Joint Annotation of Morphology and Syntax in Dependency Treebanks

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, Yixuan Li

Joint Annotation of Morphology and Syntax in Dependency Treebanks

Bruno Guillaume[1], **Kim Gerdes**[2], **Kirian Guiller**[3], **Sylvain Kahane**[3], **Yixuan Li**[4]
[1]LORIA, INRIA, Université de Lorraine
[2]LISN, Université Paris-Saclay, Orsay, France
[3]Modyco, Université Paris Nanterre & CNRS
[4]LPP (CNRS) et Sorbonne Nouvelle
Bruno.Guillaume@inria.fr, gerdes@lisn.fr, guiller.kirian@parisnanterre.fr, sylvain@kahane.fr, yixuan.li@sorbonne-nouvelle.fr

Abstract

In this paper, we compare different ways to annotate both syntactic and morphological relations in a dependency treebank. We propose new formats we call mSUD and mUD, compatible with the Universal Dependencies (UD) schema for syntactic treebanks. We emphasize on mSUD rather than mUD, the former being based on distributional criteria for the choice of the head of any combination, which allows us to clearly encode the internal structure of a word, that is, the derivational path. We investigate different problems posed by a morph-based annotation, concerning tokenization, choice of the head of a morph combination, relations between morphs, additional features needed, such as the token type differentiating roots and derivational and inflectional affixes. We show how our annotation schema can be applied to different languages from polysynthetic languages such as Yupik to isolating languages such as Chinese.

Keywords: Morph, Morpheme, Morph-based treebank, Derivational affix, Derivational path, Compound, Word structure, Universal Dependencies

1. Introduction

Syntactic treebanks have been in development since the 1970s. Originally, they were created primarily for languages with a written tradition where segmentation into orthographic words[1] was intuitive. These initial treebanks mostly focused on Indo-European languages, which generally exhibit limited inflectional morphology. The pioneering treebanks in this field were the Talbanken for Swedish (Einarsson, 1976), the Penn Treebank for English (Marcus et al., 1993), and the Prague Dependency Treebank for Czech (Hajič, 1998). Teams specializing in Natural Language Processing (NLP) developed these resources with the objective of training various NLP tools and parsers from the early 2000s (Klein and Manning, 2003; Nivre et al., 2006) used the orthographic form as their primary reference.

However, the landscape has shifted dramatically in the last decade. Treebanks are now available for a vast array of languages from across the globe, exemplified by the Universal Dependencies initiative which boasts treebanks for over 150 languages[2]. These modern treebanks encompass languages with intricate morphology, such as Japanese (Tanaka et al., 2016), Turkish (Çöltekin, 2016), and Yupik (Park et al., 2021). Furthermore, there are now treebanks dedicated to languages with only an oral tradition, such as Beja (Kahane et al., 2021) or Mbyá Guaraní (Thomas, 2019). For these languages, there is often no conventional understanding of what constitutes a 'word'. Linguists studying and transcribing these languages tend to segment text at the morphemic level, in line with the Interlinear Glossed Texts (IGT) tradition. In some cases, decisions about certain morphs – specifically whether they should be classified as affixes or standalone words – are deferred until a clearer understanding is reached.[3] Additionally, languages with written traditions but employing a *scriptio continua* (a writing system without spaces), such as Chinese (Li, 2023) or Japanese (Tanaka et al., 2016), present challenges in defining word boundaries.

The current standard for syntactic treebanks is the Universal Dependencies (UD) annotation schema (De Marneffe et al., 2021), where word-level annotation is required. But some UD treebanks have been first developed at the morph-based level, before being converted to UD, or contains some morph-level nodes (Section 2).

In this study, we introduce an adaptation of UD, termed mUD, which emphasizes morph-based annotation. While UD predominantly underscores the dependency between content words and positions function words as leaves of the dependency tree, mUD emphasizes the relationship between roots and their respective affixes. In this context, we de-

[1]Here, by *orthographic words*, we refer to segments of text separated by spaces.

[2]https://universaldependencies.org

[3]To understand the concept of a morph, refer to Mel'čuk (2006); Haspelmath (2020). A morph is a linguistic sign the signifier of which cannot be decomposed. A morpheme is a collection of (allo)morphs.

fine *roots* as the central lexical part of a word to which components like affixes attach.[4]

However, transitioning from mUD to UD poses challenges, as explained in (Kahane et al., 2021), where a morph-based treebank for Beja has been converted into a word-based treebank. To address these difficulties, we introduce mSUD, a morph-based version of SUD (Surface-syntactic UD). In mSUD, the designation of heads is steered by distributional criteria, frequently giving precedence to affixes over roots. We further elaborate on the process of converting mSUD to both SUD and UD within this paper.

