as translation, which is relevant for our work. They defined fine-grained classes of termhood and framing and combined an identification and an understandability investigation to a classification task. With the help of the termhood classes, they predicted information about compounds (based on information about their constituents). Their data consist of 206 compounds from the semantic domain of cooking, which were manually extracted from cooking recipes. Unlike us, they only used small data and from a specific area to classify German compounds. Their model consists of four classes – "non-term" (not a domain term), "sim-term" (a term which is semantically related to the domain), "term" (a prototypical and understandable term of the domain)

and "spec-term" (a prototypical and non-understandable term of the domain). After splitting the compounds and computing their features, the classification is done by a RNN classifier.

3.2.2 Translation and alignment

We considered several NLP theses and papers about German compound translation and terms aligning. The translation is provided mainly into English, eventually into other languages – French, Italian and Spanish. This works are related to the first experimental part of our thesis, which is identification of equivalents of German compounds. However, all described works translate compounds into English or other languages, not into Czech.

Clematide et al. [2018] describe a word alignment gold standard for German nominal compounds and their multi-word translation equivalents in English, French, Italian and Spanish. They used data from the parallel Europarl corpus (see Section 4.2.1) extended with aligned speaker turns. They used automatic tools for tagging and word alignment. After that, they extracted nominal compounds from the German corpus based on their POS tag ("NN" - common noun) and computed segmentation of the lemmas with different types of boundaries (compounding, weak and derivation boundaries). The pre-alignments were done by Giza++, after that they were automatically filtered (only alignments directed to compounds without function words were allowed). The last step was a manual validation with the help of their text-based software.

Weller and Heid [2012] present an approach for the alignment of German nominal compounds with equivalent English terms using comparable corpora from technical domains. They try to relate German compounds to their translation in a list of English terms using a bilingual dictionary. English and German terms were extracted from corpora containing texts on wind energy an mechanics (in English only nominal phrases). German compounds from the data were split – the splitter was trained on the domain-specific corpus and on the Europarl corpus (see Section 4.2.1). Also data from the German-English dictionary were used. They compared two methods of compounds alignment. Firstly, they individually translated constituents of the compounds (only one-word translations) into English and recombined the translations and after that searched for matching English terms. In the second approach, they also considered the word order in both terms by using term-equivalent patterns for instance, N+N – N+N (equivalents in opposite order than origin term) or N+N – N+prep+N (where nouns and their equivalents are in the same order).

Stymne [2009] focuses on merging strategies for translation of German compounds. The compounds are split into parts, constituents are translated individually into English and then merged. She compares eight different merging strategies. There are three types of merging strategies – the first one is based only on external knowledge sources (for example, frequency lists of words or compounds), strategies of the second type use symbols to guide merging (inspired by morphology merging) and the last type is based on POS tags. The system was trained on the Europarl corpus (see Section 4.2.1).

4. Data resources used in the experiments

In this chapter, we describe the data sources used in our experiments – namely, German compounds data used as the input data to our experiments, data from parallel corpora used for identification of Czech equivalents, and some other monolingual sources, which supported our experiments and analysis.

4.1 Resources of German compounds

Our work on the topic began with a selection of a suitable set of German compounds. We considered the size of the dataset, information about splitting of the compounds and additional information about their structure. We describe the two selected datasets below.

4.1.1 GermaNet

GermaNet is a lexical-semantic net which groups German nouns, verbs and adjectives with respect to their meaning. Units that express the same concept are grouped into synsets. Semantic relations are defined between the units in the synsets. We used GermaNet v15.0 (last updated 15.06.2020), which includes 99,094 split nominal compounds (Henrich and Hinrichs [2011]). Each of them is split into two parts: into a head and a modifier (first part of the compound). Both parts can be split further recursively, but GermNet does not contain compounds made of more different stems (*Brenn-stoff-lagerung-s-behälter* – four stems that can not be divided into two parts). All parts are lemmatized and, in the case an ambiguous modifier occurs, both possibilities are listed (separated by "|").

We found out that there are several duplicate values in the dataset: one total duplicate (*Laufwerk*), which was deleted, 18 compounds with two different splitting possibilities (see examples 62, 63), that were reduced to a single occurrence, and 81 duplicit values in the second and the third column (see example 64 where the compounds differ only in the presence of linking *-s-*), which remained in the dataset. After that, the dataset consisted of 99,075 nominal compounds.

The compounds included in the dataset are two- or more-stem compounds (N+N, V+N etc.), but also words formed from prepositions and nouns (for example, *Mit-glied* 'member', *Vor-teil* 'advantage').

(62)	a.	Blütenhüllblatt	Blüte	Hüllblatt
	b.	Blütenhüllblatt	Blütenhülle	Blatt
(63)	a.	Grundlinie	Grund	Linie
	b.	Grundlinie	grund	Linie
(64)	a.	Abfahrtgleis	Abfahrt	Gleis
	b.	Abfahrt**s**gleis	Abfahrt	Gleis

If we consider only unique lemmas and analyse modifiers and heads of compounds, we can find 2,053 compounds between 10,065 different heads and 2,921 compound words between 13,796 different modifiers, which can be recursively split.

4.1.2 CELEX

The CELEX Lexical Database is a database of Dutch Centre for Lexical Information, that includes data from three languages – English, German and Dutch (Baayen et al. [1993]). For each language, the dataset contains information on ortography, phonology, morphology, syntax and word frequency. The data are distributed to folders according to languages and for each language, each type of information is stored in separate subdirectory. The data are availible in two formats of tokens – lemmas or word forms. We used data from German morphology in format of lemmas (file *gml.cd* in identically named subdirectory – German Morphology Lemmas).

There are words of different POS categories and not just compounds. Therefore, we needed to extract just the nominal compounds. The whole German Morphology dataset contains 51,728 lemmas with 20 fields including ID, lemma and other pieces of morphological-structural information.

For the experiments, we extracted only several features from 20 columns for better orientation:

- 2 – Head (Lemma)
- 7 – Comp (Compound: Y/N)
- 9 – Imm (Immediate segmentation)
- 10 – ImmClass (POS of elements)
- 14 – StrucLab (Complete hierarchical analysis)
- 19 – InflPar (Inflectional paradigm)

In order to filter only the nominal compounds, we chose only lines which met following conditions:

- had "Y" in the Compound column
- the Immediate segmentation column was not empty
- had at least two non-affixal stems in POS of elements
- were nouns that means the column of the Inflectional paradigm had value like S_/P_ (for example, S1/P2)

After that, we gained 12,476 nominal compounds.

If we compare the sets of nominal compounds extracted from GermaNet (99,075 words) and CELEX (12 476 words), they overlap in 9,070 words. As the GermaNet dataset is much bigger and there are not much extra words in CELEX, we decided to use data from GermaNet for the automatic processing, although CELEX contains more pieces of information about the structure of the compounds. However, we found out that the details about the structure are not so relevant to us mainly because they are not consistent. We used CELEX data only for manually created selection dataset (in Section 5.1).

4.2 Parallel data

Parallel corpora consist of data from two or more languages that correspond to each other (either one is a translation of the second, or both of them are translations of a particular original text). That means that the first sentence in one language corresponds with the first sentence of the second language (data are stored for language pairs). The data can be gained from different domains of sources – mainly from translated texts.

4.2.1 OPUS corpus

OPUS corpus is a growing language resource of parallel subcorpora (Tiedemann [2012]). The project is focused on providing freely available data in various formats with basic annotation. In OPUS corpus, there are over 100 language pairs available. The subcorpora of OPUS corpus consist mainly of legislative and administrative texts (mostly from the European Union), translated movie subtitles and data from open-source software projects. Also newspaper texts and collections from various online sources represent a fundamental part in OPUS domains.

We used Czech-German data (cs-de) from 8 different subcorpora: OpenSubtitles (v2018), JRC-Acquis (v3.0), WikiMatrix (v1), DGT (v2019), EUbookshop (v2), Europarl (v8), EMEa (v3) and News-Commentary (v14). Their sizes are compared in Table 4.1.[1] We downloaded the data in Moses format. It provides sentence aligned data from both languages separately.

OpenSubtitles is a collection of translated movie subtitles.[2] It is a multilingual parallel corpus, which includes data from over 50 languages. WikiMatrix is compiled only from Wikimedia by Facebook Research.[3] The dataset consists of parallel sentences from 85 language mutations of Wikipedia.

DGT is a collection of translation memories. It is extracted from European Union's legislative documents in 24 EU languages.[4]JRC-Acquis also consists of legislative texts of the European Union[4] that were written between the 1950s and now. Another source is a corpus of documents from the EU bookshop. Next to that, Europarl was extracted from European Parliament web site in order to get more data for statistical translation research[5] and EMEa is made out of PDF documents from the European Medicines Agency.[6] The documents were firstly converted from PDF to plain text. The corpus consists of data from 22 languages.

News-Commentary is the only source based on newspaper articles. It is also intended for training of statistical machine translation tools. It provides data from 15 languages.

[1] `https://opus.nlpl.eu/`
[2] `http://www.opensubtitles.org/`
[3] `https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix`
[4] `https://ec.europa.eu/jrc/en/language-technologies/`
[5] `http://www.statmt.org/europarl`
[6] `http://www.emea.europa.eu/`

corpus	doc's	sent's	cs tokens	de tokens
WikiMatrix	1	1.6M	106.0M	443.1M
OpenSubtitles	21,805	18.0M	119.2M	134.2M
DGT	38,187	5.0M	93.9M	92.3M
JRC-Acquis	19,801	1.2M	55.5M	55.6M
EUbookshop	1,153	0.4M	15.3M	15.5M
Europarl	8,766	0.6M	13.6M	15.2M
EMEA	1,915	1.1M	14.2M	11.2M
News-Commentary	5,184	0.2M	4.8M	5.1M
total	**96,812**	**28.1M**	**422.5M**	**772.2M**

Table 4.1: OPUS corpora – sizes

4.2.2 InterCorp

InterCorp is a parallel corpus including texts in Czech and other 27 languages (Čermák and Rosen [2012]). We accessed it via the searching interface KonText[7] which allows us to search the corpus by simple or complex queries (see example of query in Figure 4.1), to browse the results as concordance rows, to count frequency distribution and to perform other task on the data. We used the German InterCorp (v13), which includes 107 million tokens (lemmatized to 956 thousand different lemmas) together with the Czech InterCorp (v13).

4.3 Monolingual data

4.3.1 Araneum Germanicum

Another resource we used in our experiments as a source of German monolingual data is the corpus Araneum Germanicum Maius (version 15.02 – Benko [2014]). It is a corpus from Aranea, a family of comparable gigaword web corpora. The corpus contains 1.2 billion tokens (lemmatized to 12 million different lemmas) and it is accessible via KonText.[7]

4.3.2 Czech National Corpus

The Czech National Corpus (CNC) provides the corpus SYN2020 released in December 2020 (Křen et al. [2020]). It contains 121 million tokens (100 million without punctuation, lemmatized to 726 thousand different lemmas) and it is also accessible via KonText[7]. We used this corpus for checking of phrase collocability therefore we downloaded bigrams data with the structure of N+N, A+N and N+A ordered by their frequency (see the query used for A+N in Figure 4.1).

[7] `https://www.korpus.cz/kontext`

Figure 4.1: A query for searching A+N bigrams in the SYN2020 corpus by using the KonText online tool

5. Experiments in identifying Czech counterparts

In this chapter, we describe experiments that we proposed and implemented for the identification of Czech equivalents of German compounds. Before starting the experiments, we selected a subset of data and created a manually annotated dataset of German compounds and their Czech equivalents in order to gain a better insight into the phenomena to be modeled (Section 5.1).

For the automatic identification of Czech equivalents of German compounds, we prepared the data from parallel corpora, extracted phrases from them (see Section 5.2), and then created an algorithm for the selection of Czech equivalents (see Section 5.3). We used resources of German compounds as the input data and selected correct Czech equivalents with the help of the obtained phrase table (translations of extracted phrases). We present all the tools and methods used in the experiments below.

5.1 Our hand-annotated data

Before we started with automatic identification, we manually created a dataset from a selection of German compounds that was later used as a referential dataset.

The dataset consists of 150 words and their Czech equivalents – 50 of them were selected manually from the intersection of the GermaNet list of compounds and the CELEX nominal compounds in order to choose some examples from each category (POS structure of compounds) and from different frequency bands (frequency extracted from the InterCorp) and the remaining 100 compounds were selected randomly from the whole GermaNet data.

