Modeling Compounding for Multilingual Data Resources

Thesis Proposal

Emil Svoboda
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
svoboda@ufal.mff.cuni.cz

Abstract

This proposal presents the progress that has been made toward finishing our dissertation. We have set out to model compounding multilingually; to that end, we describe a series of experiments in determining the parent words of compounds. The first experiment presents a neural model of compounds called Czech Compound Splitter, which returns the motivating words of a given compound and discriminates compounds from non-compounds. Its successor, Word Formation Analyzer for Czech, returns the motivating word(s) of any input word, and is able to discriminate compounds from both derivatives and unmotivated words. The final experiment in the series is PaReNT, which performs the same tasks, but supports 8 languages including Czech. Additionally, we propose a pipeline, partially based on already-existing tools, that endows morphologically segmented compounds with a dependency structure, modeling their internal composition.

1 Introduction

Compounds are words immediately motivated by at least two words. Coverage of this word-formation process in computational data resources has varied widely in quality and quantity. This is at least in part because there is a lack of computational infrastructure applicable to the purpose of building and harmonizing multilingual data resources. This is the niche we intend to fill.

We use Czech as a starting point because we happen to have the most insight into the language, coupled with access to high-quality data. However, we aim for a multilingual setting, so we branch off into other languages from there. We have so far been able to cover Czech, English, German, Dutch, Russian, French, and Spanish, which represent three genera (Slavic, Romance, Germanic) of the Indo-European language family.

We model compounds from a static perspective. This means that we look at already-existing compounds and determine their compoundhood, find their ancestor words, and analyze their morphological structure. This is in contrast with a dynamic perspective, in which the procedure of compounding would be modeled. To reflect this decision, the examples in this proposal are formatted with the compound on the left, followed by a left arrow, with the compound ancestors (parent words or parents) on the right, alongside their glosses and part-of-speech (POS) (cf. ex. 1, 2):

\[
\begin{align*}
(1) & \quad \text{Compound} \leftarrow \text{Parent}_1 + \text{Parent}_1 \quad \text{LANG} \\
& \quad \text{gloss.POS} \quad \text{gloss.POS} \quad \text{gloss.POS}
\end{align*}
\]

\[
\begin{align*}
(2) & \quad \text{clairsemé} \leftarrow \text{clair} + \text{semé} \quad \text{FR} \\
& \quad \text{thiny scattered.A} \quad \text{clear.A} \quad \text{spread.A}
\end{align*}
\]

The proposal begins in Section 2 by introducing basic concepts and terminology pertaining to compounding and taxonomy frameworks relevant to this proposal. We continue in Section 3 by providing an overview of compounding-relevant data resources and computational tools. We show our progress in Section 4. First, we present Czech Compound Splitter (CCS), a tool capable of splitting Czech compounds into words it was motivated by (its parents), as well as discriminating them from non-compounds. We continue by presenting Word Formation Analyzer for Czech (WFA.ces), which performs the more general task of parent retrieval, which returns the immediate word-formation ancestors of not just compounds, but also derivatives. Additionally, it performs word-formation classification, which places any given word into one of three categories – Compound, Derivative, Unmotivated. Next, we describe PaReNT, a tool that performs parent retrieval and word-formation analysis on all of the eight languages in scope. Finally, in Section 5, we propose a pipeline that could endow morphologically segmented compounds with a dependency tree structure.
2 Theoretical considerations regarding compounding

2.1 Classification of compounds

Proposals on the taxonomy of compounds vary significantly, and there exist too many to list exhaustively. Indeed, at the beginning of Chapter 3 of The Oxford Handbook of Compounding (Lieber and Štěkauer, 2011), Bisetto and Scalise remark that “from the beginning, almost every scholar dealing with composition has proposed his/her own view”.

Bozděchová (1997) proposes a hierarchical classification of compounds within the onomasiological theory of word formation. The classification is applied to a dataset of 3000 Czech compounds. The highest level is classified by the POS of the compound. The middle level is classified by the type of referent that the compound names – e.g. person, property bearer, place name for nouns. The lowest level is the formal division of compounding into three categories – simple compounding proper, complex compounding proper, and compounding improper (juxtaposition). Compounding proper refers to the spontaneous coining of a two-rooted word by a speaker, spurred on by the need to name a particular object in a particular speech situation, which makes it a genuine word-formation phenomenon. This is contrasted with compounding improper, which is the phenomenon of syntactic expression gradually solidifying over time, which places it in the domain of syntax. In Czech, the phrase that is encoded by an improper compound can be reconstructed solely by finding an appropriate split-point and splitting the compound there with no morphological adjustments. For example, splitting the improper compound vždy zelený “evergreen” yields the valid, correctly formed phrase vždy zelený “always green”, whereas the proper compound bělobřichý “white-bellied” does not yield a correctly formed phrase when split without adjustment. Complex compounding proper corresponds to what we call parasynthetic compounding, and is described in detail in 2.2.3.