We do not advocate for the universal adoption of morphological level annotation across all languages or treebanks. For certain languages, such as polysynthetic or agglutinative languages and those without a written tradition, a morphological approach is almost indispensable. For others, such as Chinese or Japanese, we demonstrate that morphological annotation offers an efficient strategy to navigate the complexities of word-level tokenization where clear delineations are not always apparent.

While the principles of mSUD, detailed later in this paper, could technically be applied to languages like English (and we demonstrate this with select examples), the benefits might not justify the effort required. Our central proposition is that each language—or even each individual treebank—has the flexibility to embrace an mSUD-style analysis, while still maintaining compatibility with prevalent word-level frameworks like UD or SUD through an automatic conversion process.

2. Related works

The question of annotation at some subword level has been discussed in many previous studies (see Gross 2010 for a first attempt). We focus here on the main discussions on this topic in the context of the Universal Dependencies project and we mention several treebanks implementing some subword analysis.

Yupik, a polysynthetic language prevalent in Alaska and the Chukotka region of Russia, is characterized by its intricate morphology. In some instances, words can encompass up to seven derivational morphemes. An illustrative sentence from the UD_Yupik-SLI treebank demonstrates this linguistic phenomenon. In sentence (1), the entirety of its information, barring the concluding period, is encapsulated in just one token, when interpreted using the UD guidelines.

(1) *Mangteghaghllangllaghyugtukut.*
house-big-to.make-to.want.to-IND.INTR-1PL
'We want to make a big house.'

Yet, the prevailing UD annotation by (Park et al., 2021) adopts a semblance of the mUD analytical approach. The elongated orthographic word is segmented into six distinct subunits (refer to Figure 1). This syntactic dissection employs traditional UD relations such as `nmod`, `obj`, and `xcomp`. Furthermore, it introduces the specific relation `dep:infl` designated for inflectional suffixes.

An alternate version of the treebank, automatically produced, adhering strictly to UD guidelines, can be accessed in the `not-to-release` directory of the associated GitHub repository.

In (Kahane et al., 2021), a treebank for Beja is presented. The Beja language does not have a writing tradition and the treebank is built from already existing IGT. The annotation at the morphological level is then more natural. The corpus is annotated in the SUD framework with a morph-level tokenisation and, like for the Yupik language, an automatic conversion is used to produce a version which follows UD requirements.

In a previous study (Li et al., 2019), an attempt was made to enhance four Chinese UD treebanks with morphological information through manual annotation and rule-based methods. The parser trained on these character-level treebanks demonstrated state-of-the-art performance. Subsequently, in a more recent work (Li, 2023), a character-level Chinese patent treebank, manually annotated and consisting of 100 sentences with five types of inter-character relations, was introduced. Additionally, this treebank was converted into a conventional UD format at the word-level.

A similar problem arises for Japanese that has no obvious word boundaries as it uses a *scriptio continua*, without whitespace, just like Chinese. The Japanese UD project struggles to apply the general UD annotation guide to their language and at the same time attempts to foster different demands towards the annotation standard by considering three levels of word segmentation: Short Unit Word (SUW), Long Unit Word (LUW), and *bunsetsu*. "SUW is a minimal language unit that has a morphological function. SUW almost always corresponds to an entry in traditional Japanese dictionaries." (Tanaka et al., 2016). SUWs can be detected by parsers based on morphological dictionaries. Combining compound nouns and light verb constructions into a single token gives LUW, whereas case markers and inflectional affixes remain separate tokens on this level. On the other end of the segmentation options is the *bunsetsu*, a unit that includes all of its clitics and affixes.

The Japanese tokenization within the UD project is notable for its exceptional approach, particularly

[4] It is crucial to clarify that by *root*, we are referring to the core segment of a word. This definition is distinct from the root that denotes the head of a sentence.