Czech equivalents were determined manually with the help of the InterCorp corpus and the Czech-German dictionary provided by Google.[1] When multiple acceptable Czech translations were found for a single German compound, all of them were stored. For each equivalent, its type and morphological structure were given (see Table 5.1). We differentiated between several categories of Czech counterparts:

- two-word collocations of an adjective and a noun (A+N, see example 65)
- two-word collocations of two nouns (N+N, see 66)
- phrases made up of more than two words, at this point without given POS structure (Multi-word, see example 67)
- one-word equivalents, which were further divided into:
  - compounds (ex. 68)
  - words where the second constituent of the German compound corresponds with a suffix in Czech counterpart (Particular suffix, see ex. 69)

[1] `https://translate.google.cz/`

– independent Czech words, in which no part corresponded exactly with any constituent of the particular German compound (Word, see ex. 70)

(65) Finanzkrise – finanční krize ‘financial crisis’

(66) Sodbrennen – pálení žáhy ‘heartburn’

(67) Pflanzenschutzmittel – přípravek na ochranu rostlin ‘plant protection product’

(68) Gleichberechtigung – rovno-právnost ‘equality of rights’

(69) Königreich – králov-ství ‘kingdom’

(70) Freitag – pátek ‘friday’

Compound	**Freq.**	**Czech equiv.**	**Type**	**Segmentation**
Forschungsrat	82	rada pro výzkum	Multi-word	rad/a/ pro/ vý/zkum
Sodbrennen	40	pálení žáhy	N+N	pál/ení/ žáh/y
Gleichgewicht	2996	rovnováha	Compound	rovn/o/váh/a
Freitag	1562	pátek	Word	pát/ek

Table 5.1: Hand-annotated dataset – example

For each German compound, the most common equivalent was chosen and classified into one of the categories. Figure 5.1 shows the distribution of types in the dataset according to this classification.

Most of the Czech equivalents were bigrams with the structures A+N and N+N (57%, see Figure 5.1). Other 15% of the equivalents were phrases made up of three or more words. Only 7% of the German compounds were translated to Czech as a compound. Based on this distribution, we got an estimate about the results of automatic experiments with bigger data. It seems that most of the German compounds will be translated to Czech by two- or more-word equivalents and that only few percents of the German compounds will have compositional counterparts in Czech.

For further steps, all the Czech equivalents (the column Type, see Table 5.1) were lemmatized with the UDPipe toolkit.

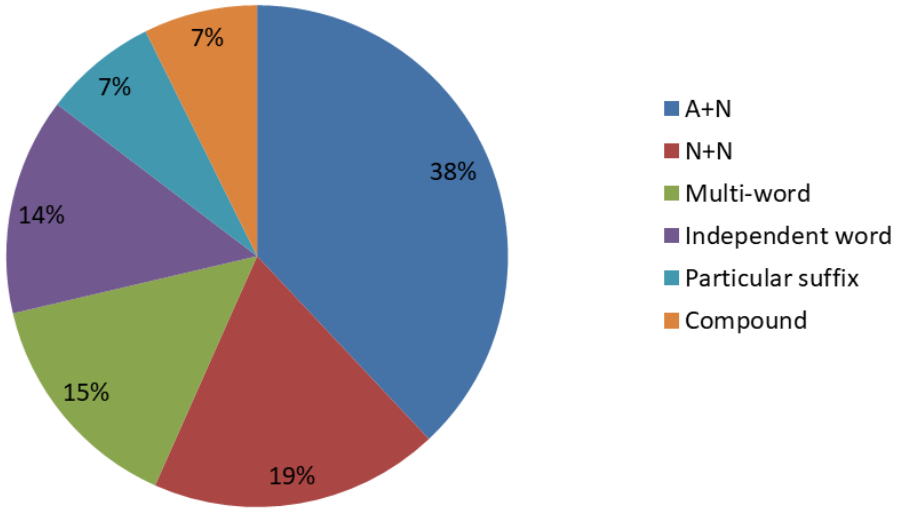

Figure 5.1: Distribution of categories of Czech counterparts in the referential dataset formed by 150 selected German compounds

5.2 Data preprocessing

We used several tools for the preprocessing of the parallel data. For this purpose, the data were lemmatized and tokenized, after that the word alignment was done to identify semantic relations between Czech and German sentences, and the last part of the preprocessing was building a German-Czech phrase table, which is an important input for the subsequent selection of Czech counterparts of German compounds. All these steps were provided automatically by NLP tools, which are described below.

5.2.1 Lemmatisation and tagging by the UDPipe tool

Lemmatization is an NLP task that takes a text with inflected forms as an input and returns lemmas (canonical forms) of the words from text. During the lemmatization process, the input words are marked with information about their grammatical properties (so-called tag), the process is also known as tagging and is mostly done together with lemmatization.

We used the UDPipe tool (Straka et al. [2019]) for the lemmatization and tagging of our data. It is a trainable pipeline for tagging, lemmatization and syntactic analysis of CoNLL-U input (see example in Figure 5.2).[2] Pre-trained models for the toolkit are available online.[3]

After downloading the parallel data from the OPUS corpus (see Section 4.2.1) tagging and lemmatization was provided by this toolkit. The UDPipe toolkit was used for each part of each corpus from the dataset separately (Czech and German

[2] `https://universaldependencies.org/format.html`
[3] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131`

data). The models were downloaded from the web.[3] The output format (*.conllu*) of this toolkit contains several features, such as the original word, lemma, POS, tag etc., for each word numbered in each sentence separately. For each sentence, there is its ID, text followed by all words with the features (see Figure 5.2).

```
# sent_id = 1
# text = Zlato za 10 000 dolarù?
1         Zlato     zlato     NOUN      NNNS1-----A----     Case=Nom|Gender=Neut|Number=Sing
2         za        za        ADP       RR--4----------     AdpType=Prep|Case=Acc          _
3         10        10        NUM       C=-------------     NumForm=Digit|NumType=Card     _
4         000       000       NUM       C=-------------     NumForm=Digit|NumType=Card     _
5         dolarù    dolar     NOUN      NNIP2-----A----     Animacy=Inan|Case=Gen|Gender=Masc
6         ?         ?         PUNCT     Z:-------------     _         _         _         _
```

Figure 5.2: Output of the UDPipe toolkit in CoNLL-U format – example

We extracted only columns 3 (lemma), 4 (POS) and 5 (tag). After that we selected only the POS categories we are interested in (adjectives, adverbs, nouns, verbs, pronouns, prepositions, numbers and particles) and rewrote lemmas from each sentence into one line for further processing with Moses (see Section 5.2.2).

Some modifications of the Czech output data were done – empty lines and extra lines with sentences IDs and text information, unimportant tokens (punctuation, conjunctions etc.) were deleted in order to downsize the data. In order to keep information about tag and POS of Czech data, all files (from all subcorpora of OPUS dataset) with the three extracted columns were merged into one file. Specification of proper nouns is not needed for our purpose, so we substituted all "PROPN" marks (proper noun) with "NOUN".

5.2.2 Data processing by the Moses toolkit

Moses is an open source toolkit for statistical machine translation based on phrases (Koehn et al. [2007]). It is a very complex toolkit and we used it for tokenization, lowercasting or truecasting (conditionally upon the corpus), word alignment and phrase extraction and scoring. Other functionalities provided by the toolkit such as the reordering model or evaluation are redundant for our present task. Although the data processed by the UDPipe tool (above in Section 5.2.1) were already tokenized, we used Moses to retokenize the data in order to make Moses run properly.

The Moses toolkit contains a lot of different scripts and options for training. In order to launch all needed scripts at once, we used a generic script (*experiment.perl*) with a configuration file, where we specified all paths to the files and scripts and the training options. For our experiment, German is assigned as the source language and Czech as the target language.

For each OPUS subcorpus, a working directory for Moses was created. The folder contained only a configuration file at the beginning. All the prepared data sources (lemmatized with UDPipe) were stored in one directory and the path to them was specified in the configuration files.

After running the Moses script from the command line, the *steps* directory was created where the progress of individual steps of the script with standard error output and standard output are stored (each of them in a separate file).

This directory is useful when the experiment crashes, because it can be launched again and continued from the last successful step. During the run of the experiment script, more directories were created in the working directory (*training, corpus, model* etc.), which contained results of the individual phases of the experiment.

Word alignment

Word alignment is an NLP task that identifies the relation between words in sentences in two languages. The output of the task are tuples of indexes of words in sentences that say which word in the first sentence is aligned to which word in the second sentence (both sentences are supposed to correlate with each other).

Word alignment is a part of the Moses experiment script. Moses provides several tools for this task (e.g. Giza++, which is very popular), out of which FastAlign was chosen for its speed, simplicity and better results (as reported by Dyer et al. [2013]). This toolkit must be downloaded and installed separately and its path and options specified in the configuration files for the Moses script.

The alignment was processed separately for the direct and the inverse direction (from the source to the target language and backwards) and, after that, the Moses script for symmetrization was used. The symmetrization was done with the *grow-diag-final-and* method. The resulting file contains only alignments (for example, *0-0*) of words in German and Czech sentences, the identification of the exact words (not only indices of them) had to be done afterwards.

Phrase extraction and scoring

In order to build a phrase translation table for the German and Czech language pair, phrases (one to five subsequent words) had to be extracted and scored. Phrases were extracted based on word alignments (it helps to discover relations between words that follow one another in data) and stored in a table (see Figure 5.3). After that, the extracted phrases were scored – a maximum likelihood lexical translation table for both direction was estimated, next to the distribution of the probabilities, phrase lexical weighting was computed and, after that, it was included in translation probabilities.

The results of the phrase extraction and the scoring steps were stored into the file *phrase-table*. In that file, all phrases and their equivalent phrases were listed. The file was sorted according to the source phrase, so that all translations of one source phrase were next to each other.

Each line of the phrase table contains a source phrase, a target phrase, scores (inverse phrase translation probability, inverse lexical weighting, direct phrase translation probability, direct lexical weighting), counts (count of inverse translations, count of source phrase, count of direct translations), word alignment of the source and the target phrase (see Figure 5.4).

```
Beschluß 3500 zusätzlich ||| další 3500 ||| 2-0 1-1
Beschluß 3500 ||| 3500 ||| 1-0
Beschluß Gericht ||| americký soudní ||| 1-0 1-1
Beschluß Sicherheitsrat Folge leisten ||| podvolit rezoluce rada bezpečnost ||| 0-0 1-1 1-2 1-3 3-3
Beschluß Sicherheitsrat ||| rozhodnutí rada bezpečnost z ||| 0-0 1-1 1-2
Beschluß Sicherheitsrat ||| rozhodnutí rada bezpečnost ||| 0-0 1-1 1-2
```

Figure 5.3: Table of extracted phrases – example

```
Busstation ||| autobusový nádraží vládní ||| 0.0511543 0.196429 0.0255771 0.0001154 ||| 0-0 0-1
↪  ||| 1 2 1
Busstation ||| autobusový nádraží ||| 0.0511543 0.196429 0.0255771 0.25 ||| 0-0 0-1 ||| 1 2 1
Bussystem ||| autobus silně omezit ||| 0.0511543 0.0204082 0.0170514 4.5855e-08 ||| 0-0 ||| 1 3 1
Bussystem ||| autobus silně ||| 0.0511543 0.0204082 0.0170514 0.0002468 ||| 0-0 ||| 1 3 1
Bussystem ||| autobus ||| 0.00138255 0.0204082 0.0170514 1 ||| 0-0 ||| 37 3 1
```

Figure 5.4: Phrase table with scores – example

We extracted only one-word German phrases and their equivalents. In each subcorpus of OPUS, we searched for compounds from GermaNet. The resulting counts of all one-word phrases and compounds found in the data are displayed in Table 5.2. After extracting one-word German phrases from the phrase table, we saved them as a dictionary (keys are German words and values are dictionaries containing Czech equivalents and its scores). The dictionary was dumped into a file for further work.

Subcorpus	**Count of one-word phrases**	**Count of compounds in subcorpus**
DGT	486 799	22 329
OpenSubtitles	478 021	30 496
JRC-Acquis	171 387	11 553
WikiMatrix	129 467	14 120
EUbookshop	123 023	9 723
Europarl	76 249	10 732
News-Commentary	67 175	7 812
EMEA	62 930	2 495
Unification		**51 144**

Table 5.2: OPUS subcorpora – counts of extracted one-word phrases and of found nominal compounds from GermaNet

5.3 Identification of Czech equivalents

Identification of Czech equivalents is based on the phrase table (see Section 5.2.2). In this section, we describe the process of identification in detail.