Bisetto and Scalise (2005) take a different approach and use two criteria to classify compounds – the relation between the constituents (subordinate, coordinate, attributive) and centricity (exocentric, endocentric), which in their view are independent of each other. Coordinate compounds are characterized by a symmetric relationship between the constituents; the relation in subordinate and attributive compounds is in contrast asymmetric. Subordinate compounds are those whose relation is that of <complement>, whereas in attributive compounds the relation is that of <attributive>. Compounds of any relation may be exocentric or endocentric. We explain this terminology in detail and discuss it in the context of other authors’ classification systems:

Constituent relation. In most taxonomies, the way that the constituents of a given compound are related is given importance. Most scholars who use the relation concept list one or more relation that is symmetric (often calling such a relation some variant of coordinate or coordinative) – Fabb (1998), for example, considers coordinative compounds to be synonymous with two-headed compounds, as the heads modify each other – and one or more asymmetric relations, such as the subordinative and attributive relations of Bisetto and Scalise (2005).

Headedness. Most scholars (e.g. Fabb 1998; Haspelmath 2002; Bisetto and Scalise 2005; Bozděchová 1997; Štichauer 2013) operate with some notion of a head; that is, some compounds have a constituent that in some way governs the properties of the given compound. For example in 3, the resulting noun is masculine, inheriting its gender from its right constituent, and refers to a particular city.

3 Волгоград ← Волга + град (RU)


Those compounds that do have a head are often analyzed as to whether the head is followed by the non-head element, or vice versa. Compounds in which the head precedes the modifier are termed left-headed; compounds in which the head follows are termed right-headed; this suggests that ex. 3 is right-headed – otherwise, the word would inherit the feminine gender of Volga, and would probably denote some part of the river. Fabb (1998) additionally considers two-headed compounds. Some authors distinguish between syntactic heads and semantic heads, the former of which governs the compound’s formal properties, like gender; and the latter of which governs the compound’s meaning.

Centricity. Centricity evaluates whether a given compound inherits form or meaning from one of its constituents by means of subsetting (endocentric; an apple cake is a type of cake) or not (exocentric; a cutthroat is not a type of throat). This is
usually understood as being related to headedness; in fact, Fabb (1998) considers exocentricity to be synonymous with headlessness.

Štichauer (2013) proposes a three-level taxonomy based on Bisetto and Scalise (2005) and applies it to Czech. The additional level of analysis describes the structure of a given compound by the parts-of-speech of the constituents enclosed in square brackets, with a symbol in between them denoting the kind of relation that the constituents have with each other. Usually, the POS of the resulting compound is shown in the subscript to the right. Štěrkopísek “mixture of sand and gravel” can thus be described as \([N + N]_N\), as an example of an endocentric coordinate nominal compound.

2.2 Challenges

2.2.1 Edge cases

When dealing with compounds, an immediate problem arises – where lies the boundary between compounding and other word-formation processes, and between compounding and syntax? This is a heated topic in morphology because the question begs the answer to other unsolved questions, such as the precise definitions of wordhood and morpheme boundness. As a result, numerous edge cases exist – and for computational purposes, these need to be resolved one way or another. For instance, on the boundary between derivation (or the so-called \(micro\) question; Lieber and Štekauer 2011) and compounding lies combinations of verbs and prepositions.

\[
(4) \text{unterstehen} \leftarrow \text{unter} + \text{stehen} \quad \text{(DE)}
\]

In ex. 4, unterstehen can either be considered a compound, or it can be understood as a derivative of stehlen with the prefix unter-. The case for the compounding interpretation can be made by observing that unter syntactically behaves like a free word in German. However, the productivity pattern of compounds with unter is much more reminiscent of derivation. Furthermore, Lieber and Štekauer (2011) propose that roots should have more \(semantic\) substance than affixes, but it is difficult to argue that unter has more semantic substance than for example the undisputed affix pre-.

