Compounds in Universal Dependencies: A Survey in Five European Languages

Emil Svoboda, Magda Ševčíková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{svoboda,sevcikova}@ufal.mff.cuni.cz

Abstract

In Universal Dependencies, compounds, which we understand as words containing two or more roots, are represented according to tokenization, which reflects the orthographic conventions of the language. A closed compound corresponds to a single word in Universal Dependencies (e.g. *waterfall*) while a hyphenated compound (*father-in-law*) and an open compound (*apple pie*) to multiple words. The aim of this paper is to open a discussion on how to move towards a more consistent annotation of compounds. The solution we argue for is to represent the internal structure of all compound types analogously to syntactic phrases, which would not only increase the comparability of compounding within and across languages, but also allow comparisons of compounds and syntactic phrases.

1 Introduction

Compounding, as a word-formation process in which two or more words (bases, roots, or stems) are combined to form a new word (Lieber, 2010, p. 43), is used across languages (Štekauer et al., 2012, pp. 51–100). However, the term compound is not only used to refer to words that result from the combination of two words (cf. *flowerpot*) or are outputs of recursive compounding (e.g. German *Jahresabschlussprüfung* 'end-of-the-year audit'), but also to words that are results of compounding happening in conjunction with derivation or conversion (e.g. the German adjective *blauäugig* 'blue-eyed'),[1] and to words that are both direct and indirect derivatives of these compounds (German *Blauäugigkeit* 'blue-eyedness/naiveté'); cf. Bauer et al. (2013, p. 442).

The criteria for defining compounds (and especially distinguishing them from syntactic phrases) vary from language to language, but features with cross-linguistic validity include, besides the requirement of at least two roots, syntactic and semantic compactness. What is not decisive, on the other hand, is spelling. Compounds are spelled as a single word (closed compounds; e.g. *waterfall*), or as several orthographic words joined by hyphens (hyphenated compounds; e.g. *cyan-magenta-yellow-key*) or separated by spaces (open compounds; e.g. *apple pie*).

The present paper surveys how compounds are treated in Universal Dependencies (UD; version 2.12, Zeman et al. 2023). Five languages, namely English, German, Czech, Latin, and Russian, have been chosen for this pilot survey based on the working criteria that for each of the languages (a) at least one treebank is available in UD, (b) a lexical database exists that contains a non-negligible number of compounds (and can be used to identify compounds in the treebanks), and (c) the authors have a sufficient command of it. We show that the current treatment of compounds in UD, which is determined by the languages' orthographic conventions and by UD's tokenization rules, renders compounds difficult to identify in the data, hindering their comparison within and across languages. However, this paper is not limited to the mere unification of compound annotation according to the existing guidelines. Our proposal is to annotate the relations between the compound's component parts by using the syntactic relations already implemented in UD, making the analogy between compounds and multi-word expressions and syntactic phrases explicit, which has already been pointed out in the literature.

The paper is structured as follows. Section 2 briefly summarizes those aspects of the linguistic discussion on compounding that are necessary for understanding the issues presented. An overview

[1] The compound cannot be traced back to *blau* 'blue' and **äugig* '*eyed', because the latter item does not exist in isolation. It is rather analysed as being formed by combining the adjective *blau* 'blue' and the noun *Auge* 'eye' and simultaneously adding the *-ed* suffix to get the compound adjective.

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 88–99
March 22, 2024 ©2024 Association for Computational Linguistics

of the language data resources that contain compounds and are used in the paper is also provided. In Section 3, we describe how compounds are currently handled in UD, exemplifying the general and language-specific problems of compounds. In Section 4, we discuss steps that can be taken to make the annotation of compounds more coherent and to bring it closer to the way syntactic relations are annotated, but without losing the difference between compounding and syntax. Future directions regarding the automation of compound identification and annotation are outlined to some extent. Section 5 concludes the paper.

2 Background

2.1 Compounds in the linguistic literature

Besides the spelling differences mentioned above, the debate over compounding and compounds has been centered around the following topics:

– boundary between compounding vs. derivation, with a special focus on neo-classical formations (cf. ten Hacken 1994, Bauer 2005, among others), and between compounds vs. syntactic phrases and multi-word expressions in particular (Olsen, 2001; Schlücker, 2019);

– part-of-speech (POS) category of the compound and its components: if the components obtained by splitting the compound do not correspond to independently existing words, the POS of the component is determined according to the closest word; if this applies to the head, the compound's POS is different from its head's POS (cf. the distinctions below; for examples, see Section 2.2);

– headedness: if one of the components plays a prominent role, it is considered the head; left-headed compounds and right-headed compounds are distinguished;

– endocentricity vs. exocentricity: the head determines the POS and meaning in endocentric compounds; an exocentric compound is headless or, as Bauer (2001, p. 70) puts it, it is "a compound which is not a hyponym of its own head element";

– relations between the compound's components: in the literature cited below, the compound's internal structure is indicated by brackets, in analogy to syntactic constituent trees;

– syntactic type of the relation between the compound parts: the crucial distinction is whether the components are independent of each other (coordinate, coordinative, additive or copulative are some of the terms used) or whether one depends on the other (subordinate, determinative, etc.).

