Universal Feature-based Morphological Trees

Federica Gamba, Abishek Stephen, Zdeněk Žabokrtský
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{gamba, stephen, zabokrtsky}@ufal.mff.cuni.cz

Abstract

The paper proposes a novel data representation inspired by Universal Dependencies (UD) syntactic trees, which are extended to capture the internal morphological structure of word forms. As a result, morphological segmentation is incorporated within the UD representation of syntactic dependencies. To derive the proposed data structure we leverage existing annotation of UD treebanks as well as available resources for segmentation, and we select 10 languages to work with in the presented case study. Additionally, statistical analysis reveals a robust correlation between morphs and sets of morphological features of words. We thus align the morphs to the observed feature inventories capturing the morphological meaning of morphs. Through the beneficial exploitation of cross-lingual correspondence of morphs, the proposed syntactic representation based on morphological segmentation proves to enhance the comparability of sentence structures across languages.

Keywords: Morphs, Universal Segmentations, Universal Dependencies

1. Introduction

Universal Dependencies (UD) (de Marneffe et al., 2021) is a framework for consistent annotation of natural language data across languages. The UD project develops cross-linguistically consistent treebanks to facilitate multilingual and cross-lingual parsing research from a typological perspective.[1] However, the syntactic annotation proposed by UD, along with the standard tokenization often based on white-space,[2] poses some challenges to actual comparability across languages, as different languages may adopt different strategies to express the same phenomenon. Consider, for instance, the English sentence *I will go through a forest*, translatable in Czech as *Půjdu lesem*.

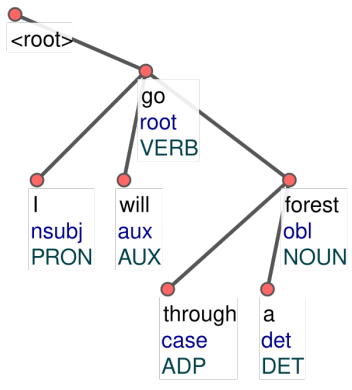


Figure 1: UD tree for the English sentence *I will go through a forest.*

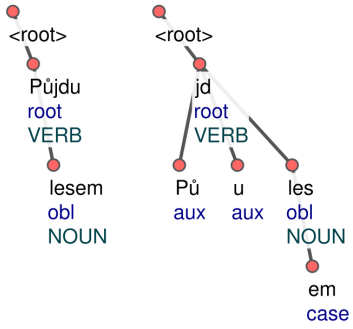


Figure 2: UD and morphological tree for the Czech sentence *Půjdu lesem. Pů* – a prefix expressing future tense, *jd* – the room morph for 'to go', *u* – a 1st pers. sg. conjugation ending, *les* – the root morph for 'forest', *em* – instr. sg. masc. declination ending.

These two equivalent sentences exhibit noticeable differences already in the token count, and their dissimilarity is reflected in their respective dependency tree structures. Nonetheless, a closer look at the sentences reveals that splitting word forms based on their morphological segmentation leads to a better mapping concerning isomorphy of trees and alignment of nodes,[3] allowing for greater comparability. Notably, in this example, Czech encodes future tense through the prefix *pů*, whereas the ending *em* for instrumental case in *lesem* expresses movement through (Figure 1, 2). Similarly, at the surface level the German compound *Finanzkrise* 'financial crisis' does not correspond –

[1] https://universaldependencies.org/.

[2] At least in the case of languages with the alphabetic writing system.

[3] At the word level, we observe a 3:1, 3:1 node alignment; at the morph level, node alignment is 1:1, 1:1, 1:1, 1:1, 1:0 (article unexpressed in Czech), 1:1.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 125–137. May 25, 2024.

in terms of structure and token count – to its Czech counterpart *finanční krize*. However, if we segment the two members that led to the formation of the compound (*Finanzen* + *Krise*), we obtain a clear correspondence of the German and Czech forms. A syntactic representation based on morphological segmentation could thus enhance the cross-lingual comparability of languages that e.g. exhibit different amounts of inflection or productivity in compounding.

Additionally, what emerges from the observation of segmented morphs[4] is that morphological features often tend to be associated to specific morphs. For instance, in the English word *letters* the morph *s* can be morphologically interpreted as an encoding for plurality. The morphological specification of a (syntactic) word form is encoded by a set of features in UD representing the lexical and grammatical properties. UD differentiates between lexical and inflectional features, where the former are an attribute of lemmas and the latter of word forms. This approach is convenient and productive in capturing the morphosyntactic functions of word forms, which fits the goal of UD, but it will not be incorrect to postulate that such lexical or grammatical functions can be encapsulated within morphs in a word form.

Thus, this study aims to propose a novel data representation, which exploits UD-like trees to represent simultaneously the UD-like syntactic sentence representation as well as the internal structure of word forms (hence taking the Item-and-Arrangement perspective on morphology (Bram, 2012)), which is merged within a single dependency tree. Using the inventory of universal morphological features in UD, we also investigate whether a strong correlation can be found between a given morph and a feature value, and then align the morphs to the observed feature that captures the lexical and grammatical functions of morphs. We thus propose a data structure that intertwines syntax and morphology with the goal of increasing comparability across languages.

The remainder of the paper is structured as follows. In Section 2 we present the related work, while Section 3 offers an overview of the resources that we employ for the present study. Section 4 details how such resources are exploited, focusing on the manipulation of treebank nodes and feature extraction, as well as discussing the strategy devised to comply with the UD schema. Section 5 shows the UD-like morphological trees that result from the present work, while Section 6 concludes the paper and outlines future research directions.

[4]Due to the ambiguous usage of the term 'morphemes', we use the term 'morphs' henceforth based on Haspelmath (2020).

2. Related Work

The idea of representing the internal structure of words has been previously explored, especially for non-alphabetic languages such as Chinese. In these kinds of languages, the issue of delimiting word boundaries is far from trivial and requires alternative strategies to be inspected. For instance, Zhao (2009) investigates internal character dependencies inside a word as a result of the attempt to handle word boundaries by identifying character-level dependencies.