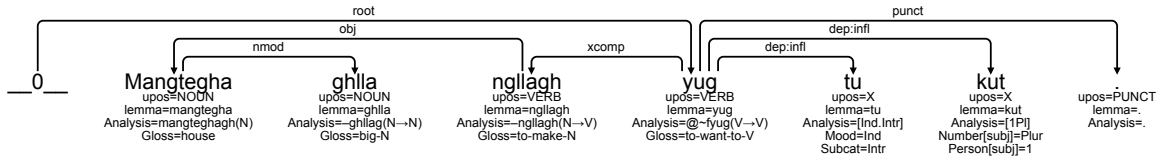

Figure 1: The UD analysis of sentence (1)

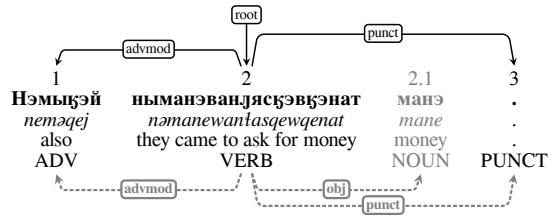

Figure 2: Annotation example for Chukchi (taken from Tyers and Mishchenkova)

when compared to the analyses of other agglutinative languages. It utilizes morphs, without morphological features, as its fundamental units, thereby contravening UD's Tokenization and Word Segmentation guidelines, which state: "morphological features are encoded as properties of words and there is no attempt at segmenting words into morphemes." A recent proposal by Taguchi and Chiang (2023) advocates reassessing how Japanese treebanks' morph combinations align with UD's tokenization standards. The proposal introduces two new levels of morph combination to accommodate the distinctive nature of Japanese verbal inflection, which is more fusional, in contrast to its less synthetic case marking.

These challenges highlight that the determination of word boundaries operates independently and is orthogonal to the process of dependency annotation.

For some languages, previous studies have proposed to have a partial annotation of the morphological level. This is the case in Chukchi and in Turkish.

For Chukchi, a polysynthetic language, Tyers and Mishchenkova (2020) are specifically interested into the annotation of noun incorporation. When a noun is incorporated in a complex morphological compound in which it plays the role of an object of a verb of the same compound, making the object relation explicit is important to account for the semantics of the construction. Hence, they propose an encoding (see Figure 2) where both the full compound (token 2) and the extracted noun (token 2.1) are represented (but not the other parts of the compound). For this, they misuse enhanced dependencies to encode a two-layer annotation in a context that does not correspond to the intended use of enhanced dependencies.

For the Turkish language, Çöltekin (2016) explains that in some examples, it is difficult to avoid annotation at a subword level. They give example (2), in which the second token, *arabadakiler*, stands for two entities carrying different inflections. Following current UD word conventions, it is impossible to have a sensible annotation because the ADJ *Mavi* refers only to the sub-word *arabada* whereas the subject of the verb *uyu* is the subword *kiler*. Çöltekin proposes to have a partial annotation at the subword level: a word is split into smaller pieces (named *inflectional groups* in Turkish literature) only if "(a) Parts of the word may have potentially conflicting inflectional features" or "(b) Parts of the word may participate in different syntactic relations." If none of these conditions are met, the annotation is kept at the word level.

(2) *Mavi* *arabadakiler* *uyuyorlar*
Blue car.LOC-ki.PL sleep.PROG.1P
'The ones in the blue car are sleeping.'

3. Annotation at the morph level

3.1. Morph level

We have seen in the previous sections that there are several motivations for annotating structure below the word level. The different papers mentioned above propose several ways of encoding syntactic relations within a word. We propose here a common way to unify these different proposals.

Our proposal is then to provide the annotation at the morphological level and, from this first annotation, to automatically generate the syntactic word level annotation expected in the UD or SUD framework. We call this annotation mUD or mSUD to explicitly place it at a different level.

Concerning SUD, the main differences in mSUD (note that the differences between UD and mUD are similar) are as follows:

- In mSUD, there are two types of dependency relations: relations between syntactic words (noted as in regular SUD) and relations at the

morphological level (within syntactic words), which are written with the suffix `/m` as in `subj/m`.

- Token may contain the feature `TokenType` with one of the values `DerAff`, `InflAff`, `Root`, `Word`, or `Punct`, as well as `Break` for spoken corpora using special symbols for prosodic breaks.
- A derivational affix may have a feature `DerPos` to indicate the final `upos` on the corresponding word level entity, while a compound can have a feature `CpdPos`, when the compound has a `upos` different from its head (as *do-it-yourself* in *the do-it-yourself department*.

Subword analysis predominantly falls into three key categories: Derivation, Composition, and Inflection. In the ensuing sections, we delineate the specific mSUD annotations associated with each category, supplemented with examples from various languages. Additionally, we will delve into instances where categorizing a particular case becomes difficult. But first, let's briefly discuss the topic of tokenization.