These features, assigned varying degrees of importance and priority, have been employed to classify compounds. The classifications proposed by Bloomfield (1933), Bally (1944), Marchand (1969), Spencer (1991), Fabb (1998), Olsen (2001), Haspelmath (2002), Bauer (2001), and Booij (2005) are compared by Bisetto and Scalise (2005), who come up with yet another classification, where the relation between the components is used as the first-level criterion[2] and it is followed by the distinction between endocentric and exocentric compounds. Bisetto and Scalise's classification was implemented in annotation scheme of the MorboComp database, which is one of the resources reported on below.

2.2 Compounds in language data resources

The selective list presented here contains language data sources that include a substantial number of compounds along with annotations reflecting various features discussed in the literature.

MorboComp is a multilingual database of compounds covering 20 languages, including the ones in scope except for Czech (Guevara et al., 2006). In Table 1, the annotation provided in MorboComp is exemplified by three nominal Italian compounds composed of words from different POS categories (cf. 2nd and 3rd column). While the first compound (*madrelingua* 'mother tongue') is endocentric with the right component playing the role of head, the latter two are exocentric (and headless). The components are listed as they occur in the compound (8th and 9th column), they may not be existing words (cf. the third compound in the table). While potentially highly useful for the purposes of this paper, as of 2023 the project seems to have been discontinued and the data are not publicly available.

Compounds are also covered by CELEX2, which is a lexical database of English, German, and Dutch (Baayen et al., 2014). Out of all the linguistic annotations provided in this resource, delimitation of the components (and the linking element, interfix, if present), POS of the components, and annotation of the internal structure using nested brackets (cf. (1) to (3)) were the most important for our survey. In the bracketed structures in German, some

[2] The authors speak of grammatical relations: "The grammatical relations holding between the two constituents of a compound are basically the relations that hold in syntactic constructions: subordination, coordination and attribution".

Compound	POS	Struc	Class	End	Head-C	Head-S	1st-C	2nd-C	Gloss
madrelingua	N	[N+N]	SUB	Tru	right	right	madre	lingua	mother+tongue
mano lesta	N	[N+A]	ATT	Fal	none	none	mano	lesta	quick+hand = thief
dormiveglia	N	[V+V]	CRD	Fal	none	none	dormi	veglia	sleep+be awake = dozing

Table 1: Annotation of Italian compounds in the MorboComp database. The compound's lemma (1st column) is followed by its POS category (2nd column), the POS categories of the components (column Struct[ure]), syntactic relation between the components (Class: subordinate/attributive/coordinate), endocentricity (End[ocentric]: True/False), placement of the semantic head (Head-C), placement of the syntactic head (H-S), the form of the first component (1st-C) and of the second one (2nd-C), and the gloss.

morphs are replaced with a representative form (cf. *gang* substituted by *geh*, which occurs in the infinitive *gehen* 'to go' in (1); but in the English example (3) *woman* is not used instead of *women*). Based on these features, $19,304$ compounds were extracted from the German section of CELEX and $6,267$ compounds from the resource's English section.

(1) `Umgangssprache ... Umgang+s+Sprache NxN ... ((((um)[V|.V],(geh)[V])[V])[N], (s)[N|N.N],((sprech)[V])[N])[N] ...`

(2) `Grossmachtpolitik ... Grossmacht+Politik NN ... (((gross)[A],(Macht)[N])[N], ((polit)[R],(ik)[N|R.])[N])[N] ...`

(3) `womenfolk ... women+folk NN ((women)[N],(folk)[N])[N] ...`

The GermaNet compound list (Henrich and Hinrichs, 2011) contains more than 120,000 compounds in its 2023 edition. This source lists for each compound the lemmas of two immediate ancestors from which it was composed ((4) to (6)). The ancestors provided are existing words, not just strings occurring in the compound (cf. (5) where the verb *abbiegen* 'to turn' is given, because **Abbiege* is not a separate word in German). Compounds with more than two roots are split in succession; see (6) where the second ancestor is a compound which is analyzed in a separate entry in the resource. For the first component, two possibilities are given, if both are equally relevant (cf. the action noun *Umfrage* 'survey' and the verb *umfragen* 'to survey' in (6)).

