



HAL
open science

Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, Amir Zeldes

► To cite this version:

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, et al.. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*, 2022, 57 (2), pp.493-544. 10.1007/s10579-022-09581-9. hal-04629571

HAL Id: hal-04629571

<https://hal.science/hal-04629571>


Submitted on 29 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations

Manuela Sanguinetti¹  · Cristina Bosco² · Lauren Cassidy³ ·
Özlem Çetinoglu⁴ · Alessandra Teresa Cignarella^{2,5} · Teresa Lynn³ ·
Ines Rehbein⁶ · Josef Ruppenhofer⁷ · Djamé Seddah⁸ · Amir Zeldes⁹

Accepted: 13 January 2022 / Published online: 20 February 2022
© The Author(s) 2022

Abstract This article presents a discussion on the main linguistic phenomena which cause difficulties in the analysis of user-generated texts found on the web and in social media, and proposes a set of annotation guidelines for their treatment within the Universal Dependencies (UD) framework of syntactic analysis. Given on the one hand the increasing number of treebanks featuring user-generated content, and its somewhat inconsistent treatment in these resources on the other, the aim of this article is twofold: (1) to provide a condensed, though comprehensive, overview of such treebanks—based on available literature—along with their main features and a comparative analysis of their annotation criteria, and (2) to propose a set of tentative UD-based annotation guidelines, to promote consistent treatment of the particular phenomena found in these types of texts. The overarching goal of this article is to provide a common framework for researchers interested in developing

Manuela Sanguinetti and Lauren Cassidy are joint first authors.

✉ Manuela Sanguinetti
manuela.sanguinetti@unica.it

- ¹ Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Cagliari, Italy
- ² Dipartimento di Informatica, Università degli Studi di Torino, Turin, Italy
- ³ ADAPT Centre, Dublin City University, Dublin 9, Ireland
- ⁴ IMS, University of Stuttgart, Stuttgart, Germany
- ⁵ PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
- ⁶ University of Mannheim, Mannheim, Germany
- ⁷ Leibniz-Institut für Deutsche Sprache, Mannheim, Germany
- ⁸ INRIA, Paris, France
- ⁹ Georgetown University, Washington, USA

similar resources in UD, thus promoting cross-linguistic consistency, which is a principle that has always been central to the spirit of UD.

Keywords Web · Social media · Treebanks · Universal Dependencies · Annotation guidelines · UGC

1 Introduction

The immense popularity gained by social media in the last decade has made it an attractive source of data for a large number of research fields and applications, especially for sentiment analysis and opinion mining (Balahur, 2013; Severyn et al., 2016). In order to successfully process the data available from such sources, linguistic analysis is often helpful (Mataoui et al., 2018; Vilares et al., 2017), which in turn prompts the use of NLP tools to that end. Despite the ever increasing number of contributions, especially on part-of-speech tagging (Behzad & Zeldes, 2020; Bosco et al., 2016; Çetinoğlu & Çöltekin, 2016; Gimpel et al., 2011; Lynn et al., 2015; Owoputi et al., 2013; Proisl, 2018; Rehbein et al., 2018) and parsing (Foster, 2010; Kong et al., 2014; Liu et al., 2018; Petrov & McDonald, 2012; Sanguinetti et al., 2018), automatic processing of user-generated content (UGC) still represents a challenging task, as it is shown by some tracks of the workshop series on noisy user-generated text (W-NUT).¹ UGC is a continuum of text sub-domains that vary considerably according to the specific conventions and limitations posed by the medium used (blog, discussion forum, online chat, microblog, etc.), the degree of “canonicalness” with respect to a more standard language, as well as the linguistic devices² adopted to convey a message. Overall, however, there are some well-recognized phenomena that characterize UGC as a whole (Eisenstein, 2013; Foster, 2010; Seddah et al., 2012), and that continue to make its treatment a difficult task.

As the availability of training resources developed on an ad hoc basis remains an essential factor for the analysis of these texts, the last decade has seen numerous resources of this type being developed. A good proportion of those resources that contain syntactic analyses have been annotated according to the Universal Dependencies (UD) scheme (Nivre et al., 2020), a dependency-based scheme which has become a popular standard reference for treebank annotation because of its adaptability to different domains and genres. At the time of writing (in UD version 2.8), as many as 114 languages are represented within this vast project, with 202 treebanks dealing with extremely varied genres, ranging from news to fiction, medical, legal, religious texts, etc. This linguistic and textual variety demonstrates the generality and adaptability of the annotation scheme.

On the one hand, this flexibility opens up the possibility of also adopting the UD scheme for a broad range of user-generated text types, since a framework which is proven to be readily adaptable is more likely to fit the needs of diverse UGC data

¹ <https://noisy-text.github.io/>.

² This phrase is used here in a broad sense to indicate all those orthographic, lexical as well as structural choices adopted by a user, often for expressive purposes.

sources, and the wealth of existing materials makes it potentially easier to find precedents for analysis whenever difficult or uncommon constructions are encountered. On the other hand, the current UD guidelines do not fully account for some of the specifics of UGC domains, thus leaving it to the discretion of the individual annotator (or teams of annotators) to interpret the guidelines and identify the most appropriate representation of these phenomena. This article therefore draws attention to the annotation issues of UGC, while attempting to find a cross-linguistically consistent representation, all within a single coherent framework. It is also worth pointing out that inconsistencies may be found even among multiple resources in the same language (see e.g. Aufrant et al. (2017) and Björkelund et al. (2017)).³ Therefore, even on the level of standardizing a common solution for UGC and other treebanks in one language, some more common guidance taking UGC phenomena into account is likely to be useful. This article first provides an overview of the existing resources—treebanks in particular—of user-generated texts from the Web, with a focus on comparing their varying annotation choices with respect to certain phenomena typical of this domain. Next, we present a systematic analysis of some of these phenomena within the context of the framework of UD, surveying previous solutions, and propose, where possible, guidelines aimed at overcoming the inconsistencies found among the existing resources (see also the “[Appendix](#)” where our proposal is summarized).

Given the nature of the phenomena covered and the fact that the existing relevant resources only cover a handful of languages, we are aware that the debate on their annotation is still wide open; therefore the primary intent of this article is not to prescribe, but rather, propose guidelines. That said, the proposals in this article represent the consensus of a fairly large group of UD contributors working on diverse languages and media, with the goal of building a critical mass of resources that are annotated in a consistent way. As such, it can be used as a reference when considering alternative solutions, and it is hoped that the survey of treatments of similar phenomena across resources will help future projects in making choices that are as comparable as possible to common practices in the existing datasets.

The present paper is an extended version of a manuscript accepted at the 12th Language Resources and Evaluation Conference (LREC 2020) (Sanguinetti et al., 2020). With respect to the latter, the current annotation proposals have been partially revised and expanded.

2 Linguistics of UGC

Describing all challenges brought about by UGC for all languages is beyond the scope of this work. Nevertheless, following Foster (2010), Seddah et al. (2012) and Eisenstein (2013) we can characterize UGC’s idiosyncrasies along a few major dimensions defined by the intentionality or communicative needs that motivate

³ Additionally, the different French resources currently available in the repository are also an example of such inconsistencies, which are mostly due to different annotation choices inherited from different linguistic traditions (e.g. Fr_ParTUT vs Fr_Sequoia) or different annotator teams (e.g. SynTagRus vs GSD_Russian).

linguistic variation. It should be stressed that one and the same utterance, and indeed often a single word, can instantiate multiple categories from the selection below, and that their occurrence can be either intentional or unintentional.⁴

The major dimensions defined to characterize UGC include the following:

- *Encoding simplification* This category covers ergographic phenomena, i.e. phenomena aiming to reduce the effort of writing, such as diacritic or vowel omissions (EN *people* → *ppl*).
- *Boundary shifting* Some phenomena affect the number of tokens, compared to standard orthography, either by replacing several standard language tokens by only one, which we will refer to as *contraction* (FR *n'importe* → *nimp* 'whatever, rubbish') or conversely by splitting one standard language token into several tokens, which we will refer to as *over-splitting* (FR *c'était* → *c t*, 'it was'). In some cases, the resulting non-standard tokens might even be homographs of existing words, creating more ambiguities if not properly analyzed. Such phenomena are frequent in the corpora of UGC surveyed below, and they require specific annotation guidelines.
- *Marks of expressiveness* orthographic variation is often used as a mark of expressiveness, e.g., graphical stretching (*yes* → *yesssss*), replication of punctuation marks (? → ?????), as well as emoticons, which can also take the place of standard language words, e.g. a noun or verb (FR *Je t'aime* → *Je t' < 3*, 'I love you', with the heart emoticon representing the verb 'love'). These phenomena often emulate sentiment expressed through prosody, facial expression and gesture in direct interaction; however the written nature of UGC data means that they need to be assigned analyses in terms of tokens, parts of speech and dependency functions. Many of the symbols involved also contain punctuation, which can lead to spurious tokenization and problems in lemmatization (see below).
- *Foreign language influence* UGC is often produced in highly multilingual settings and we often find evidence of the influence of foreign language(s) on the users' text productions, especially in code-switching (CS) scenarios, in domain-specific conversations (video game chat logs) or in the productions of L2 speakers, all of which complicate the typically monolingual context for which syntactic annotation guidelines are developed. In some cases, foreign words are imported as is from a donor language (e.g. IT *non fare la bad girl* 'don't be a bad girl' instead of *non fare la cattiva ragazza*). In other cases, foreign influence can create novel words: a good example is an Irish term coined by one user to mean 'awkward', *áicbheaird*, whose pronunciation mimics the English word (instead of the equivalent standard Irish term *amscaí*).
- *Medium-dependent phenomena* Some deviations from standard language are direct results of the electronic medium, including client-side automatic error correction, masking or replacement of taboo words by the server, artifacts of the

⁴ In fact, because of the inherent uncertainty in interpreting corpus utterances, coupled with the often highly contextual nature of UGC, it is important to apply analyses that are as independent as possible from definitions referring to speaker or writer intentions.

keyboard or other user input devices, and more. In some cases, and especially for languages other than English, some apparent English words in UGC represent automatic ‘corrections’ of non-English inputs, such as Irish *coicise* ‘fortnight’ → *concise*. These cases raise questions relating to the degree of interpretation, such as reconstructing likely UGC inputs before error correction, which may need to be annotated either as typos (in UD, the annotation `Typo=Yes`), or at an even greater level of detail in lemmatization.

- *Context dependency* Given the conversational nature of most social media, UGC data often exhibits high context-dependence (much like dialogue-based interaction). Speaker turns in UGC are often marked by the thread structure in a Web interface or app, and information from across a thread may provide a rich context for varying levels of ellipsis and anaphora that are much less frequent or complex in standard written language. In addition, multimedia content, pictures or game events can serve as a basis for discussion and are used as external context points, acting, so to speak, as non-linguistic antecedents or targets for deixis and argument structure. This can make the annotation task more difficult and prone to interpretation errors—especially if the actual thread context is not available—and requires establishing specific annotation conventions. .

As a supplementary material, we have included in “[Appendix](#)” the diagram that displays the hierarchy we followed to describe UGC phenomena (see Fig. 31), along with a number of examples in the different languages of such phenomena (Table 2). Our focus in this paper is on the noncanonical linguistic phenomena prevalent in UGC which do not yet have standardized annotation guidelines within the UD framework.

3 UGC treebanks: an overview

In order to provide an account of the resources described in the literature, we carried out a semi-systematic search on Google Scholar using the following set of keywords (*treebank web social media*) and (*universal dependencies web social media*), limiting to the first five pages, sorted by relevance, and without time filters.⁵ We selected only open-access papers describing either a novel resource or an already-existing one that has been expanded or altered in such a way that it gained the status of a new one. In the few cases of multiple publications referring to the same resource, we chose the most recent one, assuming it contained the most up-to-date information regarding the status of the resource. We also included in our collection five papers that we were aware of, but which were not retrieved by the search. As the main focus of this work is on the syntactic annotation of web content and user-generated texts, we discarded all papers that presented system descriptions, parsing experiments or POS-tagged resources (without syntactic annotation). Finally, we added in the overview the treebanks available in the official UD repository featuring

⁵ The main search was carried out on October 2019, but results were last updated on June 2021.

UGC data of some kind and for which a reference paper is not available at the time of writing (therefore it could not be found with the literature search). The results of our search are summarized in Table 1.⁶

Based on the selection criteria mentioned above, we found 24 papers and a total amount of 30 resources featuring web/social media texts; most of them are freely available, either from a GitHub/BitBucket repository, a dedicated web page or upon request. Dataset sizes vary widely, ranging from 500 (DWT) to approximately 6700 tweets (Pst) for the Twitter treebanks, and from 974 (xUGC) to more than 16,000 sentences (EWT) for the other datasets.

3.1 Languages

English is the most represented language, however, some of the resources focus on different English language varieties such as African-American English (TAAE), Singaporean English (STB), and Hindi-Indian English code-switching data (Hi-En-CS). Three resources are in French (Frb, xUGC, FSMB), one includes CS data in French and transliterated dialectal North-African Arabic (NBZ), two in Finnish (TDT, OOD) and two in Italian (TWRO, Pst); the remaining ones are in Arabic (ATDT), Belarusian (HSE), Chinese (CWT), Estonian (EtWT), German (tweeDe), Irish (TwIr), Manx (Cdh), Russian (Taiga), Spanish and Latin American Spanish (LDF), Turkish (ITU) and Ukrainian (IU).

3.2 Data sources

16 out of 30 resources are either partially or entirely made up of Twitter data. Possible reasons for this are the easy retrieval of the data by means of the Twitter API and by the use of wrappers for crawling the data, as well as the policy adopted by the platform with regard to the use of data for academic and non-commercial purposes.⁷ Only four resources include data from social media other than Twitter, specifically Facebook (FSMB, Taiga), Reddit (GUM), Sina Weibo (CWT), Instagram, YouTube and VK (Taiga), and, overall, most of the remaining resources comprise texts from discussion fora of various kinds. Only three treebanks consist of texts from different sub-domains, i.e. newspaper fora (NBZ), blogs, reviews, emails, newsgroups and question answers (EWT), and Wikinews, Wikivoyage, wikiHow, Wikipedia biographies, interviews, academic writing, Creative Commons fiction (GUM). Two resources are made up of generic data automatically crawled from the web (EtWT, TDT).

⁶ A more complete table with additional information on the surveyed treebanks can be found here: <http://di.unito.it/webtreebanks>.

⁷ <https://developer.twitter.com/en/developer-terms/agreement-and-policy#c-respect-users-control-and-privacy>.

Table 1 Overview of treebanks featuring user-generated content that formed the basis of this research, along with some basic information on the data source, the languages involved and whether they are based on UD scheme or not. In non-UD treebanks, † and ★ indicate, respectively, a constituency or dependency-based syntactic representation (AAE African-American English, MAE Mainstream American English, AR Arabic, BE Belarusian, DE German, DZFR Dialectal North-African Arabic/French code-switching, EN English, ES Spanish, ET Estonian, FI Finnish, FR French, GA Irish, GV Manx, HI/EN Hindi-English code-switching, IT Italian, RU Russian, SgE Singapore English, TR Turkish, UK Ukrainian, ZH Chinese)

Name	References	Source	Language	UD-based
ATDT	Albogamy and Ramsay (2017)	Twitter	AR	Yes
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN	Yes
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE	Yes
TWITTIRO-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT	Yes
DWT	Daiber and Van Der Goot (2016)	Twitter	EN	No★
W2.0	Foster et al. (2011)	Twitter, sort fora	EN	No†
Forebank (Frb)	Kaljahi et al. (2015)	Technical fora	EN, FR	No†
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN	No★
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN	Yes
TDT	Luotolahti et al. (2015)	Various	FI	Yes
xUGC	Martínez Alonso et al. (2016)	Various	FR	Yes
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	Various	ET	Yes
ITU	Pannay et al. (2015)	n.a.	TR	No★
WDC	Read et al. (2012b)	Various	EN	No†
tweeDe	Rehbein et al. (2019)	Twitter	DE	Yes
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT	Yes
FSMB	Seddah et al. (2012)	Twitter, Facebook, discussions fora	FR	No†
Narabizi (NEZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR	Yes
EWT	Silveira et al. (2014)	Various	EN	Yes
LAS-DisFo (LDF)	Taulé et al. (2015)	Discussion fora	ES	No†
MoNoise (MNO)	Van Der Goot and van Noord (2018)	Twitter	EN	Yes
STB	Wang et al. (2017)	Discussion fora	SgE	Yes