Li (2011) elaborates on this approach by suggesting to recover word structures in morphological analysis. One of the reasons for this lies in the observation that there exist many different annotation standards for Chinese word segmentation, which could even cause inconsistency in the same corpus.[5] As we are working with alphabetical languages, their motivation for the work differs from ours. Additionally, we adopt dependency structures, while they work with constituency trees.

Concrete applications in the parsing of the approach in Li (2011) are described e.g. by Zhang et al. (2013), who annotate internal structures of words and then build a joint segmentation, part-of-speech (POS) tagging and phrase-structure parsing system. Zhang et al. (2014) integrate inter-word syntactic dependencies and intra-word dependencies, differentiating intra- and inter-word dependencies by the arc type to achieve results comparable to conventional resources.

In the case of languages with alphabetical writing systems, CELEX (Baayen et al., 1995) represents morphological word structure for Dutch, English, and German in the shape of a tree. Steiner (2017), e.g., exploits the resource in combination with GermaNet (Hamp and Feldweg, 1997). Morphological and compound information is extracted from the two resources respectively, and reused to build a so-called morphological treebank for German. However, such a morphological treebank consists of tree-shaped single tree-words only, without including any kind of syntactic information at a sentence level.

An example of integration of morphology and syntax is provided by the UD treebank for Beja (Kahane et al., 2021), a Cushitic language spoken in Sudan. In the treebank, a morph-based tokenization instead of a word-based one is adopted. All affixes are dependent on the stem and are assigned UD deprels corresponding to their functional role, with an additional `:aff` subtype (e.g., subject pronominal affixes are marked as `nsubj:aff`).

[5]For instance, *vice president* could be considered as a single word or split into two words.

3. Exploited Resources

For the present study, we exploit the resources described hereafter. UniSegments, UniMorph, and SIGMORPHON data are selected to obtain the segmentation, which we employ to manipulate UD trees. The selection of the languages primarily stems from their availability across all resources.[6]

UniSegments UniSegments (Žabokrtský et al., 2022) is a collection of harmonized versions of selected resources relevant for segmentation, whose data have been converted to a common scheme. It comprises 17 existing data resources featuring information about segmentation in 32 languages. The level of granularity of information varies across the different resources. Some of them classify segments specifying whether they are either roots, prefixes/suffixes, inflectional endings, or zero morph(eme)s; yet, despite using the same labels, they adopt different definitions of the classes. In the attempt to devise a truly shared schema, the creators of UniSegments chose to preserve the parts that require deep in-language expertise (e.g., lemmas), unify the information available in most resources (POS tags and, to some extent, segmentation), and keep as much of the language/resource-specific information as possible unchanged (Žabokrtský et al., 2022). This ensures a balance between the diverse levels of granularity observed in the resources but does not guarantee their full conformity. Inevitably, such discrepancies among the resources will be indirectly reflected in our data. At times, UniSegments includes more than one resource for the same language; in such cases, we select only one resource. We work with DeriNet (Vidra et al., 2021) for Czech, MorphoLex (Sánchez-Gutiérrez et al., 2018) for English, Demonette (Hathout and Namer, 2014) for French, DerIvaTario (Talamo et al., 2016) for Italian, WordFormationLatin (Litta et al., 2016) for Latin, and MorphyNet (Batsuren et al., 2021) for Catalan, Finnish, German, Hungarian, and Portuguese.

UniMorph The Universal Morphology (UniMorph) (McCarthy et al., 2020) project aims at providing instantiated normalized morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema. As far as the languages we include in our work are concerned, morphological information is extracted from Wiktionary (e.g., for Finnish) or derived from existing morphological dictionaries which are publicly hosted on the LINDAT/CLARIAH-CZ repository (for English, French, German, Italian).[7] Since information about vowel length is available for Latin data in UniMorph, data normalization is needed before undertaking the manipulation of nodes in treebanks.[8]

SIGMORPHON Some datasets were made available for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022). We choose to exploit Czech gold annotated data, as the quality of the results could prove to be positively affected.

Universal Dependencies A brief introduction to UD is available in Section 1. For the languages under study, we select the following treebanks from version 2.12 (Zeman, 2023). Whenever a Parallel Universal Dependencies (PUD) treebank (Zeman et al., 2017) is available we include it, as the PUD collection can provide interesting insights in terms of parallel, cross-lingual comparison. Additionally, we also select PDT (Hajič et al., 2020) for Czech, GUM (Zeldes, 2017) for English, TDT (Pyysalo et al., 2015) for Finnish, GSD (McDonald et al., 2013) for French and German, ISDT (Bosco et al., 2013) for Italian, and Bosque (Rademaker et al., 2017) for Portuguese. We employ AnCora (Taulé et al., 2008) for Catalan, Szeged (Vincze et al., 2010; Vincze et al., 2017) for Hungarian, and ITTB (Passarotti, 2019) for Latin, for which no PUD treebank is available.

4. Workflow

We now describe the strategy designed to process the selected data and extract from it all the exploitable information. It mainly revolves around two main tasks: on the one hand, the manipulation of nodes in treebanks based on the segmentation contained in the selected sources (Subsection 4.1); on the other hand, the process of alignment between universal features and morphs (Subsection 4.2). As a result, we release a set of treebanks where morphological segmentation is incorporated within the UD representation of syntactic dependencies.[9] How the morphs are integrated into the UD annotation is discussed in Subsection 4.3.

4.1. Manipulation of Treebank Nodes

As a first step, we convert the official UD treebanks to morphologically segmented treebanks, as described hereafter and illustrated in Figure 3.

To manipulate data we exploit Udapi (Popel et al., 2017), a framework providing an application programming interface for UD data. The code that performs the transformation is not language-specific,

[6] With the only exception of SIGMORPHON.

[7] Additionally, for some low-resource languages and dialects the data mainly comes from linguists who study them. Data augmentation in a semi-supervised way was also experimented with for Tagalog.