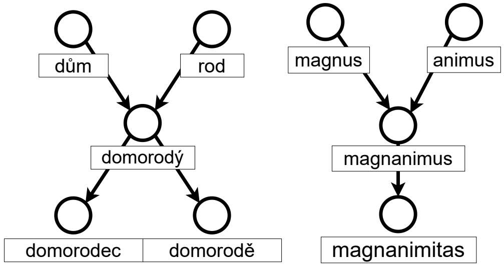
(4) `Umgangssprache Umgang Sprache`

(5) `Abbiegeassistent abbiegen Assistent`

(6) `Umfrageteilnehmer`
`Umfrage|umfragen Teilnehmer`

DeriNet is a lexical database of Czech where words that share a common root are arranged into tree-like graphs according to their morphological structure – from the morphologically simplest words (unmotivated words) to the most complex. The database contains over a million entries, of which less than a half are corpus-attested (432 thousand; only this subset is used in this study). While derivatives are linked to a single ancestor, compounds are connected to two or more ancestors. Additional compounds were identified based on heuristics and lexical lists of compound parts. When the compounds both with and without the links to their ancestors are counted (all of them having the explicit Boolean compoundhood flag set to `true`) together with the derivatives of all these compounds, the number totals to 45 thousand corpus-attested compounds available in DeriNet 2.1 (Vidra et al., 2021). The left graph in (7) shows the unmotivated nouns *dům* 'house' and *rod* 'kin' as ancestors of the adjectival compound *domorodý* 'native', from which the noun *domorodec* 'native man' and the adverb *domorodě* 'in a native way' are derived. All of *domorodý*, *domorodec* and *domorodě* are counted as compounds.

(7)



More than 3 thousand Latin compounds and their derivatives are part of the Word Formation Latin database (Litta et al., 2016). The database is organized in a way similar to DeriNet; cf. the right graph in (7) modeling the Latin adjective *magnan-*

Dataset	Language	Compounds	Total entries
CELEX (Baayen et al., 2014)	English	6,267	52,447
CELEX (Baayen et al., 2014)	German	19,304	51,728
GermaNet (Henrich and Hinrichs, 2011)	German	121,655	215,000
Derinet 2.1 (Vidra et al., 2021)	Czech	45,473	431,857
Word Formation Latin (Litta et al., 2016)	Latin	3,198	36,258
Golden Compound Analyses (Vodolazsky and Petrov, 2021)	Russian	1,699	1,699

Table 2: The databases employed in the present survey for identification of compounds in the Universal Dependencies treebanks of the five languages. The last two columns specify the number of lemmas (types).

imus 'high-spirited' as being formed by combining the adjective *magnus* 'high' and the noun *animus* 'spirit', and giving rise to the noun *magnanimitas* 'high-spiritedness'.

Golden Compound Analyses (Vodolazsky and Petrov, 2021) is a database of Russian compounds compiled for training of a compound splitter. It contains $1,699$ compounds that a re directly traced back to two or more ancestors. The annotation includes the POS category of each compound, the lemmas and POS of each of the components; cf. полувсерьёз 'half serious' in (8).

(8) полувсерьёз,adv,половина,noun,всерьёз,adv

The sources introduced in this section are, with the exception of MorboComp, further used in this survey to gain preliminary quantitative insights into how many compounds are found in the UD treebanks; cf. Table 2 for a summary.

3 Current annotation of compounds in Universal Dependencies

3.1 The annotation guidelines

We start by introducing how words considered as compounds in the literature are treated according to the UD annotation principles (de Marneffe et al., 2021).[3] The application of these rules to each of the languages under survey is described in the following subsections. Syntactic annotation in UD is based on tokenization, which in turn follows the spelling conventions of individual languages. Since the term compound covers words spelled in several ways, compounds are not annotated uniformly in UD:

– Closed compounds, appearing in the text as continuous orthographic words, are handled as discrete, internally unstructured (= atomic) items which enter into relations with other items of the sentence structure. Although the compound's components are linked by similar relations as the constituents of syntactic phrases, these intra-word relations are not captured in UD because "there is no attempt at segmenting words into morphemes".[4]

– Open compounds, which are spelled as two (or more) separate words, are treated as two (or more) items that are arranged into a subtree with the head component as the root and the less prominent item(s) as dependent node(s). The relation between the head and the other component is labeled with the dedicated syntactic relation compound. This relation is assigned to open compounds regardless of the semantic relation between the components (cf. *apple pie* = "pie made from apples" vs. *coffee cup* = "cup for coffee" vs. *water mill* = "mill powered by water", etc.). Besides the bare compound relation, there are 22 subtypes of this relation intended for language-specific phenomena,[5] of which only compound:prt is used in some languages under analysis, namely in English and German. The compound:prt is used for "[p]article verbs where the particle is realized as a separate word (which may alternate with affixed particles), for example Swedish *byta ut* ('exchange'; cf. *utbytt* 'exchanged')".

– Hyphenated compounds are treated in the same way as in open compounds. The hyphen is attached to the head, with the relation label punct.[6]

Annotation of compounds is explored for each language based on all treebanks available in the UD collection (i.e. ten treebanks for English with a total of 46K sentences, four German treebanks containing 208K sentences, six treebanks for Czech with 208K sentences, five Latin treebanks with

[3] See also https://universaldependencies.org/guidelines.html

[4] https://universaldependencies.org/u/overview/tokenization.html

[5] https://universaldependencies.org/ext-dep-index.html

[6] This is the case for the languages in scope, but the claim does not hold for all languages in UD. Swedish hyphenated compounds are for instance handled the same way as closed compounds.