On the other end of the spectrum, there is the fuzzy boundary between compounding and syntax, or what Lieber and Štekauer (2011) term the \(macro\) question. It may not be clear at which point a given syntactic phrase has “solidified” enough to be considered a word on its own. The Czech tradition would, for instance, consider the adjective vždyze-lený mentioned in Section 2 to be a single word, but this largely relies on orthographic convention, which may not be reliable in English (cf. \(flowerpot\), flower-pot, flower pot are all valid spellings).

2.2.2 Morphological variation

Some compounds are formed by the mere juxtaposition of existing words. However, this is often not the case. In ex. 5, we observe the addition of an -e- interfix between the constituents. In some languages, variation goes beyond interfix addition.

In \(\text{водопровод}\) (cf. ex. 6), the interfix replaces the ending of the first constituent \(^*\text{вод}\). Internal flexion also appears, like in the English \(\text{womenfolk}\) (ex. 7), where the first constituent is inflected for plurality.

\[
(5) \quad \text{bridegomon} \leftarrow \text{bruid} + \text{gom} \quad \text{(NL)}
\]

\[
\text{bridegonym} \leftarrow \text{bruid} + \text{gom} \quad \text{(NL)}
\]

\[
(6) \quad \text{водопровод} \leftarrow \text{вода} + \text{провод} \quad \text{(RU)}
\]

\[
(7) \quad \text{womenfolk} \leftarrow \text{woman} + \text{folk} \quad \text{(EN)}
\]

2.2.3 Parasynthetic compounding

One of the ways compounding interacts with other aspects of language is when it occurs together with some other word-formation process. Cross-linguistically, typical examples of this process involve body parts.\(^2\) This process has been cross-linguistically attested (cf. Czech ex. 8, Dutch ex. 9, English ex. 10, Latin ex. 11; Melloni and Bisetto 2010):

\[
(8) \quad \text{modrooký} \leftarrow \text{modrý} + \text{oko} \quad \text{but no} \quad \text{*oký}
\]

\[
\text{blue-eyedA} \quad \text{blue.A eye.N}
\]

\[
(9) \quad \text{blauwogig} \leftarrow \text{blauw} + \text{oog} \quad \text{but no} \quad \text{*ogig}
\]

\[
\text{blue-eyed.A} \quad \text{blue.A eye.N}
\]

\[
(10) \quad \text{blue-eyed} \leftarrow \text{blue} + \text{eye} \quad \text{but no} \quad \text{*eyed}
\]

\[
\text{A} \quad \text{A N}
\]

\[
(11) \quad \text{albicapillus} \leftarrow \text{albus} + \text{capilla} \quad \text{but no} \quad \text{*capillus}
\]

\[
\text{white-hairedA} \quad \text{white.A hair.N}
\]

Whether or not a given compound is parasynthetic may be a matter of analysis. This leads to difficulty in annotation. Similarly to the examples just discussed, the second parent of Czech přimotop (ex. 12) can only be topit, not *top, which is a bare stem and not a word, so the motivating process behind

\(^2\)In body parts, the reason words like *haired as in red-haired or *legged as in bow-legged are unattested is probably because the base assumption is all humans have these body parts, and therefore such words would carry minimal information value.
this word must be compounding together with conversion. However, the similar krvotok (ex. 12) can be either analogously understood as compounding together with conversion, assigning the verb téci “to flow” as the second parent, or we can assign the noun tok “flow” as the parent and understand the motivating process as simple compounding proper (cf. Bozdˇechová (1997)).

(12) přímotop ← přímo + topit (CZ) 
heater.N directly.ADV heat.V

(13) krvotok ← krev + téci/tok (CZ) 
bloodflow.N blood.N to flow/flow.V/N

2.2.4 Neoclassical compounding

Neoclassical compounding is a special case of compounding wherein elements borrowed from Ancient Greek and Latin are combined either with each other or with free words. We term such elements neoclassical constituents.3 For example, in the English monolog, neither the first constituent *mono- nor the second constituent *-log can be attested on their own. However, they cannot be considered to be simply affixes, because a combination of affixes without free morphemes cannot by definition form a free word.

In Bauer’s (1998) view, it is inappropriate to view neoclassical compounding as a category discrete from compounding and derivation. Instead, he proposes that lexical enrichment can be evaluated over three dimensions – [simplex-compound; native-foreign; abbreviated-nonabbreviated] – and neoclassical compounding is a fuzzy subspace of these three dimensions.

In 2021, Ološtiak and Vojteková introduced a classification of neoclassical compounds for West Slavic languages. They delimit three types of compounds according to the type of formants involved. Proper compounds are characterized as being composed of two freely-appearing bases, as in ex. 14. Semi-compounds are composed of one base and one neoclassical constituent (cf. ex. 15). Finally, quasi-compounds are composed of two neoclassical constituents (cf. ex. 16).