The extraction of phrases was done for each OPUS subcorpus independently and each of the dictionaries was also separately saved. We tried to get the best results using direct and inverse translation scores and, after that, also with the help

of POS categories of Czech words. First, we processed each subcorpus separately. We used our referential dataset for measuring an accuracy of the reached results. After that, we merged all the phrase dictionaries into one and searched for Czech equivalents of all of the nominal compounds from the GermaNet dataset.

5.3.1 Comparing the OPUS subcorpora

We compared the identified Czech equivalents obtained from all OPUS subcorpora. German compounds for identification were taken from our referential dataset. Therefore, we also had the targets and were able to measure the reached accuracy for each OPUS subcorpus.

If we considered only direct translation probabilities (based on the scores in the phrase table) and chose the best equivalents according to it, the results from the OPUS subcorpora differed substantially (see Table 5.3). We counted correctly and half-correctly (at least one part was correct) selected equivalents and measured two different accuracy metrics – one of them was counted only from the compounds that were found in the OPUS subcorpus (the count for each subcorpus is also listed in the table), and the other from all input compounds (150). We summed the count of correctly identified equivalents with 0.5 times the count of the half-correctly identified equivalents and divided the whole sum by the total count of compounds (either count of compounds that were in the subcorpora or with the total count of all input compounds).

From the OPUS subcorpora, the biggest number of compounds from our referential dataset was found in the OpenSubtitles corpus. The second best corpus was the DGT. With data from these two corpora were also achieved the best results – 25% accuracy for the OpenSubtitels corpus and 23% accurace for the DGT corpus (see Table 5.3). The worst results of the identification of Czech counterparts had the EMEA subcorpus (only 6% accuracy from all 150 compounds).

Subcorpus	**Count of compounds in subcorpus**	**Accuracy from comp. in subcorpus**	**Accuracy from all comp.**
OpenSubtitles	68	55,9%	25,3%
JRC-Acquis	40	65,0%	17,3%
WikiMatrix	43	57,0%	16,3%
DGT	61	56,6%	23,0%
EUbookshop	38	60,5%	15,3%
Europarl	42	69,0%	19,3%
EMEA	42	22,6%	6,3%
News-Commentary	33	72,7%	16,0%

Table 5.3: Comparison of OPUS subcorpora – compounds from the referential dataset and accuracy of their equivalents selection

5.3.2 Merging the phrase dictionaries

Merging the phrase dictionaries was an important step, because we needed all data together to be able to identify the best possible Czech equivalents.

For merging all phrase dictionaries into one, we abandoned the lexical weighting and considered only the translation probability in the direct and in the inverse way. The direct translation probability could be easily counted from number of occurrences of translations in the phrase tables (occurrences of each translation/equivalent was counted and divided by the sum of the occurrences of the source phrase). The inverse lexical translations probabilities were reached by weighted average (by counts of Czech equivalents) from inverse probabilities in all dictionaries.

We went through all phrase dictionaries and for each German word processed each equivalent. We looked for each equivalent of all phrases in all dictionaries if it was not already in the resulting dictionary (each word and all its equivalents were considered only once even if they were listed in more dictionaries).

5.3.3 Final result

Finally, we loaded the merged dictionary and created a scoring policy using our hand-annotated dataset. In the resulting policy, we selected always the equivalent with the best translation probability. If there were more equivalents with the same value, further scoring was made.

In order to be able to exploit the POS category for the scoring policy, we processed the prepared file of all lemmatized sentences from the OPUS corpus (see Section 5.2.1) and for each lemma, we chose the most frequent POS and stored all lemmas and their POS information to a dictionary in order to search it simply and quickly.

We considered the following factors that helped us to prefer one of the candidates with the same direct translation probability:

- the inverse translation probability
- the fact that the candidate included at least one noun
- number of words that were frequent in the list of candidates divided by the count of words in the current candidate
- the fact that the equivalent fitted into one of the most frequent categories (A+N, N+N, Multi-word with preposition in between)

According to these scoring factors (each factor was scored and the resulting score was counted as a sum of these scores), the equivalent with the highest score was chosen.

The scoring policy was based on our hand-annotated dataset. We used lemmatized Czech equivalents (see Section 5.1) to evaluate the accuracy of the result. We compared the chosen equivalents with the dataset. If one of the possibilities in the dataset matched with the selected phrase, it was counted as a correctly selected equivalent (plus one to the count of correctly identified counterparts). If only one part of the resulting phrase matched, one 0.5 was added to the count of correctly selected equivalents. The accuracy (count of correctly selected equivalents divided by the count of compounds found in OPUS subcorpora) achieved on our dataset was 63.9% (55 correct, and 14 half-correct from 97 compounds found in OPUS) and the overall accuracy (the count of correctly identified counterparts divided by the count of all compounds – 150) was 41.3%.

6. Analysis of the results

Once the selection algorithm was tuned, Czech equivalents of the German nominal compounds from GermaNet were selected from the OPUS-based phrase dictionary according to our scoring policy (described in Section 5.3.3). 51,144 out of 99,075 compounds were recognized and translated.

In this chapter, we analyze the results obtained. Firstly, we compared the OPUS dataset to the monolingual data from the Araneum corpus. Then we distinguished one-word, two-word and more-word Czech equivalents of GermaNet compounds. The equivalents were further divided into several classes according to POS categories of their parts. Some of the types (mostly the N and A categories) were reviewed and the equivalents from these types were redistributed into other categories. We made analysis for each category separately. The final analysis is provided in the last section of this chapter.

6.1 Distribution of compounds in the sources

As mentioned above, 51% of the nominal compounds listed in GermaNet were found in the OPUS. We were curious about the reason why almost a half of the German compounds was not attested in the OPUS corpus. One hypothesis was that their presence in the corpus is correlated to the frequency of the compounds. We tried to find out whether the compounds that were not found in the OPUS subcorpora have also low frequency in the monolingual corpus and whether the frequencies of compounds in the monolingual corpus correspond with the frequencies in the parallel corpus. We chose the Araneum Germanicum Maius corpus (see Section 4.3.1) as a suitable resource of the monolingual data because of its size and complexity.

Before we considered the frequencies of the compounds in the OPUS corpus and in the Araneum corpus, we compared the counts of the attested and not-attested compounds in both corpora (see Table 6.1). Most of the compounds listed in GermaNet were present in the Araneum corpus (94,714 out of 99,075 – see Table 6.1).

	# in OPUS	**# not in OPUS**	**Total**
# in Araneum	50,597	44,117	94,714
# not in Araneum	547	3,814	4,361
Total	51,144	47,931	99,075

Table 6.1: Counts of attested and not-attested compounds from GermaNet in the OPUS corpus vs. in the Araneum corpus

After that, we created two graphs showing frequencies of the compounds in both corpora. Both graphs (Figure 6.2 and Figure 6.1) have logarithmic scaling.

If we observe the curves in Figure 6.1, the compounds seem to be split into halves where one half is present in OPUS and the second one not. The curves are very similar (as of frequencies of compounds attested and not-attested

in the OPUS corpus) although the compounds included in the OPUS corpus are more frequent than the compounds not-attested in the OPUS corpus. In addition, most of the words not included in the Araneum corpus were also not present in the OPUS corpus (3,814 out of 4,316 compounds – see Table 6.1).

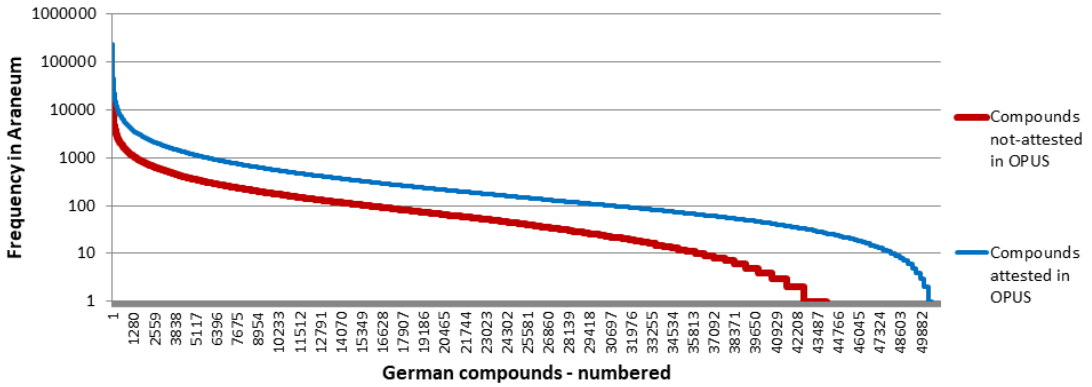

Figure 6.1: Absolute corpus frequency in the Araneum of the OPUS-attested compounds vs. of the compounds not-attested in the OPUS

Figure 6.2 shows that the average frequency in the OPUS corpus more or less corresponds to the frequency in the Araneum corpus. However, there are a lot of words which do not correspond at all – for example, compounds with high frequency in the Areneum corpus can have a very low frequency in the OPUS corpus (ex. 71) or even do not occur in the OPUS corpus at all (ex. 72). We also found words with a lot of hits in the OPUS corpus, but with a very low frequency in the Araneum corpus (see 73 and 74). The reason for these differences between the corpora might be the specificity of the OPUS corpus. It contains domain specific data (administrative texts, newspaper articles or movie subtitles, see Section 4.2.1).

(71) Schwerpunkt 'main emphasis' (frequency: 10 in OPUS, 73,773 in Araneum)

(72) Krankenkasse 'health insurance company' (frequency: 0 in OPUS, 35,191 in Araneum)

(73) Luftfahrzeug 'aircraft' (frequency: 10,121 in OPUS, 854 in Araneum)

(74) Unterabsatz 'subparagraph' (frequency: 36,087 in OPUS, 336 in Araneum)

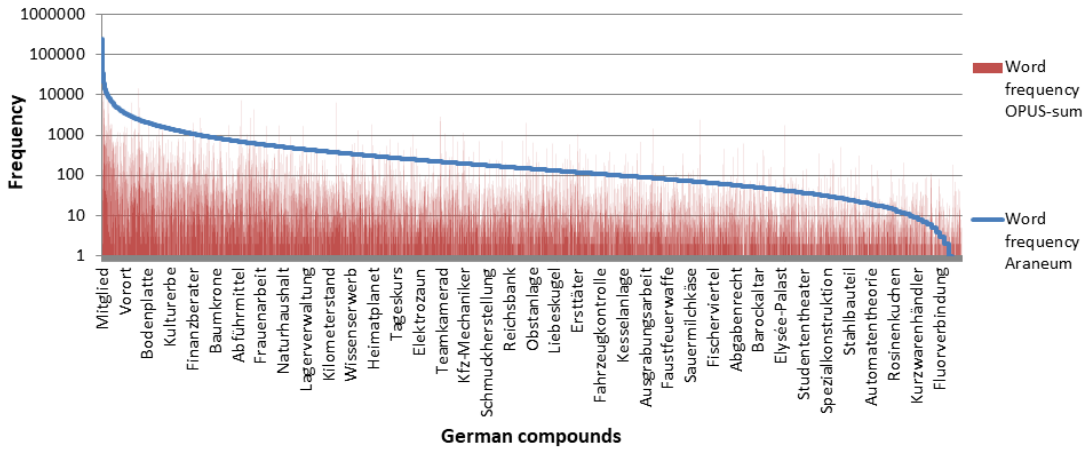

Figure 6.2: Absolute corpus frequency of the OPUS-attested compounds in the Araneum corpus vs. in the OPUS subcorpora (sum of frequencies from all subcorpora)

6.2 Classification of Czech equivalents

We divided the Czech equivalents of the German compounds from GermaNet (51,144 phrases) into one-word, two-word and more-word equivalents. In order to gain a deeper insight, the equivalents in each group were further distinguished into types based on the POS categories. In order to do that, we loaded the data from the lemmatized OPUS dictionaries, which we prepared earlier (see Section 5.2.1), and stored them into a dictionary (the most frequent POS category for each lemma was chosen, see examples in Figure 6.3). Then the dictionary was loaded and each equivalent or, in the case of two- and more-word equivalents, each item of it was assigned the POS category.

```
protokol                NOUN-41732, ADV-11
o                       ADP-1666563, NOUN-5561, ADJ-1359, ADV-7, NUM-5, VERB-2
podmínka                NOUN-168037, VERB-2
se                      PRON-4643324, ADP-59313, NOUN-52981, VERB-187, AUX-119, ADJ-98, NUM-5
smlouva                 NOUN-194074, VERB-75, ADV-18, ADJ-6
člen                    NOUN-72547, ADJ-277
být                     AUX-10071405, VERB-1191173, NOUN-38220, ADJ-668, ADP-168, PART-60
mezi                    ADP-210849, NOUN-102
stát                    NOUN-550239, VERB-187923
```

Figure 6.3: POS categories of words from the OPUS corpus – example

We chose several classes according to the count of words and the combination of the POS categories for classification of the identified equivalents. The classes were inspired by the frequent types in the data and by the categorization of our referential dataset (see Section 5.1).