Language	compound relations	Sentences with compound	compound:prt relations	Sentences with compound:prt	Total words	Total sent.
English	22,017 (3,03%)	13,459 (29.27%)	2,485 (0.34%)	2,313 (5.0%)	726K	46K
German	1,787 (0.05%)	1,418 (0.68%)	22,349 (0.59%)	21,897 (10.5%)	3,810K	208K
Czech	2,690 (0.12%)	1,356 (1.06%)	0 (0.00%)	0 (0.0%)	2,222K	128K
Latin	85 (0.01%)	82 (0.1%)	0 (0.00%)	0 (0.0%)	983K	59K
Russian	1,973 (0.11%)	1,812 (1.6%)	0 (0.00%)	0 (0,0%)	1,830K	111K

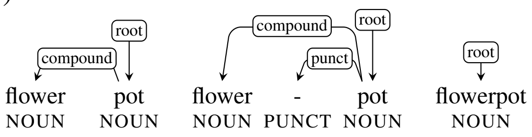
Table 3: The number of sentences containing a compound relation (assigned to open and hyphenated compounds) and sentences with the compound:prt label (with particle verbs) in the Universal Dependencies treebanks of the five languages. The percentage indicates the proportion of sentences with the labels in all sentences of the language's treebanks.

59K sentences, and five treebanks for Russian with 111K sentences). The number of sentences containing the compound relation in the languages' UD treebanks is listed in Table 3. The compound:prt relation is used only in English and German; it will not be further commented upon.

3.2 The UD treebanks for English

Out of the languages analyzed, English treebanks contain the highest number of compound relations, both in absolute numbers and in percentages, owing to the fact that in this language, NOUN+NOUN sequences are analyzed as compounds. English is also a language where these NOUN+NOUN compounds can alternatively be spelled with a hyphen or even without a space as a single graphical word (cf, Table 4), resulting in different tree structures; cf. the textbook example *flower pot* as an open compound with the hyphenated (*flower-pot*) and closed spelling alternative (*flowerpot*) annotated in line with the UD guidelines in (9).

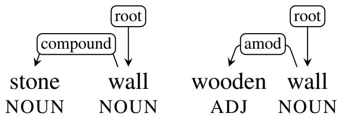
(9)



The compound relation is also assigned to NOUN+ADJ phrases (*emerald green*, *labour intensive*), as well as complex open numerals such as *twenty one*.

Even though the relationship between the components of the open compound *stone wall*, which can be paraphrased as "wall of stones", is the same as the relationship between the adjective *wooden* and the noun *wall* ("wall of wood"), the syntactic relations within these sequences are labeled differently, namely compound in the first sequence while amod in the second; cf. (10).

(10)



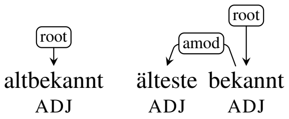
If there were an adjective to the noun *stone* (**stonen*) or if *stone* were considered also as an adjective in English, the annotation would have been no different from *wooden wall*. This is encountered in the phrase *west side*, where *west* is interpreted as an adjective (while the formally identical noun *west* and the formally different adjective *western* exist) and therefore handled as an adjectival modifier (amod) of the nominal governor.

3.3 The UD treebanks for German

German is a language where compounding is widely used, but compounds are typically spelled as compact strings. Nevertheless, both hyphenated compounds (cf. the Anglicism *Trackpad-Click*) and open compounds (NOUN+NOUN sequences, often with proper names; e.g. *Präsident Franjo* 'President Franjo') are documented in the treebanks, both types assigned the compound relation.

In German we also find cases of (here, closed) compounds with the components' relations analogous to those between words in syntactic phrases, but these analogies are not obvious in the current annotation; cf. the compound *altbekannt* 'well-known', which is represented by a single node, and the phrase *älteste bekannt* 'oldest known', which is represented as a tree headed by the second word with the first element linked by the amod relation in (11).

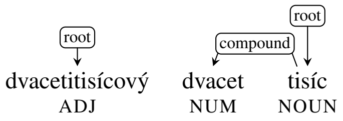
(11)



3.4 The UD treebanks for Czech

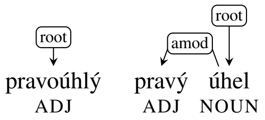
Also in Czech, compounds are commonly written as continuous strings, still a hyphen may connect the components in coordinate compounds. In the data, however, the compound relation appears not only with hyphenated compounds (*indo-australský* 'Indo-Australian'), but also with numeral expressions, which in Czech are separated by spaces.[7] The rightmost component is taken as the head and the other parts are depending on it as modifiers; cf. the right structure in (12). When a numeral construction enters derivation, the output is a closed compound and it is represented by a single node; cf. the adjective *dvacetitisícový* 'twenty-thousand' on the left in (12) which is traced back to the phrase *dvacet tisíc* 'twenty thousand'.