Table 1 continued

Name	References	Source	Language	UD-based
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH	No*
GUM	Zeldes (2017)	Various	EN	Yes
HSE	n.a.	Various	BE	Yes
OOD	n.a.	Various	FI	Yes
TwitIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA	Yes
Cadhan (Cdh)	n.a.	Various	GV	Yes
Taiga	n.a.	Various	RU	Yes
IU	n.a.	Various	UK	Yes

3.3 Syntactic frameworks

With regard to the formalism adopted to represent the syntactic structure, dependencies are by far the most used paradigm, especially among the treebanks created from 2014 onwards, though some resources include both constituent and dependency syntax versions—EWT has manually annotated constituent trees, while GUM contains automatic constituent parses based on parser output from CoreNLP (Manning et al., 2014) applied to the gold POS tags. As pointed out by Martínez Alonso et al. (2016), dependency-based annotation lends itself well to noisy texts, since it is easier to deal with disfluencies and fragmented text breaking conventional phrase structure rules, which prohibit discontinuous constituents.⁸ The increasing popularity of UD may also have a role in the prevalence of dependencies for web data, considering that 20 out of the 23 dependency treebanks are based on the UD scheme. Although not all of these corpora have been released in the official UD repository, and some of them do not strictly comply with the latest format specifications, the large number of UD resources, as well as their occasional divergences, highlight the need to converge on a single syntactic annotation framework for UGC within UD, to allow for a better degree of comparability across the resources and arrive at tested best practices.

In the next section, we provide an analysis of the guidelines of the surveyed treebanks, highlighting their similarities and differences, and a preliminary classification of the phenomena to be dealt with in UGC data from social media and the web with respect to the standard grammar framework for each language.

3.4 Annotation comparison

To explore the similarities and divergences among the resources summarized in Table 1, we carried out a comparative analysis of recurring annotation choices, taking into account a number of issues whose classification was partially inspired by the list of topics from the Special Track on the Syntactic Analysis of Non-Canonical Language (SPMRL-SANCL 2014).⁹ These issues include:

- sentential unit of analysis, i.e. whether the relevant unit for syntactic analysis is defined by typical sentence boundaries or other criteria
- tokenization, i.e. how complex cases of multi-word tokens on the one hand and separated tokens on the other are treated
- domain-specific features, such as hashtags, at-mentions, pictograms and other meta-language tokens.

The information on how such phenomena have been dealt with was gathered mostly from the reference papers cited in Table 1, and, whenever possible, by searching for the given phenomena within the resources themselves.

⁸ On the other hand, there are also a number of constituency-based annotation schemes that allow discontinuities, for example, NEGRA and TIGER for German.

⁹ <http://www.spmrl.org/sancl-posters2014.html>.

3.4.1 Sentential unit of analysis

Sentence segmentation in written text from traditional sources such as newspapers, books or scientific articles is usually defined by the authors through the use of punctuation. While it is usually treated as a more-or-less solved problem, Read et al. (2012a) showed in their overview of sentence boundary detection that performance can be significantly worse on text other than news. Recent work on sentence boundary detection in the financial and legal domains (Azzi et al., 2019; Sanchez, 2019) underscores that this assessment still applies.

The problem of segmentation is much more salient in the realm of spoken language transcription. On the one hand, there has been long-standing discussion on how and whether the notion of sentence applies at all.¹⁰ On the other hand, diverse annotation experiments have suggested that sentence segmentation cannot be done perfectly by humans and that its difficulty varies across text types (Stevenson & Gaizauskas, 2000; Westpfahl & Gorisch, 2018). Among UD corpora, the spoken French treebank is a conversion of the Rhapsodie treebank (Lacheret et al., 2014) and accordingly inherits its approach to segmentation based on the Aix school (Blanche-Benveniste et al., 1990). The spoken Slovenian UD treebank by Dobrovoljc et al. (2017) inherits its segmentation criteria from the GOS corpus from which it was sampled (Verdonik et al., 2013). They are distinct from the spoken French UD treebank's and recognize segments and turns. The recent segmentation criteria for spoken German (Westpfahl & Gorisch, 2018) are different yet again.

While the UGC data from social media that we are concerned with is written, it is frequently not well punctuated.¹¹ Often, punctuation marks may be missing, misapplied relative to the norms of written language, or used for other communicative needs altogether (e.g. emoticons such as -l, or emoticons simultaneously serving as closing brackets, etc.). In some cases, no punctuation is used whatsoever, as in Example 1 (the non-standard translation and spelling approximates the lack of punctuation in the original German text).

- (1) *Haben Menschen eigentlich nichts besseres zu tun als Suzie Grime zu haten ja einige Aktionen sind ehrenlos ich habs verstanden*
 'Don't people have anything better to do than to hate on Suzie Grime yes some things people do are a disgrace I gettit'

Against this background, it is a non-trivial task to segment social media text manually, let alone automatically. The research on spoken language segmentation also provides no widely agreed-upon applicable model. Given that many social media posts by private users tend to consist of sequences of short phrases, clauses

¹⁰ For an overview and references we refer the reader to Pietrandrea et al. (2014).

¹¹ This is not to say that there is no conventional, well-punctuated data on social media, or that sentence segmentation for other domains is trivial. For instance, many corporations and institutions employ social media managers who adhere to common editing standards. Conversely, some sentence boundaries in canonical written language are also ambiguous, e.g. in headings, tables and captions.

and fragments, it is understandable that many Twitter resources consider the entire tweet as a basic unit—though for other, longer sources, such as Reddit, using entire posts as utterances by analogy is not feasible. Further, certain types of annotations make retaining tweets as single segments more conducive. For instance, TWRO analyzed the syntactic/semantic relationships and ironic triggers across different sentences, which was more practical with tweets kept intact. In addition, annotation of inter-sentential CS (see Sect. 4) can be considered more appropriate at the tweet level. Finally, keeping tweets as single units in some treebanks saves the effort needed to develop, maintain, adapt or do post-processing on an automatic sentence segmenter.¹²

On the other hand, there are counterbalancing considerations that motivate performing medium-independent segmentation on UGC data, among these a possible overuse of syntactic relations that define side-by-side (or run-on) sentences (e.g. *parataxis* in UD); second, as mentioned previously, at least for some UGC data collections (e.g. blog posts), punctuation is found frequently enough and can be used. Third, given that Twitter doubled its character limit for posts from 140 to 280 at the end of 2017, treating tweets as single utterances might pose a usability problem for manual annotation. Fourth, some datasets, such as GUM, are multi-genre and include UGC next to canonical written text and spoken data, motivating a convergence of syntactic treatment of sentence boundaries. And finally, for NLP tools trained on multiple genres and for transfer learning, inconsistent sentence spans are likely to reduce segmentation and parsing accuracy.

Due to these considerations, *tweeDe* manually segmented tweets into sentences while introducing an ID system that enables reconstruction of complete posts, if needed. Similarly, GUM uses syntactic utterance level annotations of user IDs and addressee IDs to indicate the post-tree structure in Reddit forum posts. The CoNLL-U format used in the UD project provides the means to implement these kinds of solutions in a straightforward manner, using utterance level comment annotations, which are serialized together with each syntax tree. *tweeDe*, however, still features the use of the *parataxis* relation within a single utterance for juxtaposed clauses that are not separated by punctuation, even when they form multiple complete sentences, similar to the analysis one would find in newspaper treebanks.

For other cases authors have introduced additional conventions to cover special constructs occurring in social media. For instance, in some treebanks (sequences of) hashtags and URLs are separated out into ‘sentences’ of their own whenever they occur at the beginning or end of a tweet and do not have any syntactic function.

A third option besides not segmenting and segmenting manually is, of course, to segment automatically. In the spirit of maintaining a real-world scenario, Frb splits their forum data into sentences using NLTK (Bird & Loper, 2004), with no post-

¹² A segmenter could nevertheless be necessary e.g. if the next step is using a parser trained on sentence-split data.

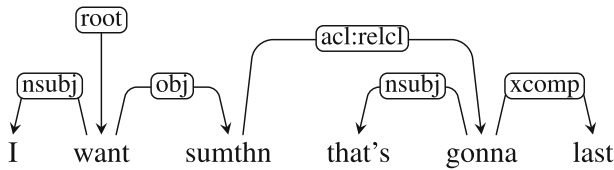


Fig. 1 Example of unsplit contraction from the TAAE treebank

corrections. Accordingly, the resource contains instances where multiple grammatical sentences are merged into one sentence due to punctuation errors such as a comma being used instead of a full stop, as in Example 2. Conversely, there are cases where a single sentence is split over multiple lines, resulting in multiple sentences (Example 3) that are not rejoined.