3.2. Tokenisation: allomorphy and portmanteau

An important rule of UD dictates that the text should be the concatenation of its tokens. In instances where a token is not succeeded by a space, it is marked with the feature `SpaceAfter=No`. In the mSUD (or mUD) framework, a word like *baby-sitter* is thus dissected into four distinct tokens: *baby*, -, *sitt*, and *-er* (see Figure 5).[5] We annotate allomorphy in the lemma column. The lemma gives the canonical form of the root or the affix. For the previous example, the lemma of *sitt* is the verb lemma *sit*, and the lemma of the suffix is *-er*.

For cases where the surface form cannot be decomposed, it is still possible to analyze the morpheme and to use the UD mechanism for portmanteaus. The English word *sung* could then be split into two "abstract" tokens of lemma *sing* and *-en*. Note that the form of these tokens is not relevant and could be empty.

[5]Note that, depending on the language, we add or not the 'dash' symbol to make suffixes explicit. We add it for languages using spaces between words or for which annotation starts from IGT which traditionally use this convention. In the other hand, we do not add the 'dash' symbol for Chinese or for Japanese.

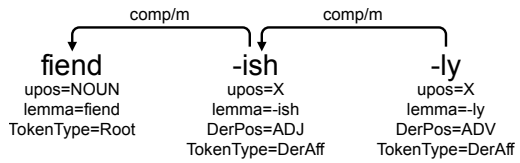

Figure 3: mSUD analysis of the English adverb *fiendishly*

4. Derivational affixes

4.1. Structure for derivation

SUD is based on distributional criteria to choose the head of a phrase (Gerdes et al., 2018). Such criteria have been formalized by Bloomfield (1933) and used in phrase-structure grammar (Jackendoff, 1977) and dependency grammar (Mel'čuk, 1988; Hudson, 1984). Roughly speaking, the head of a phrase is the element that controls the distribution of the phrase. Applied to morphology, it identifies derivational affixes as heads because it is the affix that decides what is the part of speech of the combination between a root and an affix.

Figure 3 presents the mSUD annotation of the English adverb *fiendishly*. At its root is the noun *fiend*. The affix *-ish*, signifying a derivation to an adjective, appends to it (`DerPos=ADJ`). This is further derived into an adverb with the addition of the affix *-ly* (`DerPos=ADV`). Each affix targets a segment of the word, marking it as its complement through the `comp/m` annotation.

Note that our analysis gives us the internal structure of the word. In figure 9, with the mSUD analysis (at the top), we see that the root combines first with the suffix *able* and then with the prefix *un*. Such a prefix, could not combine with a verbal root.

For derivational affixes that maintain the part of speech, the analysis is less clear. It can be more appropriate to categorize them as modifiers when the derivative shares the same semantic category as the root (an example is the prefix *sub* in *subword*, where the derivative shares the same semantic category as the root). However, in many instances, even if the POS remains unchanged by the derivational affix, the word's semantic classification does shift. Take for instance *grammarian*, which signifies a person, while *grammar* refers to a conceptual entity. Furthermore, positioning the affix *ian* as the central component allows differentiation between terms like *fiendish grammarian*—where the adjective influences the derivative—and *generative grammarian*—where the adjective impacts the root. The latter structure is notably less common across

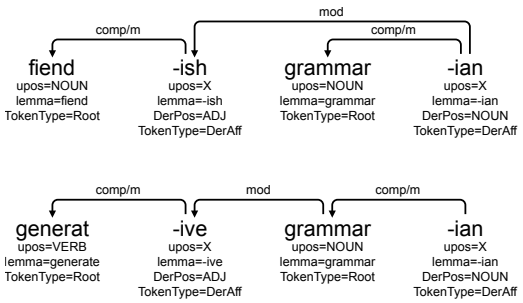

Figure 4: A derivational affix treated as the head

	DerPos			
root	VERB	NOUN	ADJ	ADV
VERB	AUX	SCONJ	SCONJ	SCONJ
NOUN	AUX	X	ADP	ADP
ADJ	AUX	X	X	X
ADV	AUX	X	X	X

Table 1: `upos` of derivational afffixes according to the `upos` of the root and the derivative

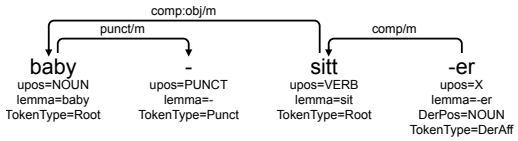

Figure 5: Compound annotation in mSUD

global languages, as highlighted by Gross (2010) (see Figure 4).