(12)



Similarly, nouns modified by adjectival modifiers can give rise to adjectives with two roots and closed spelling. Cf. the noun phrase *pravý úhel* 'right angle' and the adjectival compound *pravoúhlý* 'right-angled' in (13), which is close to the German adjective *blauäugig* 'blue-eyed' mentioned in the introductory section in that the right component does not exist as a separate adjective (**úhlý* 'angled' similar to **äugig* '*eyed').

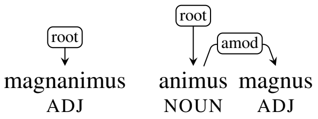
(13)



3.5 The UD treebanks for Latin

Latin treebanks contain the lowest number of compound relations, as documented in Table 3. Its current usage is limited to numeral expressions if they are spelled as separate words in a way described above for Czech, with the addendum that sometimes one of the words is *unus* 'one' labeled as a determiner and not a numeral. Example (14) is also analogous to Czech, documenting an adjectival compound (*magnanimus* 'high-spirited') that is based on a noun phrase (here, more specifically, on a phrase with the head noun preceding the adjectival modifier: *animus magnus* lit. 'spirit high' = 'high spirit').

(14)



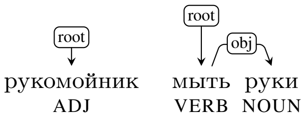
3.6 The UD treebanks for Russian

In the Russian treebanks, the compound relation is – unlike in Czech – applied to "noun compounds (e.g., стресс менеджмент 'stress management, Жар птица 'Fire bird'), but also adjective compounds (e.g., бэд блоки 'bad blocks', мини колонка 'mini speaker', Гранд отель 'Grand hotel') and some other types ("+ 1", "№ 1")".[8] Such NOUN+NOUN compounds and ADJ+NOUN compounds are often loanwords or direct translations of foreign expressions.

In addition, now similarly to Czech and also Latin, the compound relation appears also with numerals (две тысячи 'two thousand') and hyphenated constructions (город-государство; 'city-state').

Noteworthy are compounds which are analyzed as NOUN+VERB structures in the Golden Compound Analyses database. Since they are closed compounds, they are currently represented by a single node in the treebanks, but the relationship between the components resembles the obj relation of the object noun to its governing verb; cf. рукомойник 'washbasin' and the phrase мыть руки 'to wash hands' in (15), or короед 'bark beetle' traced back to есть кору 'to eat bark' and травосеяние 'grass sowing' related to сеять траву 'to sow grass'.

(15)



4 A proposal of a syntax-based annotation of compounds

4.1 Covering all types of compounds and annotating their internal structure

As we have tried to show, the current annotation does not allow to get a complex picture of compounds (as multi-root items) either within one language or across languages. On the one hand, the compound relation only applies to open and hyphenated compounds while closed compounds are

[7] The interpretation of numerals as compounds, though, does not conform to the Czech linguistic tradition.

[8] https://universaldependencies.org/ru/dep/compound.html

not marked in any way. On the other hand, the compound relation is underspecified, without capturing the different relations observed between the components in individual compounds – the exact same label is used for English NOUN+NOUN compounds, which themselves document a variety of internal relationships, and for relations between numerals in Czech, for example.

We now roughly outline a preliminary proposal for a new annotation of compounds in UD that should overcome these issues. Rather than offering an ultimate solution to each individual aspect of compound annotation, we present in our proposal one or more possible solutions based on what we have encountered in the literature or in existing language resources, with our primary goal being to initiate a discussion on this topic.

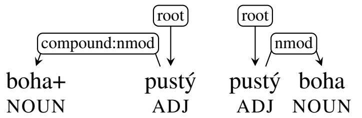
Compounds with all types of spelling should be approached as complex structures that consist of components which are linked by a relationship that is often similar to syntactic relations between words in syntactic phrases:

(a) Closed compounds should be split into their respective constituents for this purpose, and further handled in the same manner as open and hyphenated compounds. Compounds with three and more components will be divided into individual parts (e.g. the above German example *Umfrageteilnehmer* 'survey participant' into *Umfrage+ Teil+ Nehmer*) and their relationships will be captured by arranging them into a tree structure (see the next points). As illustrated, in closed compounds a "+" sign may be used on the first (or on all non-final) components to indicate the original morphological boundary, so that the information on their orthography is retained. An interfix, if contained in a compound, will be part of the preceding component (cf. *Umgangssprache* 'colloquial language' as *Umgangs+ Sprache*).