(14) séropozitivní ← sérum + pozitivní (CZ) 
seropositive.A serum.N positive.A

(15) kryptopolitika ← krypto- + politika (CZ) 
cryptopolitics.N crypto.NECONE politics.N

(16) ekologie ← eko- + logie (CZ) 
ecology.N eco.NECONE logy.NECONE

3Also known as baseoid under (Ološtiak and Vojteková, 2021)

3 Compounds in data sources and tools

3.1 Word-formation data sources covering compounding

In this section, we briefly list data sources that are relevant to the modeling of compounding across languages. One of the strongest motivations for the research described in this thesis is the fact that compounds are often underrepresented in word-formation resources, and even when they are not, their handling is inconsistent across languages or even across individual datasets.

DeriNet (Vidra, Jonáš and Žabokrtský, Zdeněk and Ševčíková, Magda and Straka, Milan, 2015) is a data resource and associated API that stores word-formation data in the form of lexemes linked to their single derivational ancestor. Word-formation families are therefore represented as a tree. Since version 2.0 (Vidra et al., 2019), support for compounding has been added in the form of a binary yes/no compound flag for each lexeme, as well as by allowing a single lexeme to have multiple parents. As a result, word-formation families containing compounds are no longer trees, but rather directed acyclic graphs (DAGs). The latest version, 2.1 (Vidra et al., 2021), contains over 2,000 compounds with assigned parents.

CELEX (Baayen et al., 2014) is a general lexical database covering English (50,964 items), Dutch (118,029 items) and German (51,278 items), which apart from word formation also covers inflection and syntactical properties of the included lexical items. The database covers compound structure as well – it includes the morphological segmentation of each word using nested parentheses, with an associated part-of-speech tag for each segment.

Golden Compound Analyses (Vodolazsky and Petrov, 2021) is a collection of around 2000 Russian compounds hand-annotated for the purposes of training a Russian compound splitter.

Universal Derivations (Kýjáněk et al., 2021) is a collection of 31 data resources harmonized so that they can be handled using the DeriNet API. These cover 21 languages, including the ones in scope. CELEX and Golden Compound Analyses are both covered by Universal Derivations, which makes their handling easier.

MORBO/COMP (Guevara et al., 2006) is a database of compounds covering 23 languages, providing information consistent with the classifica-
tion of Bisetto and Scalise (2005). The database describes the part-of-speech of each compound as well as its constituents, centricity, syntactic head- edness, semantic head (if present), linking element, and gloss in English. Unfortunately, as of the writing of this proposal, the database itself is not public- ly available.

GermaNet (Henrich and Hinrichs, 2010; Hamp and Feldweg, 1997) is a database that relates Ger- man verbs, nouns, and adjectives. It currently con- tains 215,000 lexical units, of which 121,655 are split compounds.

UniMorph (Batsuren et al., 2022) is a huge- scale coordinated effort by a team of researchers from all over the world to build a collection of mor- phological resources covering 169 different lan- guages. As a result, its coverage varies wildly. For the purposes of this proposal, only the Spanish (42,825 derivatives; 130 compounds) and French (72,789 derivatives; 161 compounds) branches of UniMorph are relevant.

Wiktionary presents a massive amount of com- pounds for many languages. Unfortunately, the data resource is very inconsistently structured, and in practice, it is difficult to extract compounds and/or descriptions thereof safely.

3.2 Compound splitters and other tools

Here, we non-exhaustively list tools that model compounding by taking a compound word as input and returning its parent words in their lemma form as output.

DériF (Dérivation en Francais, Namer 2003) is a derivational analyzer for French. Its relevance lies in the fact that even though its stated purpose is derivational analysis, it is capable of analyzing neoclassical compounds and extracting their con- stituents. The analyzer is rule-based and recur- sively returns all derivational (or compositional, in the case of neoclassical compounding) ancestors until it hits an unmotivated word, along with a set of features and morphemes for each of them.

Khaitan et al. (2009) used n-gram statistical pattern matching to build a compound splitter for English. The splitter relies on finding split-points, i.e. it finds where in the given compound the boundary between the two constituents is located. This approach is adequate in English, where interfixes and parasynthetic compounds are relatively rare, but falls short in languages where these phenomena appear more often. For example, in German, insert-

4 Experiments in compound modeling

This Section presents the experiments that have already been conducted as part of the proposed dissertation thesis. They begin by computationally modeling Czech compounding in terms of their identity (deciding whether or not a given word is a compound) and parents (finding which word a given compound is motivated by).