- (2) Combobox will start, When it is scanning don't move the mouse cursor inside the box, can cause freezing. (from Forebank)
- (3) I'm sure the devs.
can give you more details on this (from Forebank)

3.4.2 Tokenization

Tokenization problems in informal text include a wide range of cases which can sometimes require a non-trivial mapping effort to identify the correspondence between syntactic words and tokens. We may thus find multiple words that are merged into a single token, as in contractions¹³ (Example 4, which is also frequent in spoken English and can also be found in literary texts but not in newswire or academic writing) and initialisms such as the Italian example in (5), or, conversely, a single syntactic word split up into more than one token (6–7 below).

- (4) gonna ↔ going to
- (5) *tvb* ↔ *ti voglio tanto bene*
'I love you so much'

We observed a number of different tokenization strategies adopted to deal with those cases but most of the time the preferred solution seemed to involve their decomposition (Twb2, xUGC, tweeDe, FSMB, EWT,¹⁴ GUM), although a few

¹³ In this context we take into consideration only the cases encountered in informal/noisy texts, and not the traditional contractions typically present even in each standard language (such as English 'don't', the preposition-article contractions in French and German, or the verb-clitic contractions in Italian and German).

¹⁴ In Twb2 and EWT, however, some examples of phrasal contractions have been found that were not decomposed.

inconsistencies are found in the resulting lemmatization. Consider the contraction in Example 4. Twb2 reproduces the same lemma as the word form for both tokens ('gonna' → 'gon na'), while EWT and GUM instead use its normalized counterpart ('gonna' → 'go to').

Alternatively, these contractions might be either decomposed and also normalized by mapping their components onto their standard form, i.e. using 'go' and 'to' as the normalized word forms and lemmas of a multi-token unit¹⁵ 'gonna' (DWT, ITU,¹⁶ MNo), or rather left completely unsplit as a single token (and lemma) 'gonna' (TAAE, TWRO, Twb, Pst).

How these cases are annotated syntactically is not always specified in the respective papers, but the general principle seems to be that when contractions are split, the annotation is based on the normalized tokenization (Twb2, xUGC, ITU, FSMB, EWT, GUM), while when they are left unsplit, annotation is according to the edges connecting words within the phrase's subgraph (TAAE, Pst). According to this principle, Example 4 would thus be annotated based on the main role played by the verb 'go', as shown in Fig. 1.

As stated above, acronyms and initialisms may also pose a problem for tokenization, but in this case, there seems to be a higher consensus in not splitting them up into individual components, especially where an acronym is established and can be assigned a grammatical function without splitting, e.g. 'TL;DR' (too long; didn't read) is left as a single token in GUM, with the reasoning that the form is conventional and likely to be pronounced as the acronym even when read aloud.

When the opposite strategy is used, that of multi-token units, the preferable option, in most cases, is not to merge the separate tokens (TAAE, TWRO, Frb, Twb2, Pst, FSMB, EWT). As a result, one token—either the first (TAAE, TWRO, Frb, Twb2, Pst, EWT, GUM) or the last one (FSMB)—is often promoted to represent the main element of the multi-token unit. This kind of "promotion" strategy, when put into practice, could actually mean very different things. In Frb, a distinction is drawn between morphological splits (Example 6) and simple spelling errors (Example 7):

- (6) he should buy **anti vir** programs ↔ antivir (from Forebank)
 (7) **i t** keeps causing <ProductName> to lock up ...↔ it (from Forebank)

In the first case, both tokens are tagged based on the corresponding category of the intended word, i.e. as a NOUN (since 'antivirus' is a noun). In the second one, 'i t' is the erroneous split of the pronoun 'it'; the first token 'i' is here considered as a spelling error, while the second token 't' as an extraneous token. As opposed to the

¹⁵ These units, sometimes called super-tokens, have a special representation grouping several underlying tokens in the CoNLL-U format and are used to represent phenomena such as preposition-article fusion and other contractions.

¹⁶ In ITU, however, institutionalized and formal abbreviations are not expanded.

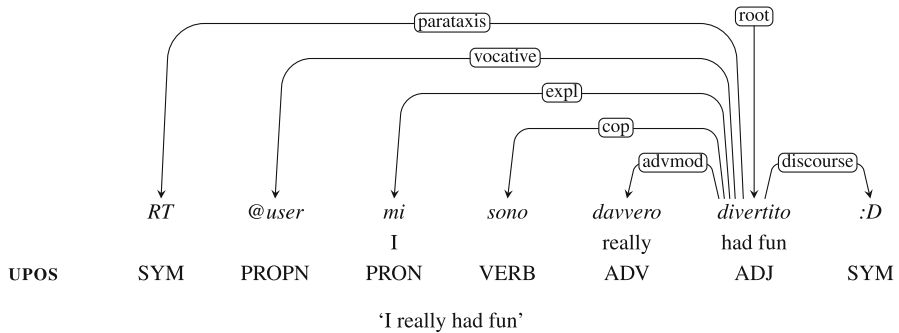
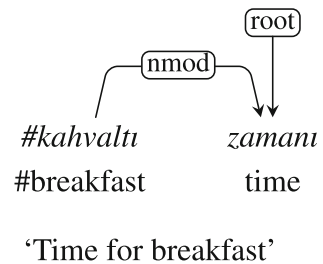


Fig. 2 Italian example of a ‘RT’ token and a ‘@’ user mention from Twitter, 2013

Fig. 3 Turkish example of a syntactically incorporated hashtag from Twitter, 2019



principles above, in LDF the effort to address tokenization issues resulted in modifying the original text, thus merging the wrongly split word (Example 8), or, conversely, splitting two words that appear joined (Example 9).

- (8) *Sistema de gener ación de bitcoins* ↔ *generación* (from LAS-DisFo)
 ‘System for **gener ating** bitcoins’ ↔ generating
- (9) *Esto estan de incrédulos* ↔ *es tan* (from LAS-DisFo)
 ‘This **isso** like incredulous people’ ↔ is so

In the remaining resources, neither explicit information nor regular/consistent patterns have been found concerning the morpho-syntactic treatment of these units. For their syntactic annotation in dependency grammar frameworks, common practice is to attach all remaining tokens to the one that has been promoted to head status. In UD corpora, the second (and subsequent) tokens in such instances are connected to the first token, and labeled with the special *goeswith* relation, which indicates superfluous whitespace between parts of an otherwise single token word.

Finally, a distinctive tokenization strategy is adopted in ATDT with respect to at-mentions, in which the ‘@’ symbol is always split apart from the username, whereas other corpora retain the unsplit username along with the ‘@’ symbol.

While we strongly urge annotators and maintainers of new resources to adopt the more common strategies outlined above, for many specific tokenization issues, as well as other issues below, it may ultimately be impossible to provide generally

valid, necessary and sufficient criteria for deciding one way or the other. What is important in such cases is to document decision (ideally in the publicly available UD language specific and universal documentation, as appropriate), and if possible to implement automatic validation tools which promote consistency by ensuring that comparable cases across a corpus or set of language corpora are annotated in the same way.

3.4.3 Other domain-specific issues

This category includes phenomena typical for social media text in general and for Twitter in particular, given that many of the treebanks in this overview contain tweets. Examples include hashtags, at-mentions, emoticons and emojis, retweet markers and URLs. These items operate on a meta-language level and are useful for communicating on a social media platform, e.g. for addressing individual users or for adding a semantic tag to a tweet that helps put short messages into context. On the syntactic level, these tokens are usually not integrated, as illustrated for Italian in Example 10 and in its syntactic tree in Fig. 2.

- (10) *RT @user mi sono davvero divertito :D*
 'RT @user I really had fun :D'

(adapted from Twitter, 2013)

It is, however, also possible for those tokens to play a syntactic role in the tweet, as shown in the Turkish example in Fig. 3.

In the different treebanks, we observe a very heterogeneous treatment of these meta-language tokens concerning their morpho-syntactic annotation. Hashtags and at-mentions, for example, are sometimes treated as nouns (DWT, ITU), as symbols (TWRO, Pst), or as elements not classifiable according to existing POS categories, or, more generically, as 'other' (Twb2, HSE,LDF, TwIr, Taiga).