[8] For instance, *ă* and *ā* are normalized as *a*.

[9] Both the code and the set of treebanks are openly available at https://github.com/fjambe/feature-based-morpho-trees/.

provided that resources featuring morphological information (e.g., about segmentation, derivation, inflection) are available. It takes as input the UD treebank to manipulate and outputs a version of it where morphological trees of segmented words are blended in UD tree-shaped sentence representation, within a well-formed CoNLL-U file.

By iterating over each node, we check whether information about morphological segmentation of the node form or lemma (as further explained later) is available in any of the exploited resources, i.e. UniSegments and UniMorph mainly, as well as SIGMORPHON gold data for Czech.[10]

Step 0: SIGMORPHON data. In the case of Czech, we exploit SIGMORPHON manually annotated data as an additional resource. As a preliminary step in the workflow, for each form we first check whether it occurs in SIGMORPHON data; if it does, we split the form according to this segmentation. Since SIGMORPHON data only provides splitting, with no additional information about the resulting morphs, deciding which morph of the word should be considered the root is not straightforward. Thus, we decide to select as root the least frequent morph among those we identify within the word. Morph frequencies were calculated initially on the whole dataset. Whenever a form is found in SIGMORPHON, we then cease looking for possible additional segmentations, since forms in SIGMORPHON data are fully segmented. If, conversely, the word form is not retrieved at this stage, we continue with the procedure valid for all languages.

Step 1: segmented lemma. The first step consists of looking up the word lemma in UniSegments. If a match is found and a segmentation is available for the retrieved lemma,[11] the information just retrieved is now stored, to be exploited subsequently to segment the node. For instance, the Czech word *prokonzul* 'proconsul' is found in UniSegments as well as provided with a segmentation (*pro* + *konzul*).

Step 2: (un)inflected form, segmented lemma. Afterward, we check if the node form corresponds to its lemma, i.e. if the token is not an inflected form. If this is the case, we proceed to split the form based on the segmentation retrieved in UniSegments, as illustrated by the *prokonzul* example. Conversely, if the form is inflected we postpone the splitting until we have gathered more information about the word ending. For this purpose, we begin by verifying whether the form is listed in UniMorph, which comprises a catalog of inflected forms. If this proves to be the circumstance, we combine the information from UniSegments with information about inflection retrieved from UniMorph. See e.g. the Catalan plural form *culturals* 'cultural', whose lemma is split in UniSegments as *cultur* + *al*, while UniMorph provides the morph *-s* for plural. If, conversely, no match is found in UniMorph, we design a strategy to obtain an approximation of the inflectional ending by comparing character by character the two strings (form and lemma) and extract as ending the substring starting after the last shared character and extending till the end of the word form. It is the case of the English verb form *shortened*, split as *short* + *en* in UniSegments, and for which we extract the ending *-ed*.

Step 3: inflected form, unsegmented lemma. If the node lemma is not found in UniSegments, we inspect whether the node form occurs in UniMorph only. If it does, we extract the information from UniMorph and proceed to segment at least the inflectional ending of the word, as in the case of the French *travaillait* '(s)he worked', third person singular form of the imperfect tense of the verb *travailler* 'to work'. The form is segmented in UniMorph as *travailler* (lemma) + *ait* (ending).

Step 4: uninflected form, unsegmented lemma. In case the word is not comprised in either UniSegments or in UniMorph, i.e. if the node lemma and the node form do not represent entries of either of the two resources respectively, we do not implement any morphological splitting of the node and we proceed to the next one. That is, for instance, what happens with the Latin form *caelum* 'sky', corresponding to nominative, accusative, and vocative singular. Since for Latin nouns the nominative singular form is chosen as lemma, the form is not split in UniMorph; given that it is not segmented in UniSegments either, no morphological splitting can be performed on such form.

Practically, in the CoNLL-U file we handle morphologically segmented words as UD multi-word tokens (MWTs). Yet, such a decision may generate ambiguity, as it could be complex to distinguish original MWTs from morphological MWTs,[12] especially when they occur jointly (i.e., a MWT which we split further). Therefore, we decide to signal

[10] At this moment, we search only for a single best segmentation for each node, without handling possible ambiguities. Considering multiple segmentations may turns out to be necessary, especially in heavily ambiguous languages such as Arabic; morphological lattices (More et al., 2018) could be then useful for representing sets of alternative segmentations.

[11] Some of the lemmas included in UniSegments are not provided with a segmentation. See, for instance, Czech words *rok* 'year' or *jazyk* 'language', for which the only segment identified is the root, spanning over the whole word.

[12] Within the expression 'morphological MWT' we intend to use 'MWT' only in the technical sense of the UD label.

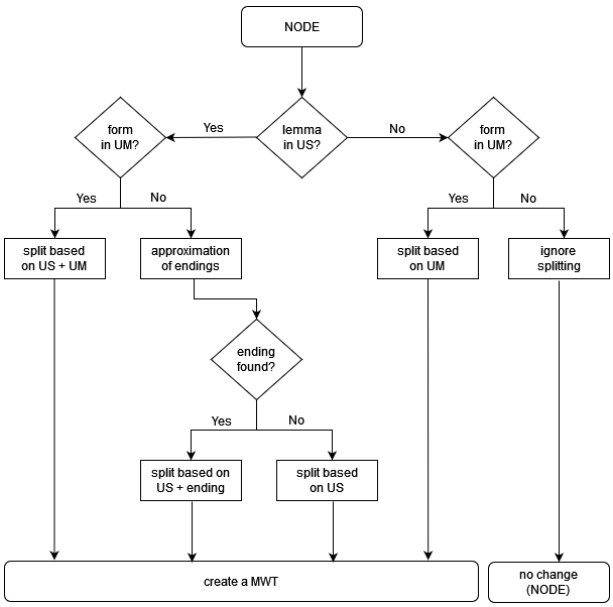


Figure 3: Flowchart of the node manipulation process (US: UniSegments, UM: UniMorph).

morphology-based split elements of MWTs through the deprel subtype `:morph` (see Subsection 4.3).