4.1 Czech Compound Splitter

CCS (Svoboda and Ševčíková, 2021) is a neu- ral compound splitter that accepts a sequence of graphemes representing a Czech compound and return a sequence of graphemes representing its par- ents separated by spaces. Unlike Vodolazsky and Petrov, it is not restricted to any specific kind of compound, but unlike Vodolazsky and Petrov’s is purely neural. It can handle compounds with any number of parents. CCS performs two tasks:

- **Compound splitting.** A generative task. CCS returns two or more parent words for each compound fed into the model. If a non- compound is fed in, it returns the word un- changed.

- **Compound identification.** A binary classifi- cation task. CCS decides whether or not the given word is a compound.

4.1.1 Data

For all experiments described, a data set of 1,500 compounds from the DeriNet 2.0.5 (Vidra et al., 2019) word-formation resource was taken. All of
these had previously been labeled as compounds, but their parent words were not linked, so these had to be annotated by hand. Less than a hundred lexemes were dropped, because some had been annotated as compounds mistakenly (levopi-
mar, a medicine brand name), are borrowings from other languages such that they have no parents in Czech (face-up) or are derivatives of compounds (e.g. the adverb velechytrě derived from the adjective velechytrý ‘very clever’). We also used all non-compounds present in DeriNet 2.0.

4.1.2 Model building and performance

CCS was created by using the Marian machine translation framework developed by Microsoft (Junczys-Dowmunt et al., 2018) to build a model and train it. The Marian framework, being built for translation, requires the input sentences to be split into words or subwords before being fed into the model. Since our data consists of isolated words, they were split character-by-character as a workaround.

For the evaluation of CCS’s performance in compound splitting, we used Accuracy, which we define as the number of times CCS returned parents string-equivalent to the label parents divided by the number of items in the test set. Additionally, we defined Family accuracy, which is the number of times CCS returned parents which

- were string-equivalent to the label parents OR
- were ALL present in the same word-formation family as the label parents.

Whether or not the model output and label belong to the same family was checked using DeriNet. CCS scored an Accuracy of 54% and a Family accuracy of 55%. To contextualize these results, we built two baselines for comparison. One is a simple procedure that tries to find an -o- interfix in the middle third of the word and splits the input there or failing that, splits in the middle, which scored an Accuracy of 11% and a Family accuracy of 11%. The other is a phonologically weighted string similarity function dubbed IML(), which tries to find strings similar to the compound in DeriNet. It scored an Accuracy of 27% and a Family accuracy of 36%.

To evaluate CCS’s performance in compound identification, we used Balanced Accuracy from scikit-learn, which is defined as \((\text{Specificity} + \text{Sensitivity})/2\). We used this metric instead of regular Accuracy, as our dataset contains many more non-compounds than compounds. In compound identification, Czech Compound Splitter achieved a Balanced Accuracy of 84% and an F1-score of 81%. In compound splitting, Czech Compound Splitter achieved an Accuracy of 54%.

4.2 Word Formation Analyzer for Czech

WFA.ces (Svoboda and Ševčíková, 2022) arose as a successor to CCS from the observation that there was nothing stopping us from training the model to be able to return parents of derivatives in addition to parents of compounds. It followed naturally to also train it to identify unmotivated words and thus end up with a general model of word formation. WFA.ces performs two tasks:

- **Parent retrieval.** A generative task that returns two or more parent words for each compound fed into the model, one parent for each derivative, and returns the given word unchanged if it is unmotivated.

- **Word-formation classification.** A ternary classification task that generalizes the binary compound identification by also considering the derivative class.

4.2.1 Data

We used DeriNet 2.1 (Vidra, Jonáš and Žabokrtský, Zdeněk and Ševčíková, Magda and Straka, Milan, 2015) to create a data set of compounds, derivatives, and unmotivated words.

We added to the data as derivative those lexemes that have a single parent, are attested in the SYN2015 (Křen et al., 2016) corpus of Czech, and are not labeled as either unmotivated or compound. Similarly, we designated as unmotivated those lexemes that had no parents, were attested in the SYN2015 corpus of Czech, and were explicitly flagged as unmotivated. The compounds used were compounds from DeriNet with both parents linked.