Some resources adopt different strategies that do not fit into this pattern: in tweeDe and GUM, for example, at-mentions referring to user names are always considered proper nouns while hashtags are tagged according to their respective part-of-speech. Multi-word hashtags are annotated as 'other' in tweeDe (e.g. *#WirSindHandball* 'We are handball'), but as proper nouns in GUM (*#IStandWithAhmed*). In Twb2, a different POS tag is assigned to at-mentions when they are used in retweets.

Similarly to hashtags and at-mentions, links can either be annotated as symbols (TWRO, Pst, TwIr), nouns (W2.0, ITU, FSMB), proper nouns (GUM), or 'other' (tweeDe, EWT, HSE, Taiga).¹⁷ Emoticons and emojis, on the other hand, are mostly classified as symbols, less often as interjections (DWT, FSMB), and in one case as a punctuation mark sub-type (ITU). Retweet markers (RT) are considered as either nouns (DWT, Pst), symbols (TwIr) or 'other' (Twb2¹⁸). On the syntactic level, these

¹⁷ In EWT the universal POS tag 'X' is used, corresponding to the concept of 'other', but a special native tag is also applied concurrently: 'ADD' (for address), which can then be used to find URLs in particular.

¹⁸ Except when they are considered an abbreviation of the verb "retweet", in which case they are annotated accordingly.

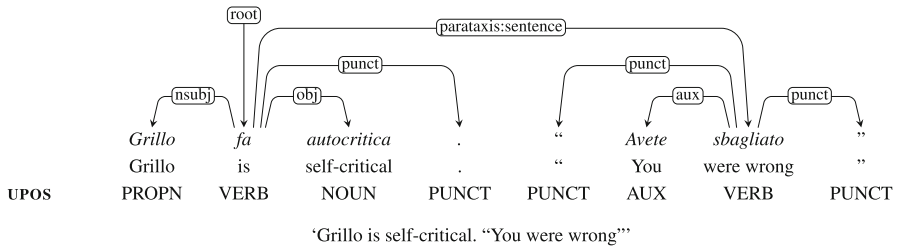


Fig. 4 Italian example from Twitter, 2015, of multiple sentential units in a tweet

meta-tokens are usually attached to the main predicate, but we also observe other solutions. As stated above, in *tweeDe* hashtags and URLs at the beginning or end of a tweet form their own sentential units, while in *Twb*, they are not included in the syntactic analysis.

Finally, in cases where meta-tokens are syntactically integrated, the recurring practice is to annotate them according to their role (TAAE, TWRO, DWT, *Twb2*, *tweeDe*, Pst, GUM). ATDT is unique in that it does not distinguish between meta-tokens at the beginning or end of the tweet and those that are syntactically integrated in the tweet, but instead always assigns a grammatical function to these tokens.

Based on the practices briefly outlined in this section, in the next section, we define an extended inventory of possible annotation issues, some of which occur in only one or a few resources, and propose a set of tentative guidelines for their proper representation within the UD framework, also summarized in “[Appendix](#)”.

4 Towards a unified representation

As is widely known, the project of UD aims at developing cross-linguistically consistent treebank annotation for many languages. Given the increasing number of treebanks featuring user-generated content, and a lack of guidelines specifically tailored for this textual genre, in this section we propose a unified approach to annotate the issues that might arise from such texts.

In the following paragraphs we will address the challenges outlined in Sect. 3.4 along with other phenomena that are often found in user-generated text, such as CS and disfluencies. As always when weighing different options, key considerations straddle a balance between maximal annotation consistency, time requirements in producing sizable treebanks, potential cognitive overload for annotators, theoretical soundness, and universal applicability across languages.

The suggestions we propose throughout this section were discussed among multiple authors—who are themselves UD contributors—taking different language scenarios into account. We thus propose a list of recommendations for UGC annotation in UD and with this proposal we look forward to receiving feedback from the community to further enhance this collaborative effort towards a unified representation of UGC in UD.

4.1 Sentential unit of analysis

In the interest of maintaining compatibility with treebanks of standard written language, we propose splitting UGC data into sentential units to the extent to which it is possible and keeping token sequences undivided only when no clear segmentation is possible. To facilitate tweet-wise annotation if desired, a subtyped parataxis label, such as `parataxis:sentence` in Fig. 4, could be used temporarily during annotation. Since some relation label will be needed to connect multiple sentential units within a tweet no matter what, this recommendation is mainly meant to help with later processing or comparison with other data sets, serving as a pointer to identify where the tweet could be split into sentences and distinguishing such junctures from other types of parataxis.¹⁹

4.2 Tokenization

As shown in the examples in Table 2, user-generated content can include a number of lexical and orthographic variants whose presence have repercussions with respect to segmentation and choices presented to annotators. The basic principle adopted in UD, for which morphological and syntactic annotation is only defined at the word level (universaldependencies.org, 2019g), can sometimes clash with the complexity of these cases, which has also been a matter of debate within the UD community.²⁰

- *Contractions* One particularly challenging issue for annotation decisions related to tokenization is contraction, i.e. when multiple linguistic tokens are contracted to form a single orthographic token (or into fewer tokens than the linguistic content would suggest). It is important to note the different types of contractions that can appear in UGC. For the cases of (i) conventionalized contractions, such as ‘don’t’ and (ii) erroneously merged words (e.g. ‘mergedwords’), it is usually easy to identify the morpheme boundary split point. In these cases, we recommend that annotators split the contraction into its component tokens, in keeping with the UD guidelines (universaldependencies.org, 2019a) already in place to deal with occurrences of such merging in standard text.

However, for instances of (iii) deliberate informal contractions, such as colloquial abbreviations and initialisms (e.g. EN ‘gonna’, ‘wanna’, ‘idk’ (‘I don’t know’)) or shorthand forms (FR *nimp*, ‘whatever’), standardized criteria are mostly inadequate, or at least insufficient to cover the whole host of possible phenomena. This is due to the ever-changing and often ambiguous nature of user-generated text, i.e. many of the colloquialisms common in UGC are also

¹⁹ Conversely, we may want to indicate that multiple sentences come from a single tweet; although the CoNLL-U does not allow relations between sentences, sentence level comments or sentence identifiers can indicate that two sentences belong to the same tweet, and in many cases tweets will correspond to individual documents in the corpus, in which case it would be clear that sentences with the same document ID belong together.

²⁰ <https://github.com/UniversalDependencies/docs/issues/641>.

increasingly conventionalized in the standard language (e.g. ‘gonna’, which is frequent in print in certain registers, and ubiquitous in spoken language), while others may fall out of use entirely. Thus, whether or not a term is considered a conventional contraction is dependent on the time of annotation, and can also be largely subjective. It is also worth noting that increased annotator effort is required if informal contractions are split, as further challenges may be introduced with regard to lemmatization and capturing information for other downstream tasks. This can create a significant overhead in treebank development. For this reason, we advise annotators to adopt an individual approach that takes both treebanking consistency and feasibility into account.

Annotators may wish to consider whether an informal contraction has reached a non-compositional status (e.g. TL;DR, LOL, WTF, idk, etc. in English), and whether it functions solely as a discourse marker or actually bears a semantic and syntactic role within the sentence which is equivalent to its potential expansion (for example, TL;DR, which means ‘too long; didn’t read’, is often used in online content creation to provide readers with a shortened summary version of a text). In cases where decomposition of a conventionalized expression is avoided, but the whole function of the phrase is equivalent, our suggested approach is in line with the principle proposed in Blodgett et al. (2018) where annotation is carried out according to the root of the subtree of the original phrase. In the example below, the conventionalized form ‘idk’ (sometimes spelled out when read aloud) is actually used in the place of a matrix verb and is therefore labeled as `root`, taking a complement clause argument `ccomp` (Fig. 5).

Some advantages of leaving deliberate, informal contractions unsplit are that less annotation effort would be required, consistency within the treebank would be easier to maintain, and fewer decisions would be left to the discretion of the annotator (such as the intention of the user and the compositionality of the term in specific instances). Additionally, treebank developers may consider this approach to be a more descriptive rather than prescriptive representation of ‘noise’ in the data.

By contrast, the benefits of splitting such tokens are that it can be considered a cleaner approach as it will result in fewer ambiguous tokens and it will also allow for more fine-grained detail in the annotation, as well as comparability with resources in which equivalent split forms appear.