Since we are proposing a novel data representation, we have no gold data to rely on to assess the quality of the output of our algorithm. In light of this, we created a random sample of 20 French words and segmented them manually, which resulted in identifying 56 morphs. Of the 56 morphs in this gold data, 8 (14%) were correctly identified by UniSegments[13] alone, 18 (32%) by UniMorph alone, and 27 (48%) by our algorithm. Even though this sample is very small, it can be argued that combining the resources using our algorithm leads to a considerable improvement in the segmentation quality.

4.2. Feature Extraction

Additionally, by exploiting the statistical measures described hereafter we investigate whether and how morphs and UD feature sets align, to assess if specific feature inventories somehow capture the morphological meaning of morphs.

Similarly to what was done for node manipulation, we exploit the information contained in segmentation resources (in this case, UniSegments only) and in UD treebanks. Specifically, if a word form occurs in the treebank under study, and its lemma is also present in UniSegments, we segment it based on the segmentation provided by UniSegments.[14] For example, in Catalan the word *estacional* 'seasonal' is present in the UD Catalan AnCora treebank and also in UniSegments, following which it is split as *estacion* and *al*.[15] After having obtained the segmentations of the word forms from UniSegments, the UD feature set that is originally attributed to the word form is associated to the individual morphs the word form has been split into. For instance, the Hungarian word *gyerekek* 'children' in the UD Hungarian-Szeged treebank has the feature set `Number=Plur|Case=Nom`. Based on the segmentation data for Hungarian in UniSegments, the word form is split as *gyerek* + *ek*; we assign the original feature set to both *gyerek* and *ek*. In the following step, the feature set is split into individual features and is assigned to the morphs. As a result, we now have two instances of *gyerek*, one with feature `Number=Plur` and the other with feature `Case=Nom`; the same applies to *ek*. In this manner, for every possible feature, we create an inventory of morphs to which the feature has been associated. For each feature-morph pair we calculate the joint frequency of locating a morph given a feature and the ΔP scores (Jenkins and Ward, 1965). According to Schneider (2020), ΔP is a measure of cue validity, i.e. it measures how strongly two events are linked.[16] ΔP can be thus used to calculate collocation strength. Since it is a unidirectional dependency measure it can be decomposed in two distinct formulae, one for the forward-directed ΔP and the other for the backward counterpart. Using ΔP, we obtain the measure of the strength of correspondence between a morph and a feature, and vice versa. It is reasonable to use such a unidirectional measure because the association of a morph and a feature is asymmetric. The ΔP scores are between -1 and 1.

$$\Delta P_{forward} = P(m|f) - P(m|\neg f) \quad (1)$$

$$\Delta P_{backward} = P(f|m) - P(f|\neg m) \quad (2)$$

In equations (1) and (2), *m* stands for morph and *f* stands for feature. P(m|f) is the conditional probability of locating a morph given a feature among the other conditional probabilities in the equations. In Table 1, we present the ΔP forward and the ΔP backward scores for the morph *ing* in English given

[13] Specifically, among the available resources for French we selected Demonette.

[14] All the steps described in this paragraph are not applied to the same files employed for manipulation of treebank nodes.

[15] The example points out how morphological segmentation still presents several open issues. *Estacional* could probably be split further, by identifying *st* as the true core of the word, and *(c)ion* as another affix. We do not address the choices made in terms of segmentation, and work with the resources in their current state, however being aware that possible alternative segmentations could be proposed.

[16] The question how to extract combinations of features (conjunctions and disjunctions), which is relevant especially with inflectional affixes, is left for future research.

different morphological features. We find that the morph *ing* has the strongest relation with the feature `VerbForm=Ger`. What this indicates is the fact that the `VerbForm=Ger` strongly correlates to the morph *ing* as indicated by ΔP forward; the ΔP backward scores show the potential feature attributes like `Tense=Pres`, `VerbForm=Part` as well as the highest ranked feature `VerbForm=Ger`. Hence by comparing the ΔP forward and backward scores some signals could be extracted for morph and feature correspondences. While for a well-resourced language like English, such findings are not surprising, interesting correspondences could emerge in the case of less described languages.

Morph	Feature	ΔP forward	ΔP backward
ing	Degree=Pos	-0.058	-0.118
ing	Number=Sing	-0.090	-0.287
ing	Number=Plur	-0.091	-0.201
ing	Mood=Ind	-0.096	-0.144
ing	Person=3	-0.094	-0.127
ing	Tense=Pres	0.139	0.148
ing	VerbForm=Fin	-0.097	-0.152
ing	VerbForm=Part	0.120	0.136
ing	**VerbForm=Ger**	**0.966**	**0.710**
ing	Polarity=Neg	-0.004	-0.001

Table 1: Probabilities of the morph *ing* in English.

In Table 2, we observe that the morph *ung* in German has the highest ΔP scores for the feature `Gender=Fem`. The association with other features is due to the co-occurrence with other morphs in a word form. For example, the feature set for the German word *Kleidung* 'clothing' is `Case=Nom|Gender=Fem|Number=Sing`. The observed co-occurrences with other features can be explained by the allocation of the original features among the morphs *kleid* and *ung*. This correlation indicates that morphs can potentially be attributed to morphological features in an empirical sense, and by using such collocation measures it is possible to extract some informative signals.

Morph	Feature	ΔP forward	ΔP backward
ung	Case=Nom	0.129	0.226
ung	**Gender=Fem**	**0.467**	**0.798**
ung	Number=Sing	0.267	0.549
ung	Case=Dat	0.230	0.389
ung	Case=Acc	0.246	0.364
ung	Gender=Masc	-0.175	-0.230
ung	Case=Gen	0.152	0.116

Table 2: Probabilities of the morph *ung* in German.