<table>
<thead>
<tr>
<th>Class</th>
<th>Oracle</th>
<th>First best</th>
<th>Lexicon</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp</td>
<td>70%</td>
<td>56%</td>
<td>55%</td>
<td>57%</td>
</tr>
<tr>
<td>Deriv</td>
<td>87%</td>
<td>69%</td>
<td>75%</td>
<td>59%</td>
</tr>
<tr>
<td>Unmot</td>
<td>91%</td>
<td>71%</td>
<td>84%</td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td>83%</td>
<td>65%</td>
<td>71%</td>
<td>61%</td>
</tr>
</tbody>
</table>

Table 1: The accuracy scores of Word Formation Analyzer for Czech in the task of parent retrieval, broken up for each word formation class, as measured on the validation set for \(\text{nb}_\text{best} = 4\).
Additionally, the compounds hand-annotated for the purposes of training CCS were used, with an extra 285 hand-annotated compounds. The data was split into a train set (60%), a test set (20%), and a validation set (20%) according to the compound class, as it was the class with the least items. The unmotivated and derivative classes were split such that there was the same number of items from each of the classes in both the test and validation sets. The rest of the derivative items and unmotivated items were added to the train set.

4.2.2 Model architecture and performance

A model was built using the Marian framework in much the same manner as in CCS. But in addition to CCS, we added a feature where instead of simply returning a parent sequence, the Marian model returns a list of the top \( n_{\text{best}} \) parent sequences, and then a reranking function is used to select the best one according to some criterion. The best parent sequence is then considered the final output of WFA.ces. We evaluated WFA.ces with respect to four reranking functions:

First best: WFA.ces simply returns the first parent sequence in the list.

Lexicon: WFA.ces uses a provided lexicon to select the first parent sequence in the parent sequence list whose elements are all attestable in that lexicon. If none such sequence can be found in the list, it uses First best.

Frequency: WFA.ces uses Derinet 2.0 to assign a relative corpus frequency to each element in each sequence. It then selects the parent sequence with the smallest sum of squared frequencies.

Oracle: This method is only available if the ground truth is already known, and as such, it is only useful for the purpose of evaluation of the other reranking methods. It returns the correct result, if present in the sequence list.

The performance of WFA.ces can be viewed in Table 1. While the Lexicon reranking method gives the best results, it carries the drawback of preventing WFA.ces from returning parents that are not present in DeriNet. Thus, with the Lexicon reranking method, the tool cannot handle derivatives or compounds motivated by words not contained in DeriNet. Parent retrieval restricted to compounds, is equivalent to compound splitting; when tested on compound splitting, WFA.ces exhibits an accuracy of 57%, which is three percentage points more than Czech Compound Splitter.

4.3 PaReNT

4.3.1 Model building and performance

The experiment series culminates in PaReNT (Parent Retrieval Neural Tool), a multilingual neural model of word formation. The tool performs the same tasks as WFA.ces, but covers all of the 8 languages in scope.

4.3.2 Data

It uses data aggregated from a range of resources, which can be viewed in Table 2. The data was split 60/20/20 into training, evaluation, and development subsets, but unlike in the case of CCS and WFA.ces, where it was done by the compound class, here it was done by so-called lexicographical block. This means that if information about the word-formation family was available (as is the case with UDer datasets), then each family had to fall into one of the subsets in its entirety. If the information was unavailable, then the items were grouped by the first three graphemes of the rightmost parent. This ensures that it was the model’s capability
to learn morphological patternings that was being measured, as opposed to recognizing already-seen word stems.

Unlike in the case of CCS and WFA.ces, we used TensorFlow (Abadi et al., 2015) to build a customized model architecture. The model utilizes multilingual semantic vector embedding provided by Heinzerling and Strube (2017), in parallel to character-level representation similar to CCS and WFA.ces. In addition, it now has two output heads – a generative Retriever head for parent retrieval and a Classifier head for word-formation classification – removing the need for the workaround solution of counting spaces in the output. The entirety of the model’s architecture can be viewed in Figure 1.

PaReNT was evaluated using Family accuracy only on Czech, since DeriNet 2.1 is the only resource at our disposal with the necessary structure and completeness. The tool was compared against two baselines. The Dummy baseline performs parent retrieval by returning the input unchanged, always guessing Unmotivated as the word-formation category. The other baseline was ChatGPT (OpenAI, 2023), which was given the following prompt:

Perform parent retrieval (predict which word or words the input lemma is motivated by.) and word formation classification (predict whether the input lemma is a compound, a derivative, or unmotivated) on the given words. For each word, you will also be given its language of origin as a language token {cs : Czech, ru : Russian, de : German, es : Spanish, fr : French, nl : Dutch, en : English}. Format the output as tsv.