- *Unconventional use of punctuation:* We recommend that unconventional use of punctuation in the form of pictograms `:-)` or strings of repeated punctuation marks `!!!!!!` be annotated as a single token rather than being split. Further, we suggest that strings of emoticons be split so that each individual emoticon is considered an individual token, such as `:):)` → `:)` + `:)` (similar to other sequences of tokens spelled without intervening spaces). As a guiding principle we advocate not splitting only in cases where there is a reason to believe that multiple glyphs amount to a morphosyntactic word together: this is not the case

for repeated exclamation points, whereas multiple emoticons or emojis can be considered to express word level meanings (cf. Sect. 4.6). An exception would be cases of rebus, such as 🌟 used to spell ‘Starbucks’ (a single token, tagged `upos=PROPN`).

- *Over-splitting* Another tokenization issue relates to the treatment of incorrectly split words. The UD guidelines already advise the use of the `goeswith` relation in cases of erroneously split words from badly edited texts (e.g. EN ‘be tween’ → ‘between’, TR *gele bilirim* ‘come I can’ → *gelebilirim* ‘I can come’). This means that the split tokens are not merged, but information on their full form is captured nonetheless, while tokens containing whitespace are avoided. In line with the specifications for erroneously split words (universaldependencies.org, 2019a)—be it due to formatting, a typo or intentional splitting—we suggest to promote the first part of the word to the role of syntactic head and apply left-right attachment, regardless of any potential morphological analysis (i.e. the head of ‘be tween’ is ‘be’). The initial token would also bear the lemma, the POS tag and the morphological features of the entire word, while the remaining split parts would only be POS-tagged as X, and leaving the lemma and features unspecified (by convention ‘_’). For instance in the Turkish example in Fig. 6, `Number` and `Person` features, as well as others, are expressed in the *bilirim* ‘I can’ part of the over-split word, but annotated in the FEATS column of the first part.

4.3 Lemmatization

With respect to the lemmatization of user-generated text, we note that the UD guidelines, specifically those referring to morphology (universaldependencies.org, 2019e), can often be applied in a straightforward manner. However, certain phenomena common to UGC can complicate this task. In the cases of contraction, over-splitting and unconventional punctuation, lemmatization will depend on the tokenization approach chosen as discussed in the previous section.

Unconventional uses of punctuation include punctuation reduplication, seemingly random strings of punctuation marks and pictograms or emoticons created using punctuation marks. Punctuation reduplication can be lemmatized by normalizing where a pattern is observed (`?!?!? → ?!`), otherwise the lemma should

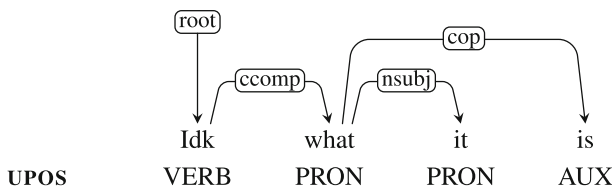


Fig. 5 Example of an unsplit contraction on Twitter, 2020

match the surface form (e.g. *!!!!!!!* → *!!!!!!!*). We also recommend that emoticons and pictograms not be normalized (*:)]* → *:)]*), as any attempt of defining a finite set of ‘conventional’ emoticon lemmas would result in a somewhat arbitrary and incomplete list. When lemmatizing neologisms or non-standard vocabulary such as transliterations, we recommend that any inflection be removed in the lemma column (TR *taymlaynda* → *taymlayn*, ‘(in) timeline’). If the token is uninflected, we suggest the lemma retain the surface form without any normalization.

4.4 Features column

UD prescribes the use of the features column to list information about the morphological features of the surface form of a word. We suggest that the feature `Abbr=Yes` be used for abbreviations such as acronyms, initialisms, character omissions, and contractions (see Fig. 8 for an example). Annotators may also choose to include the feature `Style=X`, employed by some treebanks to describe various aspects of linguistic style such as [`Coll`: colloquial, `Expr`: expressive, `Vrnc`: vernacular, `Slng`: slang]²¹ (Figs. 7, 9). Among UGC UD treebanks, only TDT currently uses this feature.

Another useful feature prescribed by UD is `Typo=Yes` (see Fig. 8) for seemingly accidental deviations from conventional spelling or grammar (used e.g. in GUM, EWT). The feature `Foreign=Yes` will be further discussed in Sect. 4.7 on CS.

4.5 MISC column

At present, aside from capturing instances of spelling variations arising from abbreviation and typos, UD prescribes no mechanism for describing the *nature* of spelling variations. For this reason, we suggest the addition of a new attribute to the UD scheme to denote the more general case of non-canonical language and to more accurately describe the nature of phenomena such as those exemplified in Table 2 (see “Appendix”). This additional attribute `NonCan=X` would be annotated in the MISC column with the following possible values (repeated for each affected token, multiple values can be joined by comma in alphabetical order as per the CoNLL-U standard, see Fig. 9):

[`AutoC`: autocorrection, `CharOm`: character omission, `Cont`: contraction, `Neo`: neologism, `OS`: over-splitting, `Phon`: phonetization, `PunctVar`: punctuation variation, `SpellVar`: spelling variation, `Stretch`: graphemic stretching, `Transl`: transliteration, `Trunc`: truncation].

Additionally, the MISC column may be used to list values corresponding to a hypothetical standard or full form of the word, i.e. the attributes `CorrectForm=X`, `FullForm=X`, `CorrectSpaceAfter=Yes` may be useful in the cases of non-

²¹ A description of the values the `Style=X` feature may take in UD is provided in the official guidelines (universaldependencies.org, 2021)

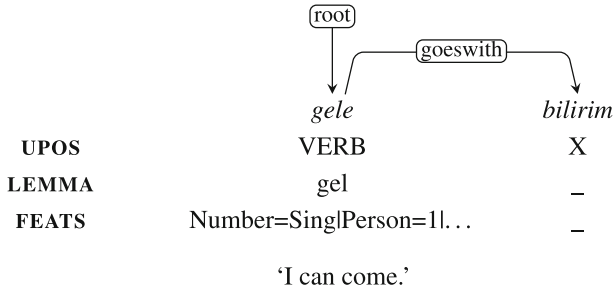
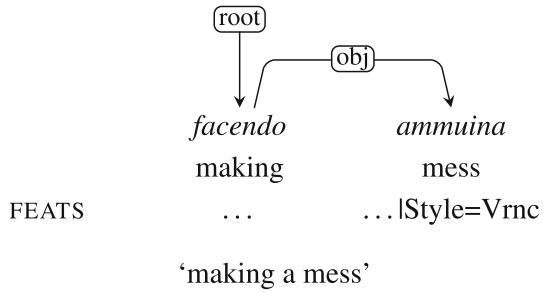


Fig. 6 Turkish example of over-splitting from Twitter, 2020

Fig. 7 Italian example in which the token 'ammuina' (derived from the neapolitan expression *facite ammuina* 'make a mess') exemplifies the use of vernacular expression as annotated in the FEATS column; from Twitter, 2012



canonical language, abbreviations and incorrectly merged words respectively (Fig. 8).²²

The attribute `Lang=x` will be further discussed in Sect. 4.7 on CS, while additional example annotations for `NonCan=X` are included in “Appendix” (see Tables 5, 6 and 7).

4.6 Domain-specific issues

UGC includes many words and symbols with domain-specific meanings. We recommend treating the various groups as follows:

- *Hashtags* are to be labeled with the tag of the actual token without the hashtag sign. If a hashtag comprises multiple words, it should be kept untokenized and the POS tag is the POS tag of the head word. e.g., *#behappy/ADJ*. Syntactically integrated hashtags should bear their standard dependencies. Classificatory hashtags at the end of tweets are to be attached to the root with the dependency subtype `parataxis:hashtag` as per the English example in Fig. 10.

²² At the same time we acknowledge a long strand of research on formulating target hypotheses for non-native and other forms of non-canonical language, which shows that establishing the ‘correct’ or intended form is often a matter of debate requiring detailed guidelines for doubtful cases. See Reznicek et al. (2013) for discussion.