In the case of Hungarian (Table 3), the morph *ek* has the strongest affinity for the feature `Number=Plur`. But there are other morphs too in Hungarian which are responsible for carrying the feature `Number=Plur`, like *ok*, *ak*, *ei* and *ai*. In the case of German too, there are multiple morphs (Table 4) that mark for the feminine gender, like *keit*, *schaft*, *enz*, and so on. Our current unsupervised approach successfully captures all the morphs attributed to a given morphological feature; we however reiterate that this finding is purely empirical given the available data resource.

Morph	Feature	ΔP forward	ΔP backward
ek	Case=Nom	-0.006	-0.146
ek	Number=Sing	-0.033	-0.431
ek	Person=3	0.031	0.427
ek	Definite=Ind	0.026	0.328
ek	PronType=Ind	0.064	0.099
ek	Mood=Ind	0.030	0.340
ek	Tense=Pres	0.032	0.344
ek	VerbForm=Fin	0.028	0.333
ek	Voice=Act	0.028	0.333
ek	**Number=Plur**	**0.163**	**0.531**

Table 3: Probabilities of the morph *ek* in Hungarian.

Morph	f(morph,feature)	ΔP forward	ΔP backward
ion	2	0.001	0.686
keit	59	0.053	0.697
heit	38	0.034	0.693
schaft	58	0.052	0.697
ung	**497**	**0.467**	**0.798**
enz	1	0.001	0.685

Table 4: Morphs for `Gender=Fem` in German.

From Table 5 and Table 6, we infer that the morphs *tunk* and *ok* both encode the features `Number=Plur` and `Person=1` in Hungarian. In the case of verbs conjugated in first person plural like *voltunk* 'we were' and *tanultunk* 'we studied' the morph *tunk* has the feature set `Number=Plur|Person=1`, whereas the morph *ok* has the feature `Number=Plur` for nouns and `Number=Plur|Person=2` for verbs (as in *tanultatok* 'you all studied'), as well as the feature `Person=1` (e.g. in *tanulok* 'I study').

Morph	f(morph,feature)	ΔP forward	ΔP backward
tunk	**1**	**0.033**	**0.972**
ok	**7**	**0.232**	**0.852**
ak	5	0.165	0.690
ek	5	0.163	0.531
ai	1	0.033	0.972

Table 5: Morphs for `Number=Plur` in Hungarian.

We do observe that a morph in Hungarian or any other language may take on multiple grammatical functions; we only cite these selected examples to highlight how polysemous morphs can be. Based on these feature sets extracted from UD it is possible to explore all the grammatical functions handled by the morphs across languages.

Based on the ΔP scores, we find that the morphological features more strongly associated with the Latin morph *us* are `Case=Nom`, `Gender=Masc` and `Number=Sing` (Table 7). The other features

Morph	f(morph,feature)	ΔP forward	ΔP backward
tunk	**1**	**0.143**	**0.994**
ok	**1**	**0.136**	**0.119**
om	1	0.141	0.328
tam	1	0.143	0.994
ttem	1	0.143	0.994

Table 6: Morphs for `Person=1` in Hungarian.

Morph	Feature	ΔP forward	ΔP backward
us	**Case=Nom**	**0.018**	**0.349**
us	Case=Acc	-0.012	-0.247
us	Case=Dat	-0.013	-0.130
us	Degree=Cmp	-0.012	-0.044
us	**Gender=Masc**	**0.017**	**0.369**
us	Gender=Fem	-0.018	-0.346
us	Gender=Neut	-0.013	-0.262
us	**Number=Sing**	**0.014**	**0.159**
us	Number=Plur	-0.018	-0.349

Table 7: Probabilities of the morph *us* in Latin.

attributed to the morph *us* are potentially due to the feature values of the lexical root morph it happens to co-occur with. The ΔP backward scores indicate the morph *us* has a strong correspondence with the feature `Gender=Masc`.[17]

Given the observations, ΔP proves to be a strong unsupervised measure that extracts features associated with morphs, which potentially indicates that morphs do carry morphological features and in any case it would be reasonable to use this information to analyze word-internal structure in more detail.

4.3. Conforming to UD

When morphologically segmenting the nodes of a treebank, a natural question that arises concerns how to annotate morphs within UD. Specifically, when creating the morphological MWT we need to assign to its elements lemma, POS, morphological features, and deprel.

In many cases when segmentation is provided, UniSegments also comprises information about morphemes; namely, a word morph is possibly associated with its corresponding morpheme. For instance, the Latin verb *auerto* 'to turn away' is split as *a* + *uerto*, with the morph *a* associated to the morpheme *a(b)*, which can indeed take both forms *a* and *ab*. When available, we adopt the provided morpheme as a lemma; otherwise, we set the morph lemma to be identical to its form. We assign the POS that the node originally has (i.e., before undergoing the segmentation) to the head of MWT, which should correspond to the stem of the word.

[17] However, this correlation comes purely from the data we have in hand. Theoretically, the morph *us* in Latin can equally express e.g. `Case=Nom`, `Gender=Masc`, and `Number=Sing`. Currently, we do not have a baseline to compare our empirical findings with theoretical facts.