The words:
<list of words>

ChatGPT formats the output differently on each query, misunderstands the task, or even outright refuses to perform it at all. Its output has to be manually checked, regenerated if needed, and then reformatted. As a result, the evaluation of Chat-GPT was performed on a small random subset (n = 300) of the development set. The subset was fed into ChatGPT in increments of 100 words, prepended by the prompt each time. The performance of PaReNT can be viewed in Table 3. The dummy balanced accuracy in classification is 0.33 for each language because there are 3 word-formation categories. Dummy accuracy for retrieval is the same as the proportion of unmotivated words in the given language’s test set. The figure in (parentheses) on the second line indicates Family accuracy, which describes how many times the system correctly identified the Czech word formation family of the correct parent(s). It is not listed for the Dummy model, because it always returns the word unchanged, and a word is trivially part of its own word family in 100% of cases.

4.3.3 Error analysis of PaReNT

Here we present a number of errors that PaReNT typically made, with analysis as to how that may have occurred. In the examples, the arrows point to the right, to indicate that the input into PaReNT is on the left and the output is on the right.

Type 1: Data conflict

The output of the model is correct in some sense, but conflicts with the label in the data.

In the data, each lexeme is assigned a single set of parents. The parents of a lexeme can however be assigned in different ways – not to mention that human error may appear. As a result, the model sometimes returns a lexeme that is correct but disagrees with the label. The consistency of this error’s appearance is reflected by the fact that Family accuracy is considerably higher than raw Accuracy, as shown in Section 4.

Type 2: Inflectional confusion

The model mishandles the contribution of inflection to word formation, which has been touched upon in Section 2. When womenfolk specifically is fed into the model, it fails to return woman and instead returns women. Similarly, it often imputes an inflectional ending in a context where it looks

<table>
<thead>
<tr>
<th>Language</th>
<th>PaReNT Retrieval acc</th>
<th>Class bal acc</th>
<th>Dummy Retrieval acc</th>
<th>Class bal acc</th>
<th>ChatGPT Retrieval acc</th>
<th>Class bal acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>0.64 (0.79)</td>
<td>0.62</td>
<td>0.05 (N/A)</td>
<td>0.33</td>
<td>0.36 (0.66)</td>
<td>0.36</td>
</tr>
<tr>
<td>German</td>
<td>0.66</td>
<td>0.82</td>
<td>0.06</td>
<td>0.33</td>
<td>0.28</td>
<td>0.71</td>
</tr>
<tr>
<td>English</td>
<td>0.72</td>
<td>0.85</td>
<td>0.27</td>
<td>0.33</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>French</td>
<td>0.45</td>
<td>0.55</td>
<td>0.10</td>
<td>0.33</td>
<td>0.4</td>
<td>0.31</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.67</td>
<td>0.86</td>
<td>0.10</td>
<td>0.33</td>
<td>0.11</td>
<td>0.69</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.76</td>
<td>0.96</td>
<td>0.17</td>
<td>0.33</td>
<td>0.3</td>
<td>0.64</td>
</tr>
<tr>
<td>Russian</td>
<td>0.70</td>
<td>0.99</td>
<td>0.09</td>
<td>0.33</td>
<td>0.25</td>
<td>0.63</td>
</tr>
<tr>
<td>Mean</td>
<td>0.66</td>
<td>0.81</td>
<td>0.12</td>
<td>0.33</td>
<td>0.45</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 3: The performance of PaReNT and baselines for each language.
like it could have been dropped (ex. 17).

(17) stearino\vý  \rightarrow  *stearina \(\text{CZ}\)

(18) holekchtivé  \rightarrow  holek  \text{girl-wanting.A}  +  chtivý  \text{wanting.A}  \(\text{CZ}\)

Type 3: Morphological ambiguity

The model fails to compensate for a difficult-to-account-for morphological process. The bulk of the model’s predictions seem to be based on the retroactive application of word-formation rules. Sometimes it is however unclear how the rule should be retroactively applied due to morphophonological changes that merge two phonemes. As an example, in Czech, the addition of a diminutive suffix can induce stem allomorphy, resulting in /u/ \rightarrow /ou/. The application of the same suffix on another word, however, can yield /u/ \rightarrow /u/. As a result, it sometimes guesses wrong, as in ex. 19, where the expected result should be ubrus “tablecloth”.

(19) ubrousek  \rightarrow  *ubrous \(\text{CZ}\)

Type 4: Neural hallucination

The model baselessly hallucinates non-existent structures. Sometimes, the model for unclear reasons simply switches two characters, generates an extra character, or does something else that is difficult to interpret. Occasionally, it even hallucinates what seems to be an entire morpheme (ex. 20; should be dèrmis “dermis”).