4.2. POS for derivation

As already mentioned, each derivational affix bears a `DerPos` feature indicating the `upos` of the stem resulting form the derivation. This feature characterizes the affix and is different from the `ExtPos` (for external POS) feature already used in some UD or SUD treebanks to indicate the POS of a unit when it is different from what can be expected from its parts. The `ExtPos` feature is mainly used for fixed expressions or for titles.

The sentence *The Shining is a 1980 horror film* is an example where the two features `DerPos` and `ExtPos` appear on the same morph *ing* with different values: The feature `DerPos=NOUN` expresses that the syntactic word *Shining* is a `NOUN`, whereas the feature `ExtPos=PROPN` expresses that the phrase *The Shining* is a film title, which is considered as a `PROPN` in the rest of the sentence.

In the UD framework, every token is expected to possess a `upos`. Several strategies can be adopted to accommodate this requirement. One approach could involve creating specific POS tags exclusively for affixes. However, as a current measure, we recommend adhering to the UD conventions and leveraging the POS tags designated for full words. One way to achieve this is to use the `DerPos` as the `upos`. This method has already been implemented for noun-to-noun and verb-to-verb derivational affixes in the Yupik treebank, as outlined by Park et al. (2021). Alternatively, setting `upos=X` for all affixes is also feasible. In the Beja morph-based treebank, Kahane et al. (2021) adopted a different `upos` from `X` when a full word with a similar syntactic function to the morpheme exists. For example, a nominalizer paired with a verb root is labeled as `SCONJ`, since a verb combined with a subordinating conjunction exhibits noun phrase-like behavior. See Table 1 for a proposition.

5. Composition

Compounds are words composed by the combination of two or more roots. In most cases, it is rather easy to decide which one is the head. This is a case for a word such as *baby-sitter*, where *baby* is an incorporated object. The fact that the derivational affix *-er* is deverbal confirms that *sit* is the head of *baby-sit* (Figure 5).

In our following analysis of compounds categories, we focus particularly, but not only, on German and Mandarin Chinese, two languages renowned for their extensive use of compounding strategies in word formation.

It can happen that the compound word has a different distribution than the head of the roots. In this case, we assign the feature `CpdPos` to the head of the compound. Like for the newly introduced `DerPos`, the `CpdPos` is not redundant with the existing `ExtPos` (see 4.2 for a more detail explanations).

Different cases of composition can be differentiated by the use of different relations. We propose to use `conj/m` when a word is composed of two roots from the same syntactic and semantic class and the meanings of the individual roots closely align with each other and with the meaning of the composite word. Examples:

- **English** *wolfhound*;
- **Mandarin** 语言 (yǔ yán) 'language', lit. *speech language* (see Figure 6).

The most frequent type of composition is the modifier-head relation, that we annotate with `mod/m`. Usually, the head is a noun, and the modifier is another noun, an adjective, or an adverb. Some examples:

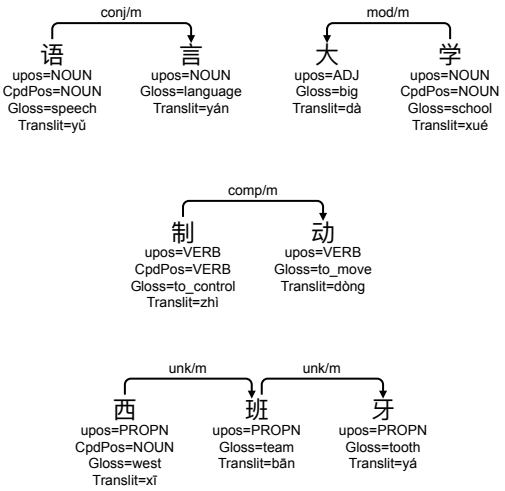

Figure 6: Some mSUD annotation in Chinese

- **English** ADJ-NOUN *blackboard*;
- **Mandarin** ADJ-NOUN 大学 (dà xué) 'university', lit. *big school* (see Figure 6).
- **German** ADJ-NOUN *Hochschule* 'university', lit. *high school*;[6]
- **Finnish** NOUN-NOUN: *kouluruokailu* 'school lunch', lit *school meal*;

We also propose a `comp/m` for predicate-complement relations:

- **Mandarin** VERB-VERB 制动 (zhì dòng) 'brake', lit. *(to) control (to) move* (see Figure 6).
- **German** NOUN-VERB *Autofahren* 'driving (a car)', lit. *car driving*.[7]

Mandarin has also words that are built by combinations of characters without clear links between them, especially for the phonetic transcription of loanwords. We us `unk/m` in this case (for *unknown* relation): 西班牙 (xībānyá) 'Spain', lit. *west team tooth* (see Figure 6).