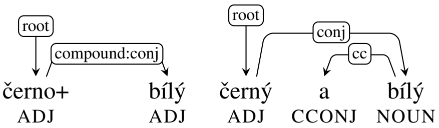
(b) Since such an approach would yield strings that do not exist as separate words (cf. **Abbiege* in *Abbiegeassistent*), we propose – in accordance with the fact that the words in syntactic phrases are treated in this way – to assign a lemma to each component. It can be a full word that is identical with the component (i.e. *Umgang* 'dealing' or *umgehen* 'to deal' and *Sprache* 'language' for *Umgangs+ Sprache*) or close to it (*abbiegen* 'to turn' and *Assistent* 'assistant' for *Abbiege+ Assistent*). Derivatives of compounds would share this lemmatization with their ancestors, e.g. *domorodec* 'native man' would be lemmatized as *domo+ rodý* 'native' (i.e. *dům* 'house' and *rod* 'kin').

(c) All types of compounds should be organized into subtrees in a way analogous to syntactic phrases in UD, making a distinction between subordinate compounds (with the compound's head as the governor and its modifier as its dependent; cf. *bohapustý* 'godless' in (16)) and coordinate compounds (with the first component as the root of the subtree and all the other conjuncts depending on it; cf. *černobílý* 'black-and-white' in (17)).

(16)



(17)



(d) Though the subtree modeling the syntactic structure of a compound's components is proposed to be as close an analogy as possible to the subtrees of syntactic phrases, the relation may retain the compound/phrase distinction. As bare compound relations are not informative, the relations within compounds could be tagged with a compound:<relation> label, where <relation> is an already-existing UD syntactic relation. This restriction regarding forcing compound subtypes into established relations should pertain solely to a) currently bare compound relations and b) closed compounds currently treated as atomic units, **not** to established, already-subtyped relations such as the compound:prt mentioned in Section 3.1. These should not be overwritten, their further usage is neither blocked nor discouraged by our proposal.

How these individual pieces of annotation could be brought into the data is discussed in the next section.

4.2 Steps towards the proposed annotation

Identification of closed compounds. To get a preliminary idea of which part of the treebank data for individual languages would be affected by the proposed annotation, the number of closed compounds in the UD treebanks needs to be estimated in addition to the number of the compound relations (which are in Table 3). In this study, we used the lists of compounds contained in the language resources discussed above in Section 2.2. The figures in Table 4 are heavily conditioned by the size of the resources used. The figures represent a lower

Language	Closed compounds	Total words	Sentences with closed compounds	Total sentences
English	5,934 (0.82%)	726K	5,286 (11.57%)	46K
German	156,629 (4.11%)	3,810K	87,104 (50.14%)	208K
Czech	47,103 (2.11%)	2,222K	34,775 (27.27%)	128K
Latin	26,271 (2.62%)	983K	18,353 (31.27%)	59K
Russian	4,803 (0.27%)	1,830K	4,460 (4.00%)	111K

Table 4: A lower bound estimate of the amount of closed compounds (tokens) in Universal Dependencies, based on searching for the known compounds (and their derivatives) extracted from the data sources listed in Table 2.

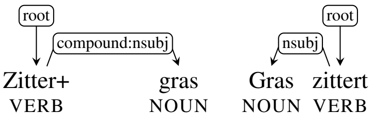
bound for the actual amount of closed compounds contained in UD, since none of the data sources list the compounds from their respective languages exhaustively.

With these limitations in mind, Table 4 suggests that the influence of such a change would be substantial, especially in German, where more than 156 thousand closed compounds were identified, which are part of 87 thousand sentences (i.e. 50% of all sentences). The least affected language by our current estimate would be Russian with less than 5 thousand closed compounds distributed over 4 thousand (4%) sentences; this is due to the relatively low coverage of the Golden Compound Analyses database used as the Russian compound data source in this study (see Table 2). The utilization of resources with higher coverage or another more sophisticated approach could render these numbers substantially higher.

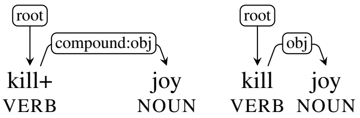
For **splitting of compounds and lemmatization of the components**, the language data sources reviewed above can be taken as a starting point, because they contain high-quality, linguistically adequate material. Whereas CELEX both divides the compounds into substrings and assigns representative forms to its individual parts (cf. *geh* for *gang* above), the other resources provide full-fledged ancestors for compounds that would fit our idea of components' lemmas. Even if the resources for some languages are limited, the existing data can – after unifying the annotation according to the proposal – be used for training automatic tools. A prototype of such a tool, *PaReNT* (Svoboda and Ševčíková, 2022), performs both compound splitting and component lemmatization with decent results on Czech.