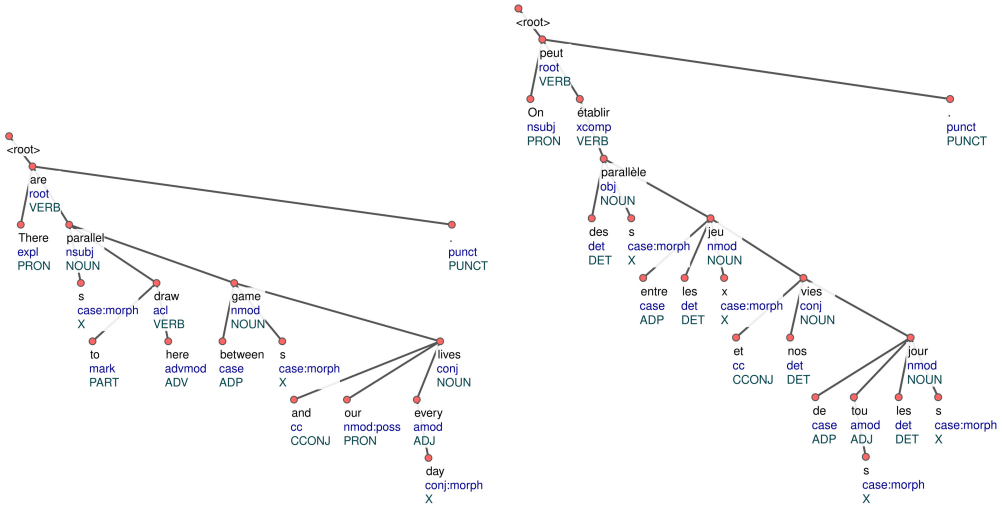
All other tokens of the MWT, i.e. morphs, receive the POS tag X. Indeed, we decide not to tag them with labels describing their position with respect to the stem (e.g., prefix, suffix) or the morphological process they convey (e.g., inflection, derivation). By assigning the X UPOS tag, we try to be as compliant as possible to UD, although without affirming that we believe morphs to have a POS.

To annotate features, we exploit the feature-based alignment presented in Subsection 4.2. Specifically, for each of the morphological segments that we identify, we search for the features that are associated with them as a result of the alignment process. If any of those features can also be found in the original feature set of the token, we assign it to the morph and remove it from the set of features of the root, as we believe it to belong to the morph instead of the root.

When assigning deprels, we handle prefixes, root(s), and suffixes in a slightly different manner. Prefixes, extracted from UniSegments, are assigned `nmod:morph` if they are substantives (NOUN/PROPN), `advmod:morph` for all other POSs. If according to UniSegments the lemma presents just a single root, it inherits the deprel that the node originally had. If more than one is found, the second (and possibly more) is annotated as `conj:morph`. It is the case of compounds, for which the choice of `conj` is justified by the fact that we want all the lexical stems to be somehow on the same level. We are aware that parataxis is not the only possible relation between words constituting a compound (cf. Svoboda and Ševčíková 2024); however, we adopt this practical solution since the type of compound structure is not annotated in the exploited resources. As of now, we intend to use `conj:morph` only as a way to point out the co-existence of two lexical roots. In the case of suffixes, we try to approximately distinguish verbal and nominal inflection. Segmented morphs of verbs and auxiliaries are assigned `aux:morph`, while `case:morph` applies to nouns, adjectives, determiners, pronouns, adverbs, numerals, and extremely rare instances of adpositions. Whenever we are not able to reasonably assign either of the two deprels, we opt for `dep:morph`. As mentioned in the previous subsection, the `:morph` subtype allows to distinguish and retrieve all instances of morphological segmentations.

5. MorphoTrees

Figures 4(a), 4(b), and 4(c) display the same sentence, corresponding to English *There are parallels to draw here between games and our everyday lives*. The sentence, extracted from PUD treebanks, is shown also in Finnish and French and provides an example of how including the internal structure of



(a) English (b) French

(c) Finnish

Figure 4: UD-morphological tree of the sentence *There are parallels to draw here between games and our everyday lives* in English, French, and Finnish respectively.

words into UD could provide interesting remarks. Indeed, parallel data available in PUD could be observed in an even more parallel perspective after morph splitting, as in different languages some features could be realized differently, but a similar approach could help align them. In Appendix A we also display the raw CoNLL-U representation of the sentences (Figures 7, 8, 9), in order for the features and the MWT-like strategy to be visible.

In the Finnish example in Figure 4(c), the word form *jokapäiväisten* 'everyday ones' is split as *jokapäiväi* and *sten*. *Jokapäiväi* gets the POS tag ADJ and the deprel `amod` and the morph *sten* gets the deprel `case:morph` as decided. In the English example in Figure 4(a), the word form *games* is split as *game* and *s* where the morph *s* gets the deprel `case:morph`. The compound *everyday* is split and *day* is attached as `conj:morph` to *every*.[18] Similar splits can be also observed in the French example in Figure 4(b).[19] Figures 5 and 6 show the integration of segmentation within non-PUD treebanks.

[18] *Everyday* clearly shows a case where the two elements of the compound are attached paratactically according to our solution, whereas *every* is actually dependent on *day* within the structure of the compound.

[19] The example can also serve to highlight how the segmentation of the exploited resources, and hence its quality and level of granularity, is inherited in our data. For instance, in the verb *établir* the infinitive marker *ir* should be segmented, while it is not. Of course, this kind of choice also strongly depends on the adopted approach to morphological segmentation, which is far from being a solved problem yet. A similar observation would probably apply to Finnish as well, where some expected segmentations may be missing.

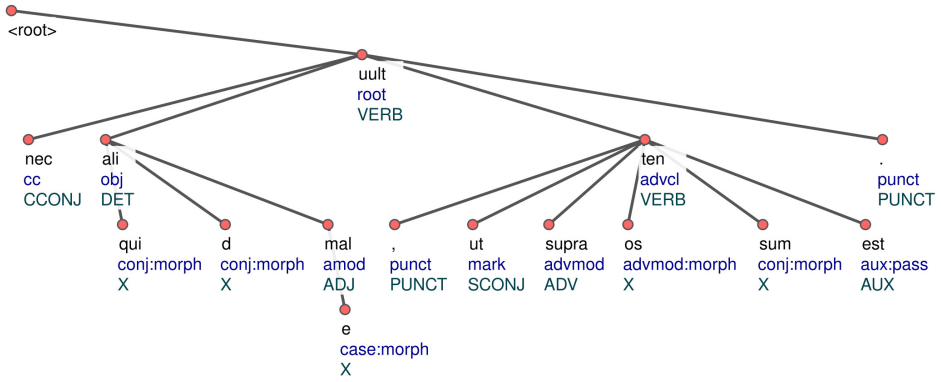


Figure 5: UD-morphological tree of the Latin sentence *Nec aliquid male uult, ut supra ostensum est.* ('Nor does he will anything evil, as we have proved.').