[6]Notably, the English equivalent *high school* is structurally similar but is represented as two separate words. This highlights the necessity of morpheme-level annotation for linguistic consistency across languages.

[7]Note that the German spelling reform of 1996 decided to separate 'Auto' and 'fahren' when the combination remains a verb and spell the nominalization as one word. This can be interpreted as an additional argument to annotate on the linguistically defined morphemes rather than on words.

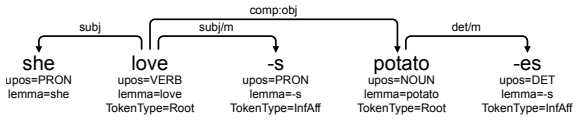

Figure 7: Example of mSUD annotation in English for inflectional affixes.

6. Inflection

In line with the case of derivational affixes, we propose that inflectional affixes govern the stem when they control the distribution of the word. This is the case of TAME affixes, i.e., affixes that indicate time, aspect, mood, or evidentiality, since a finite verb has a different distribution from an infinite verb or a participle, which can occupy positions dedicated to nouns (*I want to read/a book*) or adjectives (*the book bought by her*). We propose to use `upos=AUX` for TAME affixes, `SCONJ` for infinitive, participles and gerunds. In this way, the English word *complicated* has three analyses as an adjective, a past participle, and as past tense verb:

	complicat	*-ed*	
Analysis	upos	upos	lemma
Adjective	VERB	SCONJ, DerPos=ADJ	-ed
Participle	VERB	SCONJ	-en
Past	VERB	AUX	-ed

Note that even if the inflectional affix is the head of the word, the `upos` of the word is the `upos` of the stem (the root and the derivational affixes), as usual in UD, and we do not need a `DerPos`-like feature on inflectional affixes.

Case markers, which control the syntactic position of the noun phrase, are also treated as heads. We tag them `upos=ADP` and they take the noun as a complement.

On the other side, we consider that inflectional affixes for agreement do not change the distribution of the word and can be treated as dependents. Pronominal affixes are tagged `upos=PRON` and receives a `subj/m` or `comp:obj/m` relation depending on whether they mark agreement with the subject or the object. Number affixes on nouns are `DET` (see Figure 7).

7. Implementation

In practice, it is complicated to maintain multiple versions of the same treebank that follow different tokenization bases or different annotation principles. We then propose to use two kinds of automatic conversions (Figure 8).

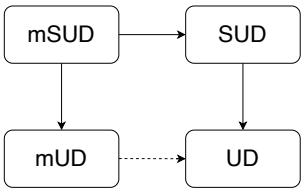

Figure 8: Conversions between formats

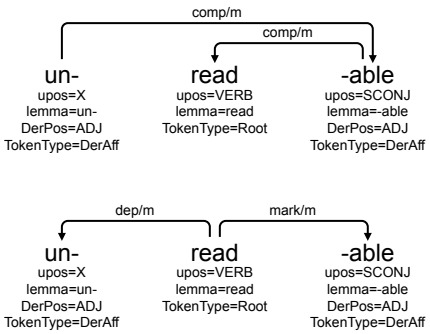

Figure 9: mSUD and mUD encoding of derivational paths

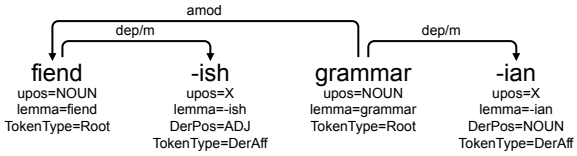

Figure 10: mUD for 'fiendlish grammarian'

First, the morph-based annotation is more detailed than the word-level annotation and it is possible to automatically construct the word-level from the morph-level (horizontal arrows in the figure). The word tokenisation is given by the `/m` annotation on dependencies: if two morphs are linked by such a relation, then they belong to the same word and they should be merged. All necessary information, such as the POS of the result of the merge, is available (with the features `DerPos` or `CpdPos`).

Note that the conversion from mUD to UD (dashed arrow) is less straightforward. As derivational paths are not encoded in mUD, it is not always possible to safely produce the final POS: if two or more derivational affixes are attached to the same root, the order in which the affixes are merged changes the output.

The second kind of conversion (vertical arrows in Figure 8) is between the SUD and UD frameworks. Starting with the SUD to UD conversion proposed in (Gerdes et al., 2018), it can be easily adapted to take into account morph-level dependencies to produce a conversion from mSUD to mUD.