Specifying the syntactic structure and assigning syntactic relation labels is another important step for which existing sources provide only very limited data (cf. the bracketed structure in CELEX). Since the pilot manual annotation was based around a mostly mechanical process of finding compound-associated phrases, feeding them into UDPipe (Straka et al., 2016), and observing the relation within the phrase, a semi-automatic procedure is being developed that follows this approach. For example, the German compound *Zittergras* 'quaking-grass' encodes the phrase *das Gras zittert*. The syntactic annotation provided for this phrase by UDPipe is then replicated in the compound, cf. the structures of the compound and of the underlying phrase both with *Gras* as `nsubj` in (18). The English example *killjoy* with the `obj` relation follows in (19).

(18)



(19)



In addition to the examples provided in this section ((16) through (19)), the envisioned annotation scheme is applied to the examples that were presented above in Section 3 – see the Appendix, where the annotation according to the current UD guidelines is shown on the left-hand side and the proposed annotation on the right.

5 Concluding remarks

In this paper, we explored the current treatment of compounds in UD in five languages. We observed that the handling of open and hyphenated compounds varies widely according to the particular language in question, and that closed compounds are taken into account in none of them. Based on these observations and also the long-standing tradition of describing compounds from a syntactic perspective present in the linguistic literature, the objective of the paper was to open a discussion on whether a multilingual annotation scheme for compounds in UD that employs the dependency relations already in use is useful and what features it should have.

The proposed scheme is currently being implemented in the data of the languages under study, and the aim is to extend it to other languages, which will inevitably result in modifications to individual aspects of the scheme.

Acknowledgments

The study was supported by the Charles University, project GA UK No. 128122, and by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

RH Baayen, R Piepenbrock, and L Gulikers. 2014. CELEX2 LDC96L14, 1995. *URL https://doi. org/10*, 35111.

Charles Bally. 1944. *Linguistique générale et linguistique francaise*. A. Francke.

Laurie Bauer. 2001. Compounding. In M. Haspelmath, editor, *Language Typology and Language Universals: An International Handbook*, pages 695–707. De Gruyter.

Laurie Bauer. 2005. The Borderline between Derivation and Compounding. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer, and Franz Rainer, editors, *Morphology and its Demarcations*, pages 97–108. John Benjamins.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford University Press, Oxford.

Antonietta Bisetto and Sergio Scalise. 2005. The classification of compounds. *Lingue and Linguaggio*, 4(2):319–332.

Leonard Bloomfield. 1933. *Language*. A. Francke.

Geert Booij. 2005. Compounding and Derivation: Evidence for Construction Morphology. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 264:109–132.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Nigel Fabb. 1998. Compounding. In A. Spencer and A. M. Zwicky, editors, *Handbook of Morphology*, pages 66–83. Blackwell.

Emiliano Guevara, Sergio Scalise, Antonietta Bisetto, and Chiara Melloni. 2006. MORBO/COMP: A Multilingual Database of Compound Words. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation*, pages 2160–2163.

Martin Haspelmath. 2002. *Understanding Morphology*. Arnold.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011*, pages 420–426.

Rochelle Lieber. 2010. *Introducing Morphology*. Cambridge University Press, Cambridge.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*, pages 185–189.

Hans Marchand. 1969. *The Categories and Types of Present Day English Word Formation*. Beck'sche Verlagsbuchhandlung.

Susan Olsen. 2001. Copulative Compounds: A Closer Look at the Interface Between Syntax and Morphology. In *Yearbook of Morphology 2000*, pages 279–320. Springer.

Barbara Schlücker, editor. 2019. *Complex Lexical Units. Compounds and Multi-Word Expressions*. De Gruyter, Berlin.

Andrew Spencer, editor. 1991. *Morphological Theory*. Blackwell, Oxford.

Pavol Štekauer, Salvador Valera, and Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: A typological survey*. Cambridge University Press.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Emil Svoboda and Magda Ševčíková. 2022. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *Prague Bulletin of Mathematical Linguistics*, 118:55–73.

Pius ten Hacken, editor. 1994. *Defining Morphology. A Principled Approach to Determining the Boundaries of Compounding, Derivation and Inflection*. Olms, Hildesheim.