(20) dèrmico  \rightarrow  *dermiar \(\text{ES}\)

dermal.A

Type 5: Overretrieval

The model does not return the parent of the input, but the parent of the parent of the input. In example 21, we would expect PaReNT to output Prüfung “exam” rather than its prüfen “to test”. This error also contributes to the rather high Family accuracy of PaReNT.

(21) Teilprüfung  \rightarrow  Teil  +  prüfen \(\text{DE}\)


Type 6: False morphemes

The model misjudges the presence of a morpheme, typically (but not exclusively) due to loaning. In ex. 22 PaReNT noticed that the um/eum originally Latin ending gets frequently dropped across many of the languages in the set, However, it does not behave like a morpheme in this case, so the expected result would be Jagd Museum “hunting museum”.

(22) Jagdmuseum  \rightarrow  *Jagdmuse \(\text{DE}\)
museum of hunting.N
A promising way to apply structure would be to adopt the dependency structures used in Universal Dependencies (UD; Nivre et al. 2020), which is a multilingual effort to create sentence-level dependency treebanks. We believe it is possible to automatically build word-level dependency trees over segmented compounds by using an extended UDeR tagset. Some of the tools needed for this task already exist, and some would need to be developed. The task requires: 

- **a)** morphological segmentation, 
- **b)** morpheme classification, 
- **c)** root morph POS-tagging, and 
- **d)** root morph relation tagging (coordinate, subordinate, attributive; see Section 2). 

The task of **a)** multilingual morphological segmentation has been covered, both by data sets like UniSegments (Žabokrtský et al., 2022) and tools like Morfessor (Smit et al., 2014). The task of **b)** morpheme classification is being developed by John and Žabokrtský (2023). MorphoDiTa (Straková et al., 2014) could be used in conjunction with PaReNT for **c)** root-morph POS-tagging, leveraging the fact that the number and order of root morphemes is the same as the number and order of parent words, so the POS tags of the parents can be transferred onto the roots. Finally, **d)** automatically labeling syntactic relations between root morphs will require dedicated development. 

The projected results can be viewed in Figure 2. We see that each morph is tagged as to whether it is a root, infix, suffix, prefix, or ending; a dependency is established for each element. Relations from root morph to other root morphs are labeled according to Bisetto and Scalise (2005), and root morphs are assigned POS based on the parents of the given compounds.

**Type 7: Semantic irrelevance**

The model retrieves a word in a formally correct manner, but in a way that no human would ever find useful or meaningful.

The problem in ex. 23 is that while all three of the returned words are relevant, the given compound is not a compound of three words, but rather a recursive compound. A human knows that, because the second parent *Fahrzeug* is a common word meaning “vehicle”, but the model has no way of knowing this. Similarly, in ex. 24, Czech speakers do not find the concept of *přidrznout* “to become a bit cheeky” to be useful enough to warrant the existence of its own lexeme. However, the neural model probably returned *přidrznout* by analogy with the paradigm *omrzlý* “frostbitten” → *omrznout* “to get frostbite”.

(23) *Straßenfahrzeug* → *Straße* + *fahren* + *Zeug* (DE) 

(24) *přidrznout* a bit cheeky.A → *přidrznout* (CZ) to become a bit cheeky V

**5 Future work and conclusion**

So far, we have only presented ways of modeling compounds in terms of their motivating words. However, we would like to go deeper and represent their internal structure in terms of their actual constituent elements with explicit relationships between them. Compounds, like any other words, can be broken up into morphemes or morphs, or morphologically segmented. It follows naturally that such a flat sequence of morph(eme)s could be endowed with some sort of structure describing how these morph(eme)s relate to one another.

In this thesis proposal, we presented a series of experiments that multilingually model some aspects of compounds, namely their status as compounds as opposed to derivatives or unmotivated words, and establish which words they were motivated by. Additionally, we proposed a pipeline that extends these capabilities by also modeling the internal structure of compounds by way of automatically building constituency trees over morphologically segmented compounds in a way that is compatible with theoretical classifications set forth in the literature. We believe that these results when developed will be sufficient for a full dissertation.
References


Laurie Bauer. 1998. Is there a class of neoclassical compounds, and if so is it productive? Linguistics, 36(3):403–422.


OpenAI. 2023. ChatGPT.


Vidra, Jonáš and Žabokrtský, Zdeněk and Ševčíková, Magda and Straka, Milan. 2015. DeriNet 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