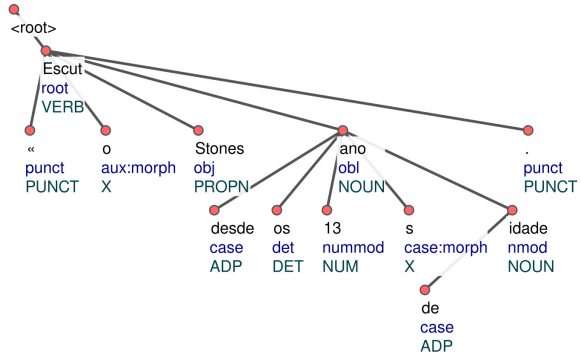


Figure 6: UD-morphological tree of the Portuguese sentence *Escuto Stones desde os 13 anos de idade.* ('I've been listening to the Stones since I was 13.').

6. Conclusion and Future Work

In the paper, we presented the proposal of a novel data structure aiming at integrating the representation of the morphological internal structure of words into Universal Dependencies. Working on 10 languages as a case study, we first devised a prototype of a methodology to manipulate UD treebanks intending to include the morphological structure of words into the canonical UD-like sentence representation. Then, we investigated the alignment between morphs and feature sets, by calculating ΔP scores that indicate the strength of the relation between a morph and a feature, and proceeded to assign relevant morphological features to morphs. Both tasks exploited already existing resources to perform segmentation. Such an approach ties the quality of our data to that of the resources we employed, for which some limitations were observed (derived e.g. from conversion from different resources).

Overall, the work we presented does not intend to suggest a reorganization of Universal Dependencies towards the inclusion of internal, morphological word structure. Our goal is to provide a resource that integrates morphology and syntax, two linguistic layers often intertwining, and that can prove beneficial in enhancing comparability of languages that express comparable meaning through different grammatical strategies[20]. The key factor for enhancing comparability lies in the cross-lingual correspondence of morphs.

In the future, we plan to improve the described workflow and expand the collection of morphological treebanks to more languages. Additionally, the extraction of the morphological trees from the sentence representation could be explored, towards their possible integration with DeriNet (Vidra et al., 2021). Moreover, in recent developments, morphological features are used to create multilingual morphological analyzers, for instance as presented by Pawar et al. (2023). We would like to carry forward our current research in that direction too by including a larger set of languages, as well as by including phenomena that we have neglected so far, such as non-concatenative morphology. We will find ways to estimate the quality of the resulting trees.

7. Acknowledgements

This work has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062). The study was supported by the Charles University, project GA UK No. 104924 and project GA UK No. 101924; and partially supported by SVV project number 260 698. We would like to thank three anonymous reviewers for their very insightful feedback.

[20] Most notably, different degrees of inflection.

8. Bibliographical References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Barli Bram. 2012. Three models of English morphology. *LLT Journal: A Journal on Language and Language Teaching*, 15(1):179–185.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

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

Robert M. Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]. *Traitement automatique des langues*, 60(2):71–95.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

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

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. In *Linguistic Issues in Language Technology, Volume 11, 2014-Theoretical and Computational Morphology: New Trends and Synergies*.

Herbert M. Jenkins and William C. Ward. 1965. Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79(1):1–17.

D. Jurafsky and J.H. Martin. 2014. *Speech and Language Processing*. Always learning. Pearson.

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.

Zhongguo Li. 2011. Parsing the internal structure of words: A new paradigm for Chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1414, Portland, Oregon, USA. Association for Computational Linguistics.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a word formation lexicon for Latin. In *CLiC-it/EVALITA*.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David

Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Amir More, Özlem Çetinoğlu, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji, and Reut Tsarfaty. 2018. Conll-ul: Universal morphological lattices for universal dependency parsing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320.

Siddhesh Pawar, Pushpak Bhattacharyya, and Partha Talukdar. 2023. Evaluating cross lingual transfer for morphological analysis: a case study of Indian languages. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 14–26, Toronto, Canada. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. 2018. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior research methods*, 50:1568–1580.

Ulrike Schneider. 2020. ΔP as a measure of collocation strength. Considerations based on analyses of hesitation placement in spontaneous speech. *Corpus Linguistics and Linguistic Theory*, 16(2):249–274.

Petra Steiner. 2017. Merging the trees - building a morphological treebank for German from two resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 146–160, Prague, Czech Republic.

Gregory T Stump. 2001. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press.

Emil Svoboda and Magda Ševčíková. 2024. Compounds in Universal Dependencies: A survey in five European languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 88–99, St. Julian's, Malta. Association for Computational Linguistics.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Veronika Vincze, Katalin Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal Dependencies and morphology for Hungarian - and on the price of universality. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 356–365, Valencia, Spain. Association for Computational Linguistics.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards Universal Segmentations: UniSegments 1.0. In *Proceedings*

of the Thirteenth Language Resources and Evaluation Conference, pages 1137–1149, Marseille, France. European Language Resources Association.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 125–134, Sofia, Bulgaria. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1326–1336, Baltimore, Maryland. Association for Computational Linguistics.

Hai Zhao. 2009. Character-level dependencies in Chinese: Usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 879–887, Athens, Greece. Association for Computational Linguistics.

9. Language Resource References

Baayen, R. Halard and Piepenbrock, Richard and Gulikers, Leon. 1995. *CELEX2*. Linguistic Data Consortium, ISLRN 204-698-863-053-1.

Vidra, Jonáš and Žabokrtský, Zdeněk and Kyjánek, Lukáš and Ševčíková, Magda and Dohnalová, Šárka and Svoboda, Emil and Bodnár, Jan. 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. PID http://hdl.handle.net/11234/1-3765.