As we said in the introduction, any language, or even any treebank can be annotated primarily at any level, and the UD version can be recovered by conversion. If annotation below word-level is required or desired, we advocate for the mSUD format, which is the richest one, and from which others can be produced.

8. Discussion

8.1. mUD

It is possible to design in a similar way an annotation based on UD principles at the morphological level. Following the UD principle which consists of choosing the semantic words as heads and functional words as dependants; in mUD, the root would be the head of the structure, each affix depending on the root. The main drawback of this annotation choice is that derivational paths are not completely encoded and the structure then contains less information than mSUD where affixes are head. When two derivational affixes are attached to the same root, the order in which they are applied is not encoded (see Figure 9).

UD relations between words are attached to the head of the morphological annotation, i.e. between roots. Hence, it is not possible to make the distinction shown in Figure 4 and the two examples are annotated in the same way (Figure 10).

8.2. Applications of mSUD to other Treebanks

When applying the principles of mSUD to various treebanks, certain intricacies and challenges emerge.

- **Yupik:** Our mSUD analysis of the one-word sentence (1) from Yupik is proposed in Figure 11. As recalled by Park et al. (2021), almost all Yupik words are constituted of a root followed by derivational suffixes and completed by inflectional suffixes. It follows from our conventions that the last inflectional TAME suffix is the head of the morphological structure.

  In the initial analysis (Figure 1), the derivational path was not explicited. It was not possible, if you did not know the general rules of Yupik morphology, to find the root and to understand that the combination between the two first morphs, both annotated as `NOUN`s, was a derivation. This information, which

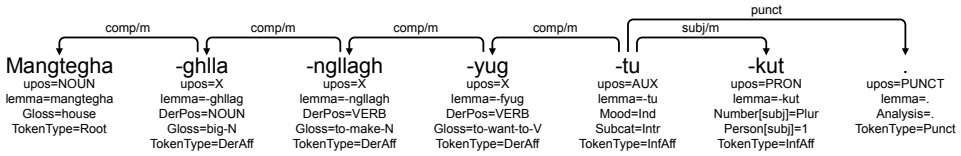

Figure 11: The mSUD analysis of sentence (1)

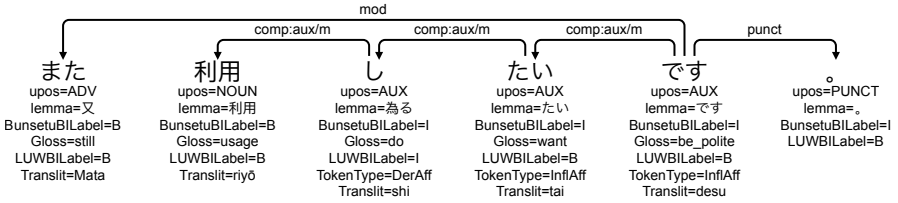

Figure 12: mSUD example in Japanese

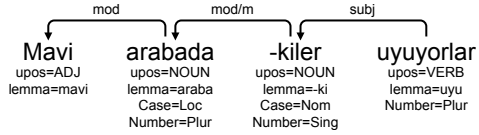

Figure 13: The (partial) mSUD analysis of sentence (2)

was present in the `Analysis` feature, is dispatched in our annotation in `lemma`, `DerPos` and `TokenType`. In other words, a mSUD annotation could be automatically inferred from the UD_Yupik-SLI treebank using the `Analysis` feature but not the current annotation.

- **Japanese:** Existing Japanese treebanks with morpheme-based tokenization can be conveniently transformed into mSUD and mUD formats. To achieve this, two primary decisions need to be made:

  1. Determine the level at which the /m relation type is introduced—whether at larger units like *bunsetsus* or restricted to within Long Word Units.

  2. Decide on the appropriate `TokenType` feature, basing the decision on the POS of individual morphemes.

  As an illustration of a conversion at the *bunsetsu* level, consider the mSUD representation of sentence test-s516 from the Japanese GSD (see Figure 12). Here, we mark relations within the *bunsetsus* using the /m notation, excluding punctuation. Notably, the POS of 利用 (riyō, 'usage') is adjusted from VERB to NOUN. This modification highlights that し (shi, 'do') serves a derivational function, converting 'usage' into the action 'use'.

8.3. Partial annotation at the morphological level

A partial mSUD could be beneficial for Turkish. The decision of how to implement this requires discussion. Our proposed method aligns with the analysis presented in (Çöltekin, 2016). Here, only a segment of the subword level undergoes analysis, making it a partial exploration. To illustrate, for the sentence (2), the analysis suggested in the referenced paper, when adapted to mSUD (Figure 13), would provide a detailed breakdown of the morphological structure.