Most one-word equivalents were either nouns (“N”) or adjectives (“A”), so they were put into the corresponding classes. The nominal counterparts were analysed for the inner morphological structure and during the analysis further divided into compounds, into words where the second constituent of the German

compound corresponds with a suffix in the Czech equivalent and into independent Czech words which are unrelated to the German compounds (see Section 6.2.5).

Two-word equivalents were divided into A+N, N+N phrases and nouns followed by an adjective (N+A). Counterparts with more than two-words that had words in expected POS categories (nouns, adjectives, numerals, adverbs and prepositions) were classified as "Multi-word".

Some of the equivalents obtained from the automatic identification did not fit into any of the categories, therefore we added one more type "Other". We noticed that there were some equivalents which had prepositions on the beginning or at the end of the phrase (not only in the "Other" but also in the "Multi-word" category). We considered that as an error that happened in the course of the identification of equivalents (after viewing enough examples) and removed the redundant prepositions.

The category of adjectives was suspicious, because an adjective is not expected to be an equivalent of a nominal compound. Based on a manual analysis of a sample of 100 adjective equivalents, it turned out that most of them are part of an A+N phrase that has not been recognized correctly by the automatic tool. In order to get a more adequate results, we decided to revise this group. We looked for words that follow these adjectives in the sentences, where the original German compound was found in the parallel data (see Section 4.2.1). We stored all the words that followed the adjectives with number of their occurrences (only in case they followed the adjective in the sentence containing the particular German compound) for each adjectival equivalent into a dictionary and, after that, the most frequent candidate was chosen. If it was a noun we added it to the adjective in order to build a complete A+N phrase.

The resulting counts of the identified A+N phrases and other adjectival equivalents are displayed in Table 6.2. Over 60% of the adjectival equivalents were added to the subset of the A+N equivalents and the rest to "Other". So the size of the subsets of the A+N counterparts was increased by 2.3% and the size of the Other subset by 1.5%.

Type	**Adjectives identified**	**% of adjectives**	**% of all equivalents**
A+N	1,179	60.4%	2.3%
Other	772	39.6%	1.5%

Table 6.2: Number of types of Czech equivalents identified in the A-category

The distribution of the categories of the Czech counterparts (after these two mentioned types of redistribution) is displayed in Figure 6.4. As we can see in Figure 6.4, the distribution is very similar to that in our manually created dataset of 150 compounds (see Figure 5.1). Almost all of the types (except for "N") have smaller percentage rates than in our manually processed dataset (Section 5.1). This can be explained by the existence of an extra category of unclassified equivalents. The phenomena that we found in the "A" category (partly identified phrases) occurs undoubtedly also in the "N" category, where phrases of A+N or N+N are hidden, which might be a clue why the percentage of this category is the same as in the referential dataset. Each type is analyzed in detail below.

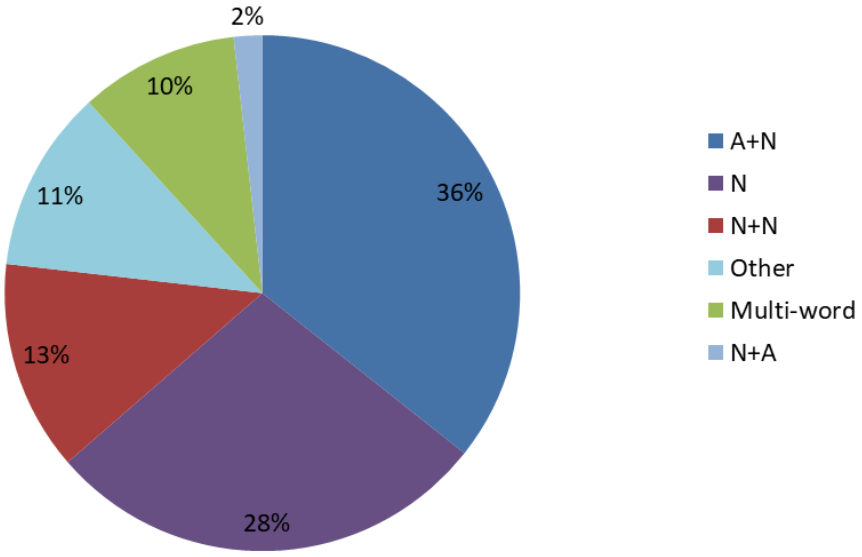

Figure 6.4: Distribution of types of Czech equivalents based on the data from the automatic identification

6.2.1 A+N

Most nominal compounds from GermaNet that were found in the OPUS corpus have a two-word equivalent in Czech made up of an adjective and a noun (36% as displayed in Figure 6.4). We wanted to find out, whether the two-word counterparts of German nominal compounds are frequent collocations or only phrases created by a speaker in order to describe the reality in the current moment. So, we investigated, how often these bigrams occur together, i.e. Their collocability in Czech. We used data from SYN2020 (Section 4.3.2) for that purpose and looked for the A+N phrases and stored their frequency.

After that we collected the data and displayed them in a histogram according to their frequency (see Figure 6.5). The distribution was counted according to the input German compounds, so if there were compounds with identical equivalents in Czech, the frequency count of the equivalents was included more than once in the histogram (see examples 75, 76 or 77). The fact that a particular Czech phrase is listed as a counterpart of different German compounds, can have different reasons:

- there are two (or more) different German compounds which are synonymous, for example, (75) – first constituents of both compounds (*niedrig* and *billig* 'low') are synonymous and have the same Czech equivalent (*nízký*)
- the identification of Czech equivalents did not find the exact equivalent for one of them,for instance, (76) – the Czech counterpart of the first compound (76a) express only part of the meaning of the German compound
- the form of the Czech phrase has more meanings, for example, (77) – the Czech equivalent *hlavní role* has the same meaning as both of the German compounds (*Hauptdarsteller* 'main actor' and *Hauptrolle* 'main role')

The frequency intervals presented in the histogram as well as in the analysis were created manually, we considered bigrams with the highest and the lowest frequency and also the distribution of their frequency. Almost 20% of the equivalents were not in the set of bigrams extracted from the SYN2020 corpus and over 70% of them had frequency 50 or lower (including zero frequency). Only 17% of the A+N bigrams had more than 100 hits in the SYN2020 corpus. The phrase *světová válka* 'word war' (frequency 5,299) had the highest frequency out of all phrases correctly identified.

The histogram shows that most of the A+N equivalents of German compounds have low collocability. It may indicate that most of the German compounds with A+N counterparts in Czech are used to refer to changing facts of the extralinguistic reality rather than to have a fixed, idiomatic meaning.

Additionally, we decided to check whether there is no problem in low-frequency bigrams that should be revised. Our analysis revealed, that phrases with zero frequency are either incorrect – wrongly lemmatized (see 79) or incorrectly identified by our algorithm, or they are not frequent collocations in Czech (see 78). We found no crucial problem in the data that could be fixed by a simple revision.

As for the syntactic relationship, all Czech A+N phrases manifested the same structure: the adjective determines the noun. So we looked into the data in order to find out whether these phrases are Czech equivalents of German compounds with the corresponding syntactic relation. If there were compounds with a different syntactic relation between their constituents, the Czech equivalents identified by our algorithm would not be correct. All the viewed examples of the German compounds were determinative (for instance, 75, 77 and 78), so there was no obvious problem in this part.

(75) a. Niedrigpreis – nízký cena 'low price'
b. Billigpreis – nízký cena 'low price'

(76) a. Hauptstadtfunktion 'function of the capital' – hlavní město 'capital'
b. Hauptstadt – hlavní město 'capital'

(77) a. Hauptdarsteller 'main actor' – hlavní role 'main role' or also 'main actor'
b. Hauptrolle 'main role' – hlavní role 'main role' or also 'main actor'

(78) Werbebudget – reklamní rozpočet 'advertising budget'

(79) Weltfrieden 'world peace' – *světový míra 'world rate'

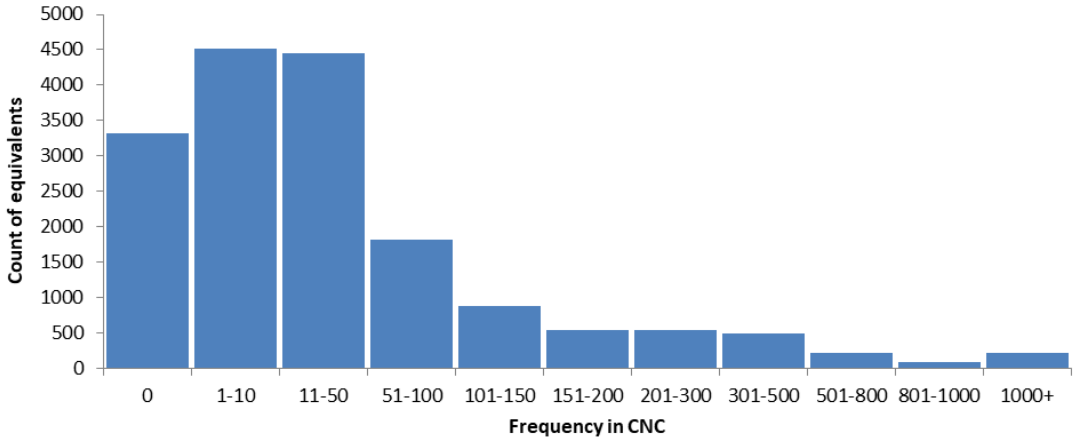

Figure 6.5: Frequency of the A+N equivalents of German compounds – histogram according to SYN2020

6.2.2 N+N

N+N is the third biggest category after the A+N bigrams and the N category (13% of the Czech equivalents – see Figure 6.4). Similarly as for the A+N category, we studied the collocability of the nouns in the bigrams and created a histogram (see Figure 6.6). The histogram is, analogously to that for A+N category, based on the counts of the Czech equivalents and their frequency in the SYN2020 corpus. There were also Czech phrases which occurred as equivalents of different German compounds. The same Czech equivalent was listed either with two synonymous German words (see 81) or with two different German words with different meanings that can be translated equally into Czech (see 80).

The frequency intervals presented in the histogram were again established manually according to the data. As we can see in Figure 6.6, a lot of bigrams have zero frequency (35%). However, these are not only incorrect phrases, but also Czech bigrams not included in the SYN2020 corpus (see 83) or phrases that have different tags in the SYN2020 corpus – one part of the phrase is not considered as a noun (see 82). However, frequency of most of the other bigrams is not much higher – 61% of them have the frequency of 10 or lower. Only 16% of the N+N bigrams have more than 50 hits in the SYN2020 corpus. The most frequent bigram is *konec roku* 'end of the year' with 2,848 hits in the corpus. Phrases with the structure of N+N (average frequency 39) are in Czech less frequent than A+N (83.5 hits on average) according to the SYN2020 corpus.

(80) a. Körperteil – část těla 'body part'
b. Leichenteil 'part of the dead body' – část těla 'body part'

(81) a. Lebensweise – způsob života 'way of life'
b. Lebensart – způsob života 'way of life'

(82) Ordinatenachse – osa y 'y-axis'

(83) Ozonkonzentration – koncentrace ozónu 'ozone concentration'

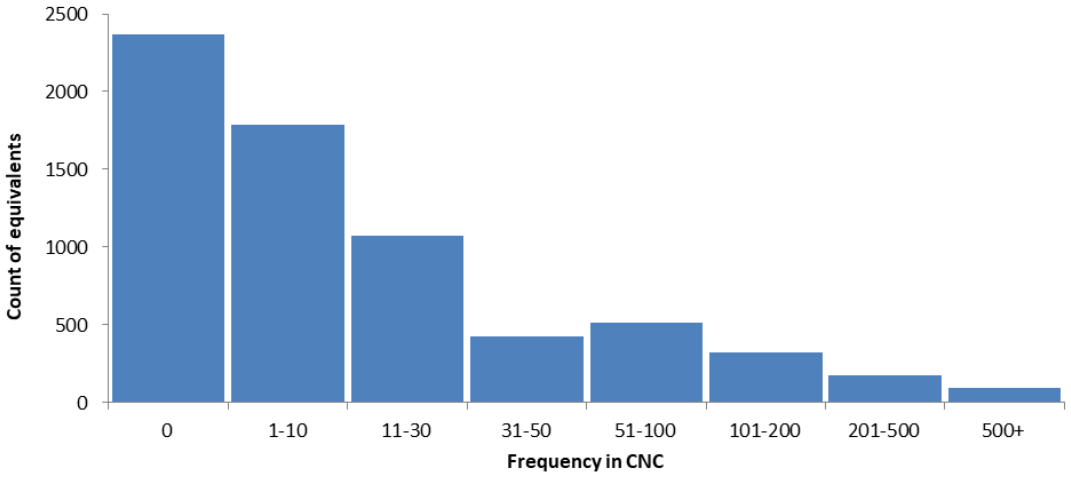

Figure 6.6: Frequency of the N+N equivalents of German compounds – histogram according to SYN2020

Not only the collocability is interesting for the N+N phrases, but we also considered the relation between these two nouns. We viewed the case of the second noun, because we expected that most of these two-word phrases have the determinative relation between their constituents and that the second part depends on the first.