Žabokrtský, Zdeněk and Bafna, Nyati and Bodnár, Jan and Kyjánek, Lukáš and Svoboda, Emil and Ševčíková, Magda and Vidra, Jonáš and Angle, Sachi and Ansari, Ebrahim and Arkhangelskiy, Timofey and Batsuren, Khuyagbaatar and Bella, Gábor and Bertinetto, Pier Marco and Bonami, Olivier and Celata, Chiara and Daniel, Michael and Fedorenko, Alexei and Filko, Matea and Giunchiglia, Fausto and Haghdoost, Hamid and Hathout, Nabil and Khomchenkova, Irina and Khurshudyan, Victoria and Levonian, Dmitri and Litta, Eleonora and Medvedeva, Maria and Muralikrishna, S. N. and Namer, Fiammetta and Nikravesh, Mahshid and Padó, Sebastian and Passarotti, Marco and Plungian, Vladimir and Polyakov, Alexey and Potapov, Mihail and Pruthwik, Mishra and Rao B, Ashwath and Rubakov, Sergei and Samar, Husain and Sharma, Dipti Misra and Šnajder, Jan and Šojat, Krešimir and Štefanec, Vanja and Talamo, Luigi and Tribout, Delphine and Vodolazsky, Daniil and Vydrin, Arseniy and Zakirova, Aigul and Zeller, Britta. 2022. *Universal Segmentations 1.0 (UniSegments 1.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zeman, Daniel et al. 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. PID http://hdl.handle.net/11234/1-5150.

A. Example sentences in ConLL-U format

```
1       There   there   PRON    EX      _       2       expl
2       are     be      VERB    VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        0
3-4     parallels       _       _       _       _       _       _
3       parallel        parallel        NOUN    NNS     _       2       nsubj
4       s       s       X       _       Number=Plur     3       case:morph
5       to      to      PART    TO      _       6       mark    5:mark  _
6       draw    draw    VERB    VB      VerbForm=Inf    3       acl
7       here    here    ADV     RB      PronType=Dem    6       advmod
8       between between ADP     IN      _       9       case
9-10    games   _       _       _       _       _       _
9       game    game    NOUN    NNS     _       3       nmod
10      s       s       X       _       Number=Plur     9       case:morph
11      and     and     CCONJ   CC      _       15      cc
12      our     we      PRON    PRP$    Number=Plur|Person=1|Poss=Yes|PronType=Prs      15      nmod:poss
13-14   everyday        _       _       _       _       _       _
13      every   every   ADJ     JJ      _       15      amod
14      day     day     X       _       Degree=Pos      13      conj:morph
15      lives   life    NOUN    NNS     Number=Plur     9       conj
16      .       .       PUNCT   .       _       2       punct
```

Figure 7: CoNLL-U representation of the English sentence *There are parallels to draw here between games and our everyday lives* (see also 4(a), 4(c), 4(b)). All three figures in the appendix allow us to better understand how morphological features have been treated. In the CoNLL-U files shown here the ninth and tenth fields have been removed, for reasons of space, as they are not strictly relevant to what is discussed in the present work.

```
1       Pelien  peli    NOUN    _       Case=Gen|Number=Plur    8       obl
2       ja      ja      CCONJ   _       _       5       cc
3-4     jokapäiväisten  _       _       _       _       _       _
3       jokapäiväi      jokapäiväi      ADJ     _       Case=Gen|Degree=Pos|Derivation=Inen|Number=Plur 5       amod
4       sten    sten    X       _       Case=Gen|Degree=Pos|Derivation=Inen|Number=Plur 3       case:morph
5       elämiemme       elämä   NOUN    _       Case=Gen|Number=Plur|Number[psor]=Plur|Person[psor]=1   1       conj
6       välillä välillä ADP     _       AdpType=Post    1       case
7       on      olla    AUX     _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 8       aux:pass
8       löydettävissä   löytää  VERB    _       Case=Ine|Number=Plur|PartForm=Pres|VerbForm=Part|Voice=Pass     0       root
9       yhtäläisyyksiä  yhtäläisyys     NOUN    _       Case=Par|Number=Plur    8       obj
10      .       .       PUNCT   _       _       8       punct
```

Figure 8: CoNLL-U representation of the Finnish sentence.

```
1       On      on      PRON    Number=Sing|Person=3|PronType=Ind   2   nsubj
2       peut    pouvoir VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root
3       établir établir VERB    VerbForm=Inf    2   xcomp
4       des     un      DET     Definite=Ind|Gender=Masc|Number=Plur|PronType=Art   5   det
5-6     parallèles  _   _   _   _   _
5       parallèle   parallèle   NOUN    Gender=Masc|Number=Plur 3   obj
6       s       s       X       Gender=Masc|Number=Plur 5   case:morph
7       entre   entre   ADP _   9   case
8       les     le      DET     Definite=Def|Gender=Masc|Number=Plur|PronType=Art   9   det
9-10    jeux    _   _   _   _   _
9       jeu     jeu     NOUN    Gender=Masc|Number=Plur 5   nmod
10      x       x       X       Gender=Masc|Number=Plur 9   case:morph
11      et      et      CCONJ   _   13  cc
12      nos     son     DET     Gender=Fem|Number=Plur|Number[psor]=Plur|Person=1|Person[psor]=1|Poss=Yes|PronType=Prs  13  det
13      vies    vie     NOUN    Gender=Fem|Number=Plur  9   conj
14      de      de      ADP     _   18  case
15-16   tous     _   _   _   _   _
15      tou     tou     ADJ     Gender=Masc|Number=Plur 18  amod
16      s       s       X       Gender=Masc|Number=Plur 15  case:morph
17      les     le      DET     Definite=Def|Gender=Masc|Number=Plur|PronType=Art   18  det
18-19   jours   _   _   _   _   _
18      jour    jour    NOUN    Gender=Masc|Number=Plur 13  nmod
19      s       s       X       Gender=Masc|Number=Plur 18  case:morph
20      .       .       PUNCT   _   2   punct
```

Figure 9: CoNLL-U representation of the French sentence.