In contrast, our proposal can not be easily adapted to partial annotation where only a strict subpart of the morphological content of words are targeted. Either we split a word into smaller units or it is kept as atomic, but it is not possible to have both a compound and one of its strict subpart as in Figure 2.

8.4. Comparison with another proposal

In (Zeman, 2023), another way of encoding for the subword level is proposed. The idea is to have in one structure both the word level and the morph level. An example in given with the German noun *Hauptrolle* which is a compound made of two words (*haupt* and *Rolle*) and the structure proposed contains three different units for the compound and its two parts. While containing all the information we can expect, this structure is difficult to annotate and to maintain. Nevertheless, if a structure contain-

ing both layers is needed, it can be automatically produced from our proposed format.

9. Conclusion

In this paper, we propose some guidelines for syntactic annotations which extends the UD or SUD current guidelines for annotating below the syntactic word level (named mUD and mSUD). We consider the distinction of three mechanisms at play (derivation, composition and inflection) and we give SUD-style criteria for deciding the internal mSUD structure of morphs in words and to encode the derivational path. We exemplify on several languages theses principles. Several treebanks are currently developed in the mSUD format. In version 2.14 (expected in May 2024) of SUD, three trebanks will be released in mSUD: mSUD_Chinese-PatentChar, mSUD_Chinese-Beginner, mSUD_Chinese-PatentChar. The mSUD format is also used in the ongoing develoopement of other treebanks for low resources languages.

We hope that this proposal would help the development on other new treebanks, mainly for low resources languages for which IGT are available.[8] We also think that this will help for better comparison of constructions across languages in the many cases where some construction is expressed at the syntactic level in some languages and at a morphological level in some others.

10. Acknowledgements

We would like to thank the three anonymous reviewers for their valuable comments. This research was supported by the French National Research Project (ANR) Autogramm (Projet-ANR-21-CE38-0017).

References

Leonard Bloomfield. 1933. *Language*. Henry Holt, New York.

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

Jan Einarsson. 1976. Talbankens skriftspråkskonkordans.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud. In *Proceedings of the second Universal Dependencies Workshop (UDW)*. Association for Computational Linguistics (ACL).

Thomas Gross. 2010. Chains in syntax and morphology. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 143–152.

Jan Hajič. 1998. Building a syntactically annotated corpus: The prague dependency treebank. In Eva Hajičová, editor, *Issues of valency and meaning: Studies in honour of Jarmila Panevová*, pages 106–132. Karolinum.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134.

Richard Hudson. 1984. *Word Grammar*. Blackwell, Oxford.

Ray Jackendoff. 1977. $\overline{X}$ *syntax: A study of phrase structure*. MIT Press, Cambridge.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics (ACL)*, pages 423–430.

Yixuan Li. 2023. Character-level dependency annotation of Chinese. In *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, pages 42–53, Washington, D.C. Association for Computational Linguistics.

Yixuan Li, Gerdes Kim, and Dong Chuanming. 2019. Character-level annotation for Chinese surface-syntactic Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 216–226, Paris, France. Association for Computational Linguistics.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. 1993. *Computational linguistics*, 19.

Igor Mel'čuk. 1988. *Dependency syntax: Theory and practice*. State University of New York Press, Albany.

[8] Another challenge for treekanks built from IGTs is to perform cumulative work and not lose the information from the IGT, in particular the tokenization into morphs and the morphosyntactic annotation at the morph level.

Igor Mel'čuk. 2006. *Aspects of the Theory of Morphology*. Mouton de Gruyter.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the fifth Conference on Language Ressource and Evaluation (LREC)*, volume 6, pages 2216–2219.

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

Chihiro Taguchi and David Chiang. 2023. Introducing morphology in Universal Dependencies Japanese. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 65–72, Washington, D.C. Association for Computational Linguistics.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal dependencies for japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (lrec'16)*, pages 1651–1658.

Guillaume Thomas. 2019. Universal dependencies for mbyá guaraní. In *Proceedings of the third workshop on Universal Dependencies (UDW), Syntaxfest*, pages 70–77.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

Daniel Zeman. 2023. Subword relations - superword features. UniDive 1st general meeting, available at https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:2023-saclay:abstracts:39_zeman_subword_relations_superword_features.pdf.

Çağrı Çöltekin. 2016. (when) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics.*