We went through all lemmas from the OPUS sources with tags (see Section 5.2.1) and searched for two nouns that follow one another. Their tags with the number of their occurrences were listed and the most frequent combination of tags for each bigram was chosen and stored in a dictionary. Finally, we looked for the N+N equivalents in that dictionary and extracted the morphological case of the second noun. We created a graph showing the distribution of the cases (Figure 6.7).

Figure 6.7 documents that most of the second nouns (almost 80%) are in genitive. There is a determinative relationship between the constituents of these phrases (see for example, 80 or 81). These phrases are counterparts of determinative German compounds, however, the heads of the German compounds (the second constituents of the compounds) correspond to the first words of the N+N phrases. In the Czech N+N phrases with determinative relationship between their parts, the second part describes the first one.

8% of the second nouns are in the nominative case. There are different reasons for that. First, we extracted the N+N phrases from the data where no sentence segmentation was provided, so there can also be phrases, where one noun comes from the end of one sentence and the second one from the beginning of the following sentence (see 84). There can even be an error in defining the tag (see 85). However, we can also find phrases where both nouns are in the nominative (see 86 and 87). The constituents of these phrases are appositive. Both parts denote the same thing but in a different way.

(84) Abwasserleitung ‘sewer’– odtok vývod ‘drain outlet’

(85) Apfelschale – slupka jablka ‘apple peel’

(86) Alphamännchen – alfa samec ‘alpha male’

(87) Apple-Computer – počítač Apple ‘Apple computer’

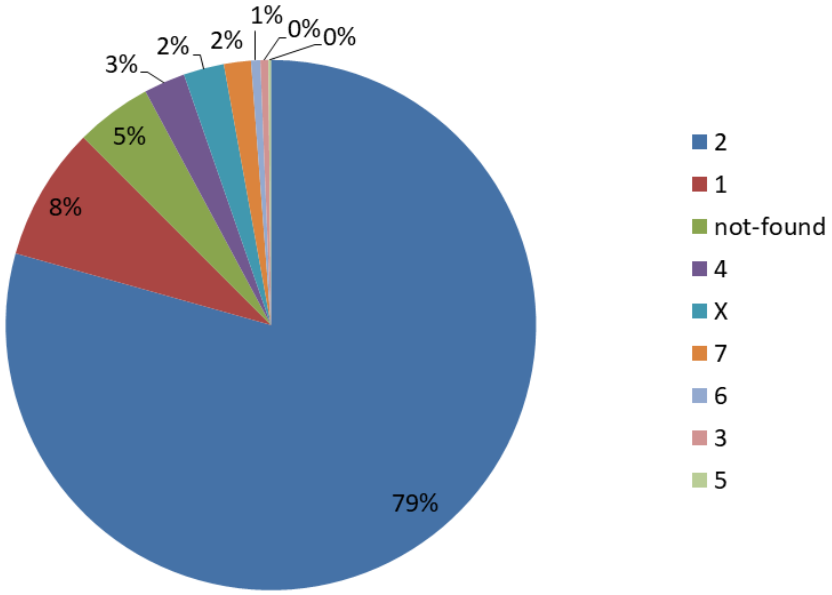

Figure 6.7: Distribution of morphological cases of the second nouns in the N+N phrases

6.2.3 N+A

Phrases with the structure of a noun followed by an adjective (N+A) occurred relatively rarely (2% of the Czech equivalents). However, we also considered the category interesting. We can find words that refer to several different phenomena – names of institutions (see 88), chemical compounds (see 89), biological names of species (example 90) or other established phrases (see 91) between them. We also discovered partially translated compounds in this category, where a multiword equivalent was needed for expressing the entire meaning of the compound. A general phrase ending with an adjective was identified, where the adjective does not determinate the meaning of the first noun, but it needs a further explanation – a noun that would express the missing part of the meaning of the phrase (see 92).

The N+A phrases in Czech, as the phrases in the A+N category, have determinative relationship between their parts. However, the constituents are in a reversed order than in the A+N phrases. The order of the parts of the N+A bigrams does not correspond to that in the German compounds. The N+A phrases are counterparts of the determinative German compounds.

(88) Karls-Universität – univerzita Karlova ‘Charles university’

(89) a. Kohlendioxid – oxid uhličitý ‘carbon dioxid’

b. Kohlenstoffdioxid – oxid uhličitý ‘carbon dioxid’

(90) Waldkiefer – borovice lesní ‘scots pine’

(91) Kubikmeter – metr krychlový ‘cubic meter’

(92) Qualitätsfrage ‘quality issue’ – otázka týkající ‘question concerning’

We created a histogram, as with the N+N and A+N categories above, in order to see the frequency of these N+A phrases (see Figure 6.8). As in the A+N or N+N subsets, there were cases of German compounds corresponding to the same Czech equivalent (for example, 89). Over 50% of the phrases were not found in the SYN2020 corpus – that are mostly not completely translated compounds such as in 92 or not so frequently used names of species. Only 16% of the phrases from this category have frequency higher than 10. The most frequent phrase is *Univerzita Karlova* ‘Charles university’ with a frequency of 1,098 (see 88). The low count of occurrences of these phrases may be a consequence of the specificity of these terms. There are a lot of chemical or biological terms, which are not so frequently used in the language.

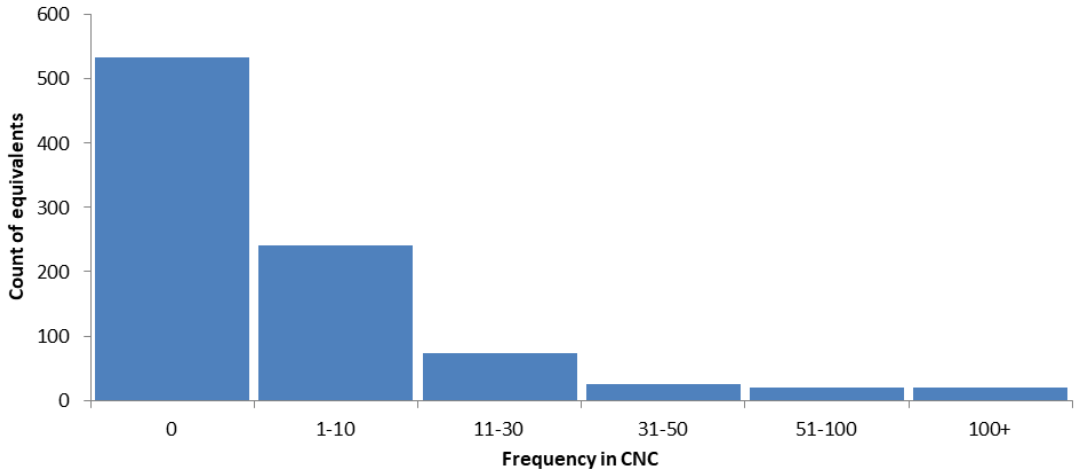

Figure 6.8: Frequency of the N+A equivalents of German compounds – histogram according to SYN2020

6.2.4 Multi-word equivalents

10% of the Czech equivalents are multi-word phrases (see Figure 6.4). Phrases in this category are made up of words from different POS categories, so we examined the exact distribution of these structures. Because we found a lot of different types, we considered only the most common structures (with more than 100 occurrences) in our data, all additional types were classified as “Other”. The distribution of the types of multi-word equivalents is displayed in Figure 6.9.

The most frequent type of the multi-word equivalents are two nouns connected by a preposition (see 94) – they account for 28% of all multi-word equivalents. Other frequent types of the equivalents were for example, N+N phrases where one of them was expanded by an adjective (see 93 and 95). A noun with two defining adjectives occurred in 6% (see 96). Also phrases where all the parts have the same POS category (see 97) or other more complicated prepositional phrases

(for example, N+prep+A+N see 98) were present among the equivalents. We classified almost 30% of all multi-word equivalents as "Other" – there are not so frequent structures (see example 99) or incorrectly selected equivalents.

In this category, there are equivalents of both determinative (for example, 94, 97 or 98) and copulative (ex. 95) German compounds. As we can see in the examples listed in 93–99, constituents of the German compound directly correspond to the parts of the Czech equivalent. For example, the head of the German compound in 93 – *Rückführung* corresponds with the word *recirkulace* 'recirculation' in the Czech equivalent. The head is described by the modifier *Abgas*, that corresponds with the part of the Czech equivalent *výfukový plyn* 'exhaust gas'. As we can see, most of the listed examples (93–99) of German compounds have their constituents in reversed order than parts of their Czech equivalents (except for 96).

(93) Abgasrückführung – recirkulace výfukového plynu 'exhaust gas recirculation'

(94) Produktinformation – informace o přípravku 'product information'

(95) UNESCO-Weltkulturerbe – světové dědictví UNESCO 'UNESCO World Heritage'

(96) Sommerfahrplan – letní jízdní řád 'summer timetable'

(97) Essstörung – porucha příjmu potravy 'eating disorder'

(98) Familienspiel – hra pro celou rodinu 'family game'

(99) Schadensabwicklung – provádění znaleckého posudku o škodě 'carrying out an expert opinion on the damage' (N+A+N+prep+N)

The most frequent category of elementary, prepositional phrases (N+prep+N) was analyzed in order to examine prepositions required specifically by the particular nouns. We counted the occurrences of the prepositions that followed the particular nouns. We created a list of all nouns that occurred multiple-times with particular prepositions (it shows that they require one). The list was sorted by the count of occurrences in our data (see Appendix A.1). We also collected nouns that were used with multiple different prepositions and listed the count of their occurrences (see Appendix A.2). 161 nouns that require a preposition were found (out of 826 different nouns in the prepositional phrases).

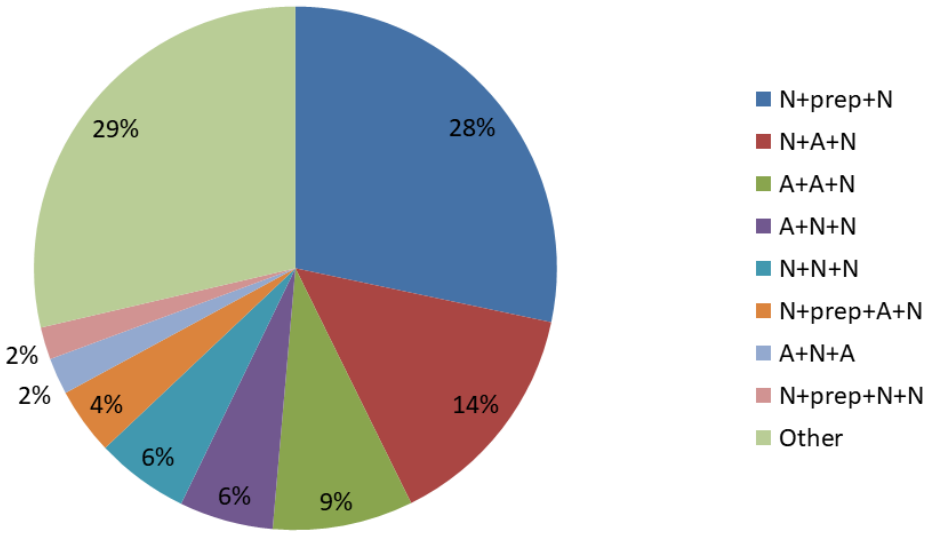

Figure 6.9: POS structures of multi-word equivalents

6.2.5 Nouns

One-word nominal phrases accounted for 28% of the equivalents (see Figure 6.4). In order to analyze this category in more detail, it was split into:

- compounds
- words where the second constituent of the German compound corresponds to a suffix in the Czech equivalent (abbreviated as words with a particular suffix)
- independent words in Czech

Because there are possibly also partly identified equivalents (only one part of the phrase was identified as a counterpart of the German compound) – for example, N+N or A+N phrases – included in this category, we also created a subcategory of possibly half-correct equivalents.