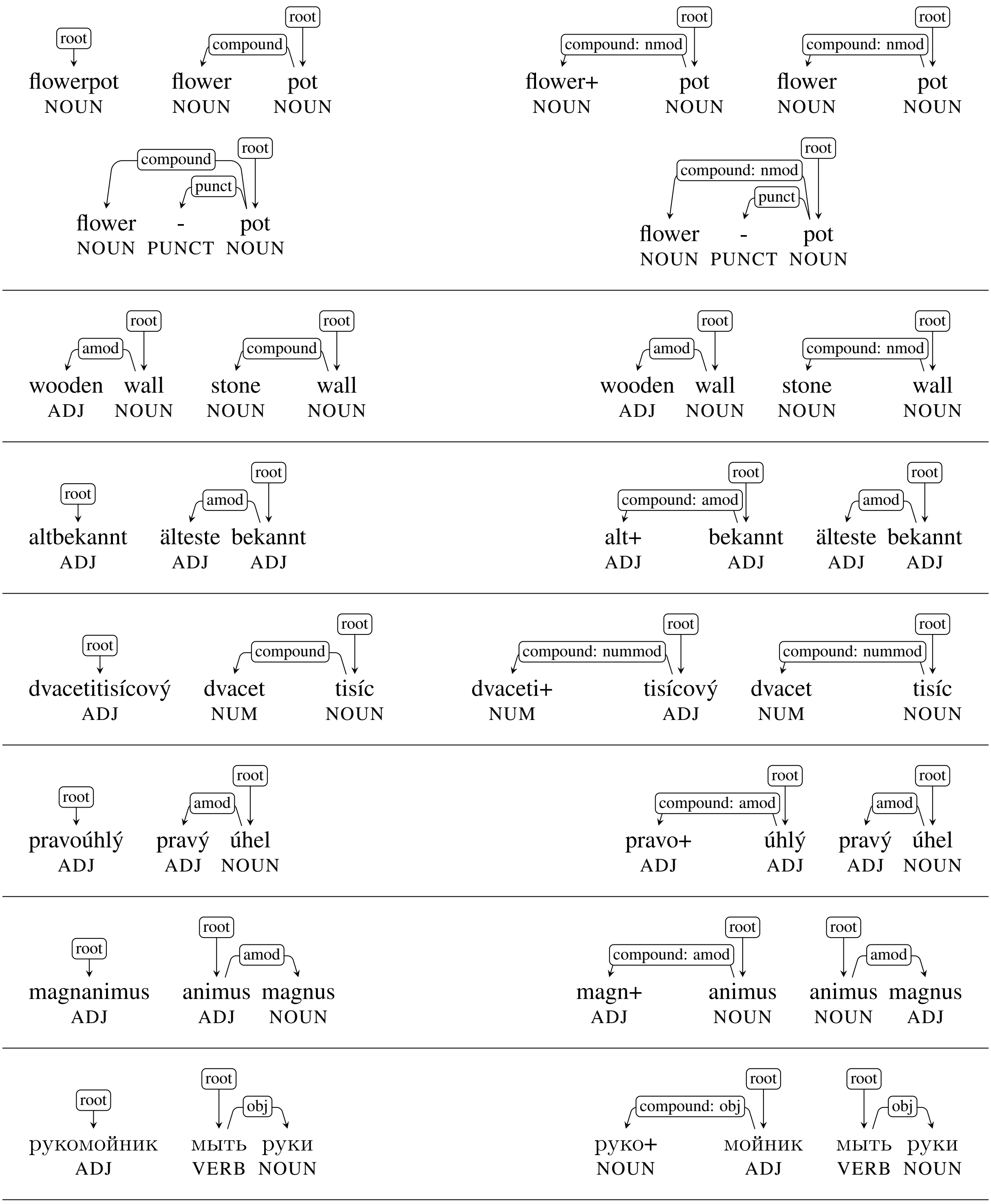
Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. 2021. DeriNet 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniil Vodolazsky and Hermann Petrov. 2021. Compound Splitting and Analysis for Russian. *Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 145–153.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Francisco Costa, Marine Courtin, Mihaela Cristescu, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóğa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio

Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinh͡ór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn H͡órðarson, Vilhjálmur H͡orsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. Universal dependencies 2.12. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix: Compounds annotated according to the current Universal Dependencies guidelines (left) vs. in line with the proposed annotation scheme (right)


root
flowerpot NOUN
compound
root
flower NOUN
pot NOUN
compound: nmod
root
flower+ NOUN
pot NOUN
compound: nmod
root
flower NOUN
pot NOUN
compound
punct
root
flower - pot
NOUN PUNCT NOUN
compound: nmod
punct
root
flower - pot
NOUN PUNCT NOUN
amod
root
wooden ADJ
wall NOUN
compound
root
stone NOUN
wall NOUN
amod
root
wooden ADJ
wall NOUN
compound: nmod
root
stone NOUN
wall NOUN
root
altbekannt ADJ
amod
root
älteste ADJ
bekannt ADJ
compound: amod
root
alt+ ADJ
bekannt ADJ
amod
root
älteste ADJ
bekannt ADJ
root
dvacetitisícový ADJ
compound
root
dvacet NUM
tisíc NOUN
compound: nummod
root
dvaceti+ NUM
tisícový ADJ
compound: nummod
root
dvacet NUM
tisíc NOUN
root
pravoúhlý ADJ
amod
root
pravý ADJ
úhel NOUN
compound: amod
root
pravo+ ADJ
úhlý ADJ
amod
root
pravý ADJ
úhel NOUN
root
magnanimus ADJ
root
amod
animus ADJ
magnus NOUN
compound: amod
root
magn+ ADJ
animus NOUN
root
amod
animus NOUN
magnus ADJ
root
рукомойник ADJ
root
obj
мыть VERB
руки NOUN
compound: obj
root
руко+ NOUN
мойник ADJ
root
obj
мыть VERB
руки NOUN