We used two methods to divide nominal equivalents into these subcategories. Firstly, we considered only German compounds and their parts with the phrase table gained from the parallel data. We took Czech translations of the constituents of the German compounds into account and compared them with the nouns from this category to see whether they correspond. The second method was implemented as a pilot study with a goal to split and identify Czech compounds (Svoboda and Ševčíková [fortcomming]). We describe both methods and their results below.

The constituents of German compounds were translated with the help of the phrase table. The best ten translations (according to the direct translation probability) for both parts of the German compounds were collected from the phrase table. For each Czech equivalent, the word was compared stringwise with the translations of the constituents of the corresponding German compound.

For determining a noun as a compound, both parts had to match (usual linking elements and possible alternations were taken in account), for example, the German compound *Acrylamid* (ex. 100) had the constituents translated to Czech as *akryl* and *amid* (these equivalents were present in the best ten gained from the phrase table) and these translations matched with the Czech equivalent *akrylamid*.

To estimate a word with a particular suffix, we created a retrograde dictionary from all nominal equivalents and extracted the most commonly used suffixes (suffixes with 30 and more occurrences were taken). Based on the extracted suffixes, we looked for a translation of the first compound constituent matching with the part of the noun without the suffix (see ex. 101 – the suffix *-ník* was found and the Czech word *mravenec* 'ant' was an equivalent of the first part of the German compound *Ameise*.

(100) Acrylamid 'acrylamide' (Acryl + Amid) – akrylamid (akryl + amid)

(101) Ameisebär 'ant bear' (Ameise + Bär) – mravenečník (mravenec + ník)

The remaining nouns were compared with the equivalents of both constituents of the German compounds. If they matched fully with the translation of one part, they were classified as "half-correct", otherwise as an "independent word". The division into these two categories is not fully correct, because there can be nouns included that correspond only to one part of the German compound, but can also be accepted as correct equivalents for the whole compound (ex. 102). This may be caused by the preference of the speaker (as inferred from the parallel data). We found examples in the data, where the Czech equivalents were really only partly identified (ex. 103) or where the counterpart was identified correctly as a one-word equivalent (ex. 104), but there were also examples that were inaccurately classified as independent words (ex. 105).

(102) Abflussmenge 'flow rate' – průtok 'flow' (independent word)

(103) A-Mannschaft 'A-team' – a (half-correct)

(104) Abendveranstaltung 'evening event' – večírek 'party'(half-correct)

(105) Abendvorstellung 'evening performance'– promítání 'projection' (independent word)

Type	**Nouns identified**	**% of nouns**	**% of all equivalents**
Independent word	6,668	47%	13%
Half-correct	6,185	43%	12%
Particular suffix	1,008	7%	2%
Compound	445	3%	1%

Table 6.3: Number of types of equivalents from N-category – obtained from our method which compares counterparts of constituents of the German compounds with the Czech equivalents

After that, we compared the counts of nouns classified into above described categories (see Table 6.3). We did not find many compounds, so we combined our

method with a tool for classification and splitting of Czech compounds(Svoboda and Ševčíková [fortcomming]).

This tool identified 778 compounds and 194 of them were also identified by our algorithm (out of 445 compounds classified by our algorithm), so we revised the counts of nouns of each category according to the results of this tool. Before we did that, we compared compounds identified by both methods. Several equivalents of German compounds were correctly classified as "half-correct" by comparing the strings, but they were also identified as compounds according to the morphological structure of the Czech equivalents (see 106). However, a lot of Czech compounds were classified only by one of the methods correctly – some of them were identified by our string comparison method (see 107) and other compounds were incorrectly classified as "word" (see 108) or "half-correct" (see 109) in comparison with the second method (Svoboda and Ševčíková [fortcomming]). After this comparison, we decided to take compounds identified at least by one of the methods and created a final distribution (see Table 6.4).

According to the distribution of the nominal equivalents (see Table 6.4), German compounds have more derivative counterparts in Czech as compared to Czech compositional equivalents – in the Czech equivalents, there are only 1,029 compounds in comparison with over 6,000 independent words and 949 words with particular suffix.

German compounds with composite equivalents in Czech are either copulative (ex. 107) or determinative (ex. 108, 109). The syntactic relationship between the constituents corresponds with the Czech counterparts. Czech words where the second part (the head) of the German compound corresponds with a suffix in Czech (ex. 110, 111) are translations of determinative compounds. The modifier and not the head of these German compound is preserved as a stem in the Czech counterpart.

The Czech words where the second part of the German compound is expressed as a suffix were further analyzed. We listed all the suffixes together with their numbers of occurrences which we identified in the subcategory of the nouns with a particular suffix. We also wanted to see the German equivalents of these suffixes, so we looked into the data and stored all of them together. After that, we went through the data manually. The suffixes were examined (with the help of the dictionary of Czech affixes Šimandl [2016]) and corrected. Several suffixes were merged (for example, *-inec* and *-ánec*). The resulting list of suffixes with the number of their occurrences is displayed in A.3. Possible equivalents of these suffixes are presented in A.4 (only some of the most frequent equivalents are displayed). Examples of words with the most frequent suffixes are presented in 112–116. The suffix *-ní* expresses some action (112), the suffix *-ště* is used for places (113), the suffix *-ctví* is on the end of the names of shops (114), *-ník* implies that the word denotes an object (115) and the suffix *-ka* signifies a women (116).

(106) Auto-aufkleber 'car sticker' – samo-lepka 'sticker' (half-correct, sám lepka)

(107) Bass-bariton – bas-baryton 'bass baritone' (compound, basbaryton)

(108) Bei-fahrer – spolu-jezdec 'co-driver' (word, spolu jezdec)

(109) Arten-vielfalt – bio-diverzita 'biodiversity' (half-correct, -bio- diverzita)

(110) Ameisen-hügel – mraven-ště ‘anthill’ (suffix)

(111) Mittels-frau – prostředn-ice ‘mediator (fem.)’ (suffix)

(112) Abschieds-gruß – rozlouče-ní ‘farewell’

(113) Arbeits-ort – pracov-ště ‘workplace’

(114) Antiquitäten-laden – starožitni-ctví ‘antique shop’

(115) Aschen-becher – popel-ník ‘ash tray’

(116) Bar-frau – barman-ka ‘bartender (fem.)’

Type	Nouns identified	% of nouns	% of all equivalents
Independent word	6,249	44%	12%
Half-correct	6,079	42%	12%
Compound	1,029	7%	2%
Particular suffix	9,49	7%	2%

Table 6.4: Distribution of types of equivalents from the N-category – provided after the compounds identification by the tool for identifying and splitting of Czech compounds

6.2.6 Other

The remaining set ”Other” encompasses equivalents that fitted none of the groups above (overall 11% of all equivalents – see Figure 6.4). The phrases were unclassified for two different reasons – 79% of them included a POS category (for example, verb) that was not expected as an equivalent of a nominal compound, and other 21% were from expected POS categories (nouns, adjectives, numerals, adverbs and prepositions) but they did not match with our chosen types of equivalents (see Section 6.2).

We carried out an error analysis of the examples of both types. Verbs were most problematic in the first category – the equivalents were either identified incorrectly (problems in word alignment, equivalent selection – for example, redundant words (mainly verbs) were added before or after the phrase (see 117) or the phrase was identified totally incorrectly) or there was a problem in tagging and/or lemmatization (for example, *večeří* ‘dinner (genitive, plural)’ also means ‘he dines’ was lemmatized as **večeřit* and tagged as verb – see 118).

Most of the 21% of equivalents from the “Other” category were incorrectly identified equivalents (the problem mostly rooted in word alignment – totally different words were selected as equivalents of German compounds), however, we also discovered some incorrectly lemmatized and tagged equivalents (see 119) or not fully translated phrases, where one or two words were missing (see 120).

(117) Abendspaziergang ‘evening walk’ – večerní procházka být ‘evening walk to be’

(118) Abendmenü ‘evening menu’ – nabídka *večeřit ‘dinner menu’

(119) Aaskrähe ‘carrion crow’– *vraný obecný (A+A)

(120) 32-Bit-Prozessor ‘32-bit processor’ – 32 bitový ‘32-bit’

6.3 Final analysis

We created a final graph showing the distribution of the individual types of the Czech equivalents (see Figure 6.10), where also the subclassification of the N category is provided (described above in Section 6.2.5).

If we consider the final distribution in Figure 6.10 generally, most of the German compounds listed in GermaNet were translated to Czech as phrases made up of more than one word. 51% of the Czech counterparts were two-word phrases (with structures A+N, N+N and N+A) and 10% of all equivalents were phrases with three or more words. In addition, 12% of the nouns were suspected to be A+N or N+N phrases, but it was not investigated in the data. Only 2% of all Czech counterparts were classified as compounds in Czech. However, 2% of the German compounds were translated as words where the second constituent of the German compound corresponds with a suffix in the Czech equivalent and 12% of Czech counterparts were independent words (consisting of parts directly unrelated to the constituents of German compounds). This fact may be seen as a piece of evidence in favor of the general linguistic assumption that the composition is less important in Czech, while the derivation prevails.

Since most of the Czech equivalents of German compounds were two-word phrases (A+N, N+N and N+A), we studied their collocability. In all of the three classes, the Czech counterparts occurred mostly with lower frequency in the corpus. It might indicate that most of the German compounds do not correspond to frequently used collocations, but are coined to describe things occurring in the actual extralinguistic reality.

We also considered syntactic relationships between the parts of the Czech equivalents in comparison to that between the constituents of the German compounds. It was possible only for the equivalents made up of more than one word and for Czech compounds and words, where the second part of the compound corresponded to a suffix in Czech (abbreviated as words with particular suffixes). We showed that the Czech equivalents mostly preserve the same syntactic relationship between the parts as the German compounds have, but the order of the parts can differ (only the phrases with parts in determinative relationship). Some types of the Czech equivalents have the same order of the parts as the German compounds, such as A+N, words with particular suffixes or compounds. Typically, the N+N phrases have the reverse order in comparison with right-headed German compounds.

Over 10% of all Czech equivalents were not recognised or did not fit into our types. While analyzing all the types one by one, an error analysis was done. We discovered problems in tagging and lemmatization and also in word alignment or equivalent selection algorithm. The examples were shown and described mainly in Section 6.2.6.

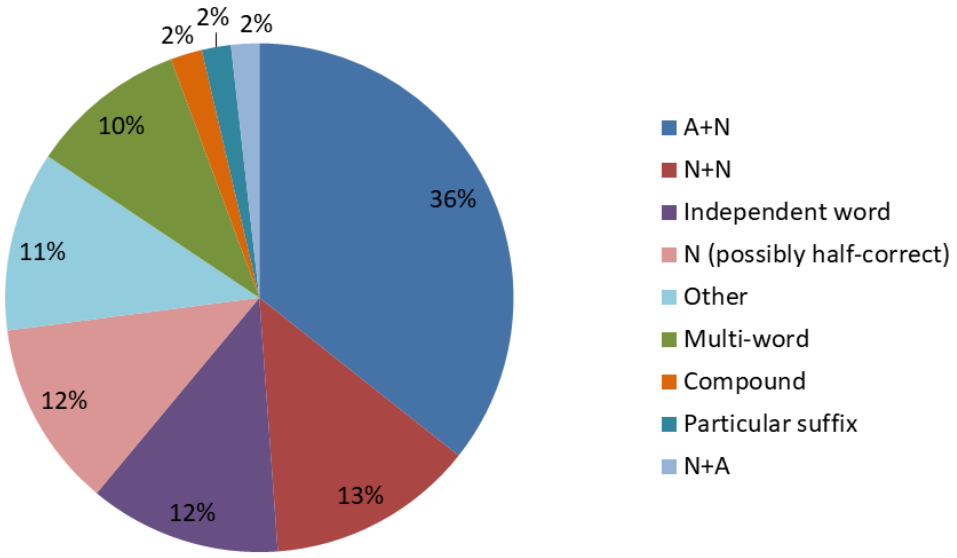

Figure 6.10: Final distribution of types of Czech counterparts of German nominal compounds listed in GermaNet

7. Conclusion

The goal of the thesis was to automatically identify Czech equivalents of German compounds, and analyze them according to the number of the items involved, to their POS structure, eventually to their morphological and syntactical structure.

Firstly, we identified suitable resources of German nominal compounds and of parallel data. We provided a brief analysis of the selected data and created a manually annotated dataset containing 150 nominal compounds from the resources. After the manual analysis, we got an insight into the topic and started with a preparation for the automatic identification. The parallel data were preprocessed with several NLP tools – lemmatization, tagging, word alignment, phrase extraction and scoring were provided. Then, the scoring policy for the selection of the correct equivalents of German compounds was tuned and counterparts of German compounds from the selected resources were identified.

After the identification of Czech counterparts of the German compounds, an analysis of the results was carried out. The distribution of the German compounds in parallel corpora and in a monolingual corpus was compared. We distinguished several types of Czech equivalents of German compounds and analyzed each type individually. While we considered the data of each type separately, we made an error analysis and described the sources of the most frequent errors. We were able to identify that most of the problems originated in lemmatization and word alignment.

The analysis seems to support the general assumption that the process of compounding is exploited differently in German and Czech. We showed that only 2% of German nominal compounds were translated as compounds to Czech. Most of the German compounds were translated to Czech by two-word phrases with determinative relationship between their parts. We found both determinative and copulative German compounds in the data, but the determinative ones prevailed. The syntactical structure of their Czech counterparts always corresponded to their structure. However, the order of the parts of Czech equivalents differed across the types of Czech equivalents. German compounds are right-headed, but the order of the corresponding parts in the determinative Czech phrases can be reversed, for example, in the determinative N+N phrases or in the N+A phrases. Despite the fact, the A+N phrases or even some N+N phrases (counterparts of copulative German compounds) have the same order of their parts as their equivalent German compounds. The syntactic relations between the constituents and their order were studied for German compounds and their Czech equivalents of all the types.

Czech equivalents made up of three or more words consisted of both types of the syntactic relation and their parts were mostly reversed compared to the constituents of the equivalent German compound. The distribution of the frequency of the two-word phrases documented that only few percent of them are frequent collocations in Czech. It signifies that the German compounds do not only correspond to commonly used collocations, but are also created to describe things occurring in the actual extralinguistic reality.

Rare, but also interesting are Czech equivalents of German compounds which

are also compounds in Czech or which are words where the head of the German compound corresponds to a suffix in the Czech counterpart. Czech compounds have the same structure as their German equivalents. The head of the German compounds was expressed as a suffix in 2% translations. The final picture shows that the composition is rarely used in Czech. In addition, Czech compounds equivalent to German compounds are less frequent as compared to derivatives (translations with the head expressed as a suffix with other derived nouns from the "independent words" type).

Despite the fact that we worked with specific data with the help of specific tools, we believe that our thesis could provide valuable insights into the structure of German compounds and their Czech counterparts, which might be relevant both for natural processing tasks and linguistic research.

Bibliography

R Harald Baayen, Richard Piepenbrock, and H Van Rijn. The CELEX lexical database (CD-ROM). Linguistic Data Consortium. *Philadelphia, PA: University of Pennsylvania*, 1993.

Irmhild Barz. German. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, volume 4, pages 2387–2410. Mouton de Gruyter, Berlin, 2016.

Vladimír Benko. Aranea: Yet another family of (comparable) web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, pages 247–256, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10816-2.

Ondřej Bojar and Daniel Zeman. Czech machine translation in the project CzechMate. *The Prague Bulletin of Mathematical Linguistics*, 101:71–96, May 2014. doi: 10.2478/pralin-2014-0005.

Ivana Bozděchová. Czech. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, volume 4, pages 2872–2891. Mouton de Gruyter, Berlin, 2016.

Edward Joseph Callow. Predicting German compound words using a recurrent neural network [online]. Master's thesis, University of Washington, 2019. URL `https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/44691/Callow_washington_02500_20779.pdf`.

Simon Clematide, Stéphanie Lehner, Johannes Graën, and Martin Volk. A multilingual gold standard for translation spotting of German compounds and their corresponding multiword units in English, French, Italian and Spanish. In *Multiword Units in Machine Translation and Translation Technology*, pages 126–145. John Benjamins, 2018. URL `https://www.jbe-platform.com/content/books/9789027264206-cilt.341.06cle`.

Michaela Cocca, Václav Řeřicha, and Elizabeth Alvarado. Comparison of formation processes in English and Czech sports terminologies. *Linguistica Pragensia*, 25:132–144, June 2015.

František Čermák and Alexandr Rosen. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13:411–427, December 2012. doi: 10.1075/ijcl.17.3.05cer.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1073`.

Michaela Ficenecová. Způsoby tvoření slov v angličtině a češtině [online]. Bachelor's thesis, University of West Bohemia, Faculty of Education, Plzeň, 2011. URL https://dspace5.zcu.cz/bitstream/11025/5711/1/BP_MF.pdf.

Rita Finkbeiner and Barbara Schlücker. *Compounds and multi-word expressions in the languages of Europe*, pages 1–44. De Gruyter, January 2019. ISBN 9783110632446. doi: 10.1515/9783110632446-001.

Anna Hätty and Sabine Schulte im Walde. Fine-grained termhood prediction for German compound terms using neural networks. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 62–73, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4909.

Kateřina Hegerová. Czech and German equivalents in the translations of the English novel Lolita by V. Nabokov (with the main focus on compounds) [online]. Master's thesis, Palacký University Olomouc, Faculty of Arts, 2009. URL https://theses.cz/id/pude5h/.

Verena Henrich and Erhard Hinrichs. Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 420–426, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/R11-1058.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, 2007.

Radka Koprdová. Anglizismen im Deutschen und Tschechischen: Ein Vergleich anhand ausgewählter journalistischer Texte über EU-Ereignisse [online]. Master's thesis, Masaryk University, Faculty of Arts, Brno, 2013. URL https://is.muni.cz/th/btk7n/.

Irina Krotova, Sergey Aksenov, and Ekaterina Artemova. A joint approach to compound splitting and idiomatic compound detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4410–4417, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.543.

M. Křen, V. Cvrček, J. Henyš, M. Hnátková, T. Jelínek, J. Kocek, D. Kovaříková, J. Křivan, J. Milička, V. Petkevič, P. Procházka, H. Skoumalová, J. Šindlerová, and M. Škrabal. SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, 2020. URL http://www.korpus.cz.

Ivana Kvapilíková. Unsupervised machine translation between Czech and German language [online]. Bachelor's thesis, Czech Technical University in Prague, Faculty of Information Technology, 2020. URL https://dspace.cvut.cz/bitstream/handle/10467/86196/F8-BP-2020-Kvapilikova-Ivana-thesis.pdf.

Věra Levová. Slovotvorný proces v českém a anglickém jazyce [online]. Bachelor's thesis, University of West Bohemia, Faculty of Education, Plzeň, 2012. URL https://dspace5.zcu.cz/bitstream/11025/5712/1/Bakalarska%20prace%20-%20Levova.pdf.

Susan Olsen. Composition. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation. An International Handbook of the Languages of Europe*, volume 1, pages 364–386. Mouton de Gruyter, Berlin, 2015.

Milan Smutný et al. Czech equivalents of English compounds. *Discourse and Interaction*, 1(2):99–108, 2008.

Milan Straka, Jana Straková, and Jan Hajič. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4212. URL https://www.aclweb.org/anthology/W19-4212.

Sara Stymne. A comparison of merging strategies for translation of German compounds. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 61–69, Athens, Greece, April 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E09-3008.

Emil Svoboda and Magda Ševčíková (fortcomming). Splitting and Identifying Czech Compounds. Submitted to the 3rd Workshop on Resources and Tools for Derivational Morphology, 2021.

Martin Šemelík. *Wortbildung im deutsch-tschechischen Wörterbuch [online].* PhD thesis, Charles University, Faculty of Arts, Prague, 2014. URL https://dspace.cuni.cz/handle/20.500.11956/56921.

Josef Šimandl, editor. *Slovník afixů užívaných v češtině.* Karolinum, Praha, 2016. ISBN 978-80-246-3544-6. URL http://www.slovnikafixu.cz.

Monika Šimková. Word-formation processes in English and Czech common and specific features (focussed on affixation, compounding and conversion) [online]. Bachelor's thesis, University of West Bohemia, Faculty of Education, Plzeň, 2011. URL https://theses.cz/id/s6c39q/.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Jitka Trachtová. Ausgewählte Aspekte der deutschen und tschechischen Wortbildung. Wie werden die Zusammensetzungen übersetzt? [online]. Master's thesis, Masaryk University, Faculty of Arts, Brno, 2012. URL `https://is.muni.cz/th/x3qw6/`.

Marion Weller and Ulrich Heid. Analyzing and aligning German compound nouns. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2395–2400, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/817_Paper.pdf`.

List of Figures

List of Tables

List of Abbreviations

POS	Part-of-speech
A	Adjective
N	Noun
V	Verb
Num	Numeral
Pron	Pronoun
Adv	Adverb
prep	Preposition
fem	Feminine
NLP	Natural language processing
MT	Machine translation
RNN	Recurrent neural network
CNC	Czech National Corpus

A. Attachments

A.1 Nouns attested in the N+prep+N phrases multiple-times with particular prepositions

obchod	s	30	oznámení	o	3	přistání	na	2
problém	s	23	osvědčení	o	3	překládka	v	2
žádost	o	22	okno	v	3	pytlík	s	2
pokus	o	14	odstoupení	od	3	pytel	na	2
dohoda	o	12	ochrana	proti	3	pruh	pro	2
zákon	o	11	návod	k	3	prostor	pro	2
továrna	na	11	nástroj	na	3	produkt	z	2
smlouva	o	11	nemoc	z	3	pozor	na	2
právo	na	11	ministr	pro	3	povlak	na	2
smysl	pro	10	lístek	na	3	potíž	s	2
daň	z	10	kukuřice	na	3	poptávka	po	2
obchodník	s	9	krabice	na;s	3;3	podnět	k	2
zpráva	o	8	jednání	o	3	pacient	s	2
péče	o	8	film	o	3	osoba	v	2
povolení	k	8	dárek	k	3	odkaz	na	2
náklad	na	8	domeček	pro	3	odchod	z	2
trh	s	7	žaloba	na	2	obvinění	z	2
prohlášení	o	7	žadatel	o	2	nástup	do	2
nárok	na	7	šťáva	z	2	nádoba	na	2
vstup	do	6	člověk	od	2	nakládání	s	2
test	na	6	čas	na	2	místnost	s	2
boj	proti;o	6;3	účast	na	2	motiv	k	2
místo	na;v	5;2	útěk	z	2	mlýnek	na	2
informace	o	5	úspora	z	2	miska	na;s	2;2
účet	za	4	úrok	z	2	medaile	za	2
stánek	s	4	zápas	o	2	med	z	2
přístup	k	4	zápach	z	2	láhev	od	2
práce	v;pro;na	4;2;2	změna	v	2	let	podle	2
požadavek	na	4	zařízení	na	2	lak	na	2
očkování	proti	4	věda	o	2	krvácení	do	2
obchodování	s	4	výpis	z	2	krmivo	pro	2
nůž	na	4	vzdálenost	mezi	2	krabička	od	2
nehoda	na	4	vražda	z	2	koupelna	pro	2
klíč	od	4	vosk	na	2	kontejner	na	2
hospodaření	s	4	volání	o	2	kolíček	na	2
hon	na	4	voda	k;z;s	2;2;2	kamarád	z	2
dveře	do	4	vakcína	proti	2	jízda	na	2
život	na	3	učitel	na	2	investice	do	2
závod	v	3	tlak	v	2	hnutí	za	2
výhled	na	3	svět	v	2	důvod	k	2
výbor	pro	3	styk	s	2	dům	na	2
válka	za;v	3;3	studie	na	2	díra	po;v	2;2
vchod	do	3	stres	z	2	diskuse	o	2
touha	po	3	strach	z	2	den	na	2
stopa	po	3	stojan	na	2	dar	od	2
srážka	s	3	souvislost	s	2	clo	z	2
spor	o	3	směrnice	o	2	chování	v	2
rada	pro	3	skvrna	od	2	chlupa	na	2
příspěvek	na	3	situace	s	2	cesta	po;na	2;2
příkaz	k	3	setkání	s	2	cena	na	2
potvrzení	o	3	rozkaz	k	2	bomba	v	2
poplatek	za	3	rezervace	v	2	bolest	v	2
pokoj	pro	3	reklama	na	2	alergie	na	2
plechovka	od	3	příjem	z	2			

A.2 Nouns attested in the N+prep+N phrases with several different prepositions

skandál	s na	1 1
test	na v z	6 1 1
přístup	k na do	4 1 1
problém	s v	23 1
škola	v bez	1 1
rok	v za	1 1
žaloba	na proti pro	2 1 1
pokus	o na s	14 1 1
obchod	s pro	30 1
boj	proti o s za	6 3 1 1
podpora	v z	1 1
stres	z v	2 1
díra	po v na	2 2 1
žadatel	o po	2 1
náklad	na k za	8 1 1
učitel	na v	2 1
důvod	k pro	2 1
opatření	proti na	1 1
telefon	v s	1 1
dveře	do od u na	4 1 1 1
člověk	od bez	2 1
reklama	na o v	2 1 1
voda	k z s v po	2 2 2 1 1
trasa	v na	1 1
lístek	na z do	3 1 1
slupka	od z	1 1
zůstatek	u na	1 1
holka	od z	1 1
dům	na pro	2 1
dřevo	kvůli pro	1 1
válka	za v o s	3 3 1 1
místo	na v pro k	5 2 1 1
daň	z za	10 1
účast	na podle v	2 1 1
důstojník	z v na	1 1 1
klid	na v	1 1
nástroj	na k pro	3 1 1
schopnost	k v	1 1
vstup	do na	6 1
zisk	za po	1 1
odchod	z do	2 1
krabice	na s od	3 3 1
nehoda	na při v o	4 1 1 1
krabička	od na	2 1
chlupa	na z v	2 1 1
bolest	v na z	2 1 1
tanec	na s	1 1
pomocník	do pro	1 1
deska	v s	1 1
cesta	po na kolem	2 2 2
pořad	v pro o	1 1 1
zařízení	na proti pro před	2 1 1 1
úraz	v během	1 1
olej	z do	1 1
pivo	v s	1 1
peníze	z na za	1 1 1
papír	do pro	1 1
vražda	z pomocí	2 1
požadavek	na o k z	4 1 1 1
zkušenost	v s	1 1
výcvik	v s	1 1
řízení	o k pod	1 1 1
společnost	s proti	1 1
závod	v na o	3 1 1
ochrana	proti před	3 1
zápas	o s v	2 1 1
práce	v pro na k z	4 2 2 1 1
otázka	k na	1 1
příběh	z v	1 1
zpráva	o v pro k z	8 1 1 1 1
pomoc	při proti	1 1
počasí	k na	1 1
licence	na k	1 1
domeček	pro z	3 1
miska	na s	2 2
dárek	k pro	3 1
čas	na za	2 1
poznámka	pod z	1 1
láhev	od s na	2 1 1
postel	pro s z	1 1 1
situace	s na v	2 1 1
věda	o v	2 1
přihrádka	pro nad	1 1
sestava	mezi na	1 1
skupina	podle pro	1 1
obal	na s	1 1
kamarád	z v	2 1
rána	z do při	1 1 1
razítko	z kvůli	1 1
fotka	z pro	1 1
běh	přes v	1 1
den	na v	2 1
informace	o s	5 1
přípravek	proti na	1 1
let	podle do na	2 1 1
krok	v pro	1 1
nádoba	na s	2 1
dítě	v od	1 1
hudba	pro do	1 1
látka	na proti	1 1
život	na v	3 1
srážka	s z	3 1
prodavač	v na	1 1
pacient	s po	2 1
návštěva	u na	1 1
skok	do na o	1 1 1
chování	v za	2 1
pytel	na s	2 1
lampa	na u	1 1
výrobek	pro z	1 1
příručka	o pro	1 1
mandát	v k	1 1
úkol	z na	1 1
příjem	z pro	2 1
prášek	k na	1 1
karta	na z	1 1
blok	s z	1 1
záznam	o do	1 1
vzorek	v na	1 1
centrum	na pro	1 1
doba	na pro	1 1
oblečení	do k	1 1
doprava	pro pod	1 1
skokan	na o do	1 1 1
příplatek	za na	1 1

A.3 Occurrences of suffixes in the Czech words where the second part of the German compound is expressed with a suffix

-ní	208
-ost	124
-tví/-ství/-ctví	118
-ník	88
-ka/-anka/-enka	68
-ina	51
-ice	48
-ace	39
-vka	37
-ště	34
-ika	23
-tel	21
-ita	21
-rna	21
-ovna	20
-ista	16
-inec/-ánec	14
-áček/-íček	12
-alka/-álka	9
-dlo	8
-or	6
-ek	3

A.4 German equivalents of suffixes in the Czech words where the second part of the German compound is expressed with a suffix

-ní -arbeit:9, -vorschrift:4, -möglichkeit:4, -tätigkeit:3, -pflicht:3, -anlage:3,
↪ -verfahren:2, -zeit:2, -tour:2, -versuch:2, -findung:2, -mittel:2, -spiel:2, -sport:2,
↪ -bildung:2, -förderung:2, -verhalten:1, -session:1, -phase:1, -bereich:1, -leistung:1,
↪ -zahlung:1, -kunst:1, -wärme:1, -störung:1, -erscheinung:1, -geschehen:1, -stadium:1,
↪ -system:1, -maßnahme:1, -anspruch:1, -regelung:1, -problem:1, -manöver:1, -gruß:1
-ost -rate:7, -grad:4, -quote:4, -vermögen:4, -kraft:3, -sein:3, -zahl:3, -sucht:3,
↪ -prüfung:3, -gefühl:3, -zustand:2, -zahlung:2, -nachweis:2, -lage:2, -fähigkeit:2,
↪ -verbesserung:2, -prinzip:2, -pflicht:2, -dauer:2, -frage:2, -kram:2, -gehalt:2,
↪ -entwicklung:2, -bereitschaft:1, -merk:1, -ausgleich:1, -betrag:1, -geld:1, -name:1,
↪ -strom:1, -wirkung:1, -neigung:1, -gabe:1, -leben:1, -termin:1, -verlust:1
-tví/-ství/-ctví -laden:9, -wesen:9, -sektor:3, -wissenschaft:3, -haltung:3, -sinn:3,
↪ -wirtschaft:3, -bau:3, -kunst:3, -branche:2, -zucht:2, -händler:2, -stand:2, -führung:2,
↪ -leben:2, -frage:2, -arbeit:2, -gesellschaft:2, -gewerbe:2, -minister:2, -buch:1, -reise:1,
↪ -tour:1, -bereich:1, -reden:1, -angebot:1, -tag:1, -tätigkeit:1, -besteigung:1, -sport:1,
↪ -sitz:1, -geschäft:1, -handlung:1 -salon:1, -betrieb:1
-ník -baum:6, -eck:4, -mann:4, -nehmer:3, -buch:3, -leute:3, -inhaber:2, -macher:2,
↪ -besitzer:2, -schirm:2, -fahrer:2, -haus:2, -bär:1, -kurs:1, -händler:1, -becher:1,
↪ -leader:1, -arbeiter:1, -gänger:1, -mitglied:1, -person:1, -fachmann:1, -makler:1, -revier:1,
↪ -sklave:1, -helfer:1, -kandidat:1, -treibende:1, -führer:1, -herr:1, -träger:1, -bund:1,
↪ -gesellschafter:1, -zeitschrift:1
-ka/-anka/-enka -frau:5, -schrank:2, - dose:2, -schein:2, -fabrik:2, -tochter:2,
↪ -waschanlage:1, -anlage:1, -eisen:1, -werk:1, -band:1, -figur:1, -mädchen:1, -dame:1,
↪ -sender:1, -kleid:1, -rock:1, -kabel:1, -kauffrau:1, -bildung:1, -mittel:1, -firma:1,
↪ -muschel:1, -rolle:1, -tier:1, -tüte:1, -film:1, -bildner:1, -rad:1, -ruhe:1, -manufaktur:1
-ina -stoff:5, -land:4, -unterricht:2, -material:2, -obst:1, -sprache:1, -waren:1,
↪ -handel:1, -diele:1, -löffel:1, -molekül:1, -besitz:1, -bild:1, -frieden:1, -ansammlung:1,
↪ -volumen:1, -anbau:1, -kultur:1, -brief:1, -raum:1, -fell:1, -leder:1, -stimmung:1,
↪ -erkrankung:1, -leiden:1
-ice -haus:5, -frau:2, -stube:2, -schuh:2, -anlage:2, -mensch:1, -bevölkerung:1, -leben:1,
↪ -bahn:1, -station:1, -stuhl:1, -wanderung:1, -schlagader:1, -tau:1, -garten:1, -wasser:1,
↪ -form:1, -stück:1, -speicher:1, -linie:1, -munition:1, -rakete:1, -signal:1
-ace -gebiet:3, -pflicht:2, -phase:2, -installation:2, -grad:1, -raum:1, -auftrag:1,
↪ -gespräch:1, -freiheit:1, -material:1, -marke:1, -betrieb:1, -anzeige:1, -bild:1, -bedarf:1,
↪ -erklärung:1, -dienst:1, -wesen:1, -rate:1, -fähigkeit:1, -störung:1, -aufgabe:1
-vka -dose:2, -wasser:2, -rakete:1, -schein:1, -note:1, -mütze:1, -zettel:1, -schirm:1,
↪ -büchse:1, -rechnung:1, -schaden:1, -tier:1, -kirsche:1, -tv:1, -fabrik:1, -gewehr:1,
↪ -pistole:1
-iště -platz:5, -stätte:5, -feld:2, -haus:2, -haufen:1, -hügel:1, -ort:1, -hafen:1,
↪ -stadt:1, -land:1, -grund:1, -revier:1, -schauplatz:1, -gelände:1, -landschaft:1, -deck:1,
↪ -garage:1, -bahn:1, -box:1, -kasten:1, -grube:1, -moor:1, -anlage:1, -aufgang:1
-ika -artikel:3, -stellung:2, -markt:1, -frage:1, -sektor:1, -übung:1, -student:1,
↪ -wissenschaft:1, -rennsport:1, -ökonomie:1, -unterricht:1, -fabrik:1, -feld:1, -technikum:1,
↪ -industrie:1, -profil:1, -technik:1, -kritik:1, -aufschwung:1, -form:1
-tel -verband:2, -inhaber:2, -täter:2, -zahler:1, -vorsteher:1, -beamter:1, -geber:1,
↪ -vater:1, -amt:1, -bildung:1, -begründer:1, -funktion:1, -verschwörer:1, -wisser:1,
↪ -lehrer:1, -gott:1, -steller:1, -writer:1
-ita -person:1, -grad:1, -steigerung:1, -mobilität:1, -sache:1, -wert:1, -sexualität:1,
↪ -problem:1, -verlust:1, -prinzip:1, -rate:1, -krise:1, -lage:1, -schwierigkeit:1,
↪ -fortschritt:1
-rna -werk:8, -fabrik:5, -café:1, -haus:1, -raum:1, -anlage:1, -leistung:1, -asyl:1,
↪ -kraftwerk:1, -saal:1,
meldung:1, -silbe:1, -bau:1, -grund:1, -belastung:1, -rechtfertigung:1, -ausflug:1
-ovna -haus:3, -raum:3, -laden:2, -einrichtung:1, -bestand:1, -regal:1, -schrank:1,
↪ -mutter:1, -auswahl:1, -grube:1, -prägestette:1, -stube:1, -agentur:1, -gesellschaft:1,
↪ -unternehmen:1
-ista -spieler:4, -fahrer:3, -sport:1, -autor:1, -schreiber:1, -springer:1, -profi:1,
↪ -gewinner:1, -beamter:1, -offizier:1, -person:1
-inec/-ánec -haus:2, -kind:2, -vertretung:1, -anteil:1, -hof:1, -paar:1, -tag:1, -woche:1,
↪ -fladen:1, -tier:1, -kranz:1, -heim:1
-áček/-íček -finger:3, -kind:1, -knopf:1, -gerät:1, -kabel:1, -äffchen:1, -kissen:1,
↪ -ankömmling:1, -bombe:1, -mantel:1
-alka/-álka -fabrik:1, -hersteller:1, -konzern:1, -marke:1, -frau:1, -probe:1,
↪ -überholung:1, -bonbon:1, -bein:1
-dlo -mittel:2, -zeug:1, -stoff:1, -stempel:1, -fläche:1, -lader:1, -werk:1
-or -projektor:1, -merkmal:1, -modulator:1, -talent:1, -urheber:1, -gesellschaft:1
-ek -erscheinung:1, -bezahlung:1, -maß:1