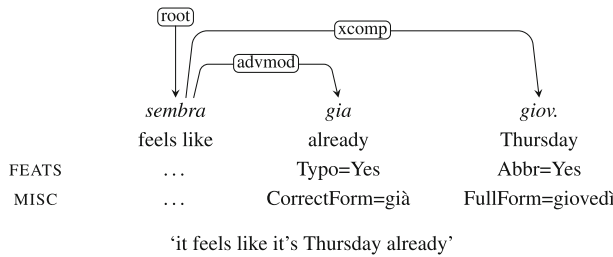


Fig. 8 Italian example of both typo and abbreviation with their corresponding correct/full form in the misc column. Adapted from Twitter, 2012

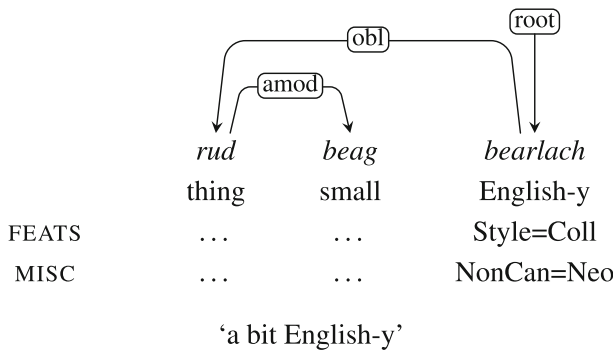


Fig. 9 Irish tweet in which the token ‘bearlach’ (derived from Béarla ‘English language’) exemplifies a colloquialism and neologism as annotated in the FEATS and MISC columns respectively. Adapted from Twitter, 2013

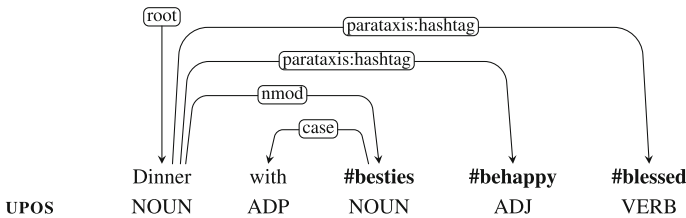


Fig. 10 English example of hashtag usage from Twitter, 2018

- *At-mentions* to be labelled as PROP. Their syntactic treatment is similar to hashtags: when in context they bear the actual syntactic role (see Fig. 11 for a Turkish example), otherwise they should be dependent on the main predicate with the *vocative* label as per the Irish example in Fig. 12.
- *URLs* are to be tagged as SYM as per UD guidelines. They are often appended at the end of the tweet without bearing any syntactic function. Throughout our explored corpora, those URLs are diversely annotated, without an obvious consensus emerging: *parataxis:url* vs *discourse:context* vs *dep*. In

cases where they are syntactically integrated in the sentence, we recommend that they be given their syntactically warranted dependency relation, as per Fig. 13. We favor using `parataxis:url` for non-syntactically integrated URLs (as in Fig. 14), or plain `parataxis` if the subtype is not used in the corpus, since by default we assume that an unintegrated URL has a status similar to a separate utterance standing within the same orthographic sentence (as opposed to emoji adding flavor to a sentence).

- *Pictograms* are often used at the end of the tweets as discourse markers. In such cases they should be POS-tagged as SYM and attached to the root with the `discourse` relation. But there are also cases where pictograms function as instances of word classes other than SYM. Thus, deviating from the UD guidelines' invariant treatment of emojis as SYM, we treat the heart emoji as a VERB in the two following examples (11–12). We believe this is more in line with UD's basic criteria for POS assignment, namely a form's occurrence in particular syntactic environments and its inflectional properties.²³ The non-SYM treatment should also be adopted for other symbols such as the dollar sign, e.g. in “How much more \$\$ does the Ice Sports Association need to raise for the Scheels IcePlex? I'll tell you now on @keloland news. ”

(11) ❤️ed it

(from Twitter, 2020)

(12) Thank you 4 All U do & ❤️ing dogs

(from Twitter, 2020)

These cases are to be annotated with the lemma, UPOS tag and dependency relation of the word they substitute. The French example in Fig. 15 demonstrates both cases.²⁴ The morphological features should reflect the intended meaning. Thus, in example 13 the feature for the pictogram/verb should be `Person=3` even though the form canonically is a non-third-person form.

(13) Go follow @IMAPCT_Zodiak he's a beast & he'll follow back. He ❤️ his followers. (from Twitter, 2020)

- *RTs* are originally used with at-mentions so that the Twitter interface interprets it as a retweet. In such cases, their UPOS should be SYM with a dependency label `parataxis` attached to the root.²⁵ However they are now more commonly used as an abbreviation for *retweet* within a tweet. The UPOS tag

²³ We take it that UD's invariant treatment of emojis (and other symbols) as SYM was chosen based on quite different data sources, in which cases like (11–12) simply did not feature. The invariant treatment of course also has the appeal of simplicity.

²⁴ An anonymous reviewer has asked why the coffee emoji is labeled ‘discourse’ rather than being treated as a lexical item (e.g. as a kind of noun modifier, apposition or dislocated). The annotation decision here is based on the annotator's judgment that the coffee emoji, unlike the heart, is not meant to be pronounced, and is therefore paradigmatic with typical emojis labeled discourse, such as smileys.

²⁵ An anonymous reviewer asked why we use `parataxis` here, assuming that that label was restricted to clauses. While such a restriction may be desirable, inspection of the UD guidelines shows that the label is also used, e.g. “to connect the parts of a news article byline”. Given the restricted label inventory of UD, there is no clearly better label to use and we therefore ask `parataxis` to also moonlight as a connecting device between certain instances of RT and their heads.

should be NOUN or VERB depending on its syntactic role and potentially its inflectional properties since we also find inflected forms, e.g. “Someday , I’ll get RTed by @jizziemcguire and it’ll be fucking awesome” or “This deserves endless RTs”. In these cases, the dependency relation depends on the functional role of the full form (see Figs. 16 and 17).

- *Markup* symbols (e.g. <, >, ++), if used as symbols that serve to delimit and organize phrases, as in the German example in Fig. 18, have the UPOS PUNCT and are attached to the head with punct.

4.7 Code-switching

As discussed in Sect. 3, capturing CS in tweets is an additional motivation for following a tweet-based unit of analysis (Çetinoğlu, 2016; Lynn & Scannell, 2019).

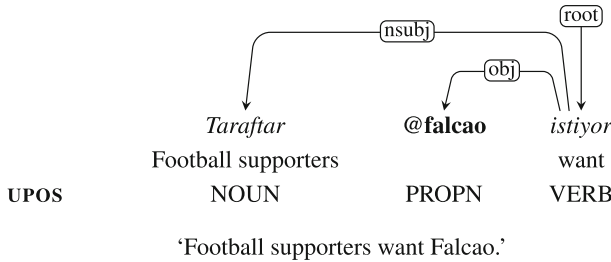


Fig. 11 Turkish example of a syntactically incorporated at-mention from Twitter, 2018

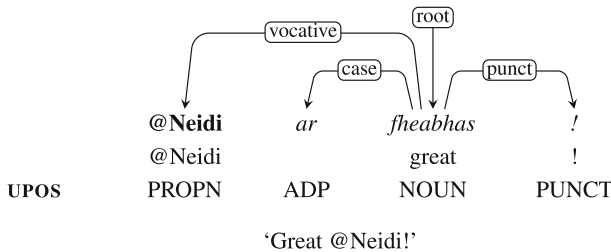


Fig. 12 Irish example of a vocative at-mention from Twitter, 2012

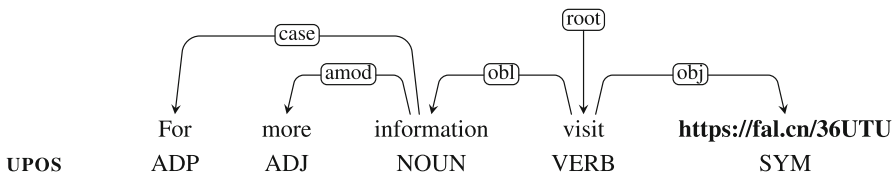


Fig. 13 English example of syntactically-integrated URL from Twitter, 2020

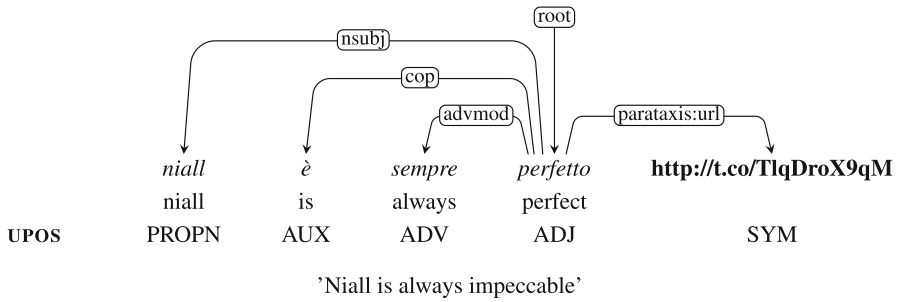


Fig. 14 Italian example of syntactically-unintegrated URL from Twitter, 2013

Code-switching—switching between languages—is an emerging topic of interest in NLP (Bhat et al., 2018; Solorio & Liu, 2008; Solorio et al., 2014) and as such should be captured in treebank data where possible. Code-switching can occur on a number of levels. Code-switching that occurs at the sentence or clause level is referred to as inter-sentential (INTER) switching as shown between English and Irish in Example 14, and German and Turkish in Example 15:

- (14) *Má tá AON Gaeilge agat, úsáid í! It's Irish Language Week.*
 If is ANY Irish at-you use it!
 'If you have ANY Irish, use it! It's Irish Language Week'. (from Twitter, 2014)
- (15) *@user Jedem das was er verdient. ;-) Yoksa Köln'den Almanca öğrenmeden mi döndün*
 To-each that what he deserves. Or Köln-from German learn-without Ques returned
 'Everyone gets what they deserve ;-) Or did you return from Cologne without learning German?'
 (from Twitter, 2014)

Inter-sentential switching can also be used to describe bilingual tweets where the switched text represents a translation of the previous segment: “Happy St Patrick’s Day! *La Fhéile Pádraig sona daoibh!*” This phenomenon is often seen in tweets of bi-/multi-lingual users.

Code-switching occurring within a clause or phrase is referred to as intra-sentential (INTRA) switching. Example 16 demonstrates intra-sentential switching between Italian and English:

- (16) *Le proposte per l'education di Confindustria*
 The proposals for the-education of
 'The proposals for the education by Confindustria' (adapted from TWITTIRÒ, 2014)

Word-level alternation (MIXED) describes the combination of morphemes from different languages or the use of inflection according to rules of one language in a word from another language. This is particularly evident in highly inflected or agglutinative languages. Example 17 shows the creation of a Turkish verb derived from the German noun *Kopie* ‘copy’.

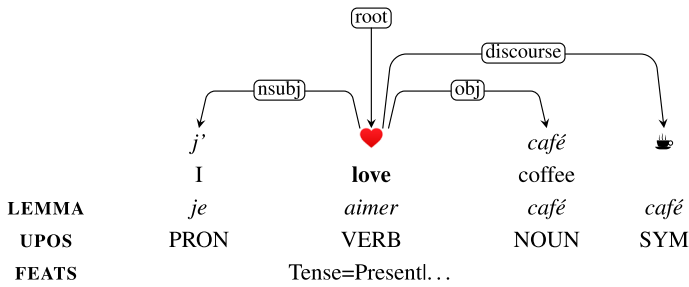


Fig. 15 French example of differing syntactic roles of pictograms from Twitter, 2013

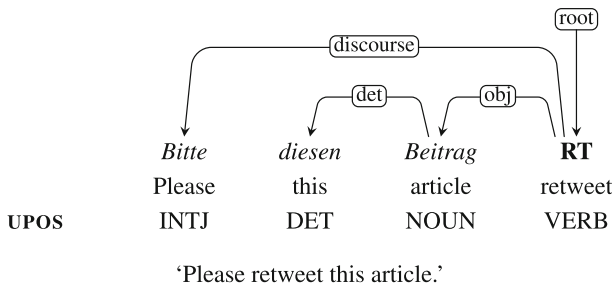


Fig. 16 German example of RT as a verb from Twitter, 2019

- (17) *Adamın 3-4 biyografisi var Kopyelenip yapıştirılmış.*
 Guy's biography exists copied pasted.
 'The guy has 3-4 biographies copied and pasted.' (adapted from Twitter, 2016)

While borrowed words can often become adopted into a language over time (e.g. 'cool' is used worldwide), when a word is still regarded as foreign in the context of code-switching, the suggested UPOS is the switched token's POS—if known or meaningful—otherwise X is used (universaldependencies.org, 2019d). The morphological feature *Foreign=Yes* should be used, and we also suggest that the language of code-switched text is captured in the MISC column, along with an indication of the code-switching type. As such, in Example 16, 'education' would have the MISC values of *CSType=INTRA|Lang=en*.²⁶

In terms of syntactic annotation, the UD guideline recommends that the *flat* or *flat:foreign* label is used to attach all words in a foreign string to the first token of that string (universaldependencies.org, 2019c). We recommend that this guideline is followed (for both inter-sentential and intra-sentential code-switching) when the grammar of the switched text is not known to annotators (see Fig. 19). Otherwise, we recommend applying the appropriate syntactic analysis for the switched language (see Fig. 20).

²⁶ In the case of the language identification feature, for the purpose of facilitating the UD validation tool, the value used should be the ISO code. However, in the case of MIXED code-switching where a combination of languages is in question, the UD developers should be advised.

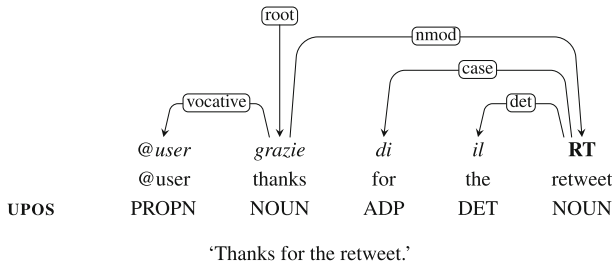


Fig. 17 Italian example of RT as a noun from Twitter, 2013

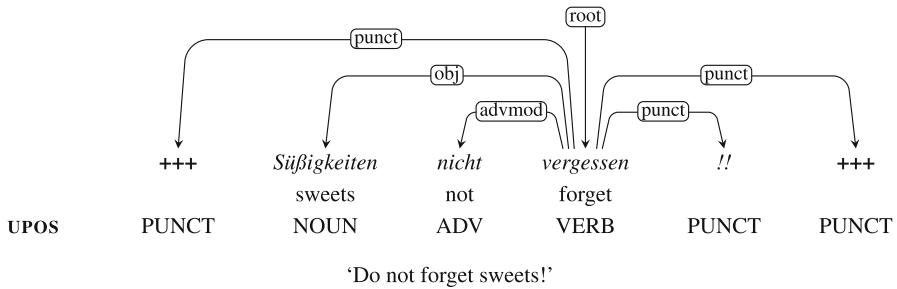


Fig. 18 German example of the use of markup symbols from Twitter, 2020

Lemmatization of code-switching tokens can prove difficult if a corpus contains multiple languages that annotators may not be familiar with. To enable more accurate cross-lingual studies, all switched tokens should be (consistently) lemmatized if the language is known to annotators and annotation is feasible within the constraints of a treebank’s development phase. Otherwise the surface form should be used and a `ProvisionalLemma=Yes` feature should be added to the `MISC` column, allowing for more comprehensive lemmatization at a later date.

4.8 Disfluencies

Similarly to spoken language, UGC often contains disfluencies such as repetitions, fillers or aborted sentences. This might be surprising, given that UGC does not pose the same pressure on cognitive processing that online spoken language production does. In UGC, however, what may seem to be a production error can in fact have a completely different function (Rehbein, 2015). Here, self-repair and hesitation markers are often used with humorous intent (Example 18 illustrates this for the use of hesitation markers and 19 for the use of self-repair).

- (18) *Kann man eigentlich bei übermäßigem Verzehr eine Schokoladenvergiftung bekommen? Ich, ähm, frage für einen Freund.*
 ‘Can you actually get chocolate poisoning if you eat too much? I’m, **uhm**, asking for a friend.’
 (from Twitter, 2016)
- (19) *Du hast den Apple Wahnsinn... äh, Spirit einfach noch nicht verstanden ;)*
 ‘You haven’t yet understood the Apple madness... **uh** spirit :)’
 (from Twitter, 2012)

Disfluencies pose a major challenge for syntactic analysis as they often result in an incomplete structure or in a tree where duplicate lexical fillers compete for the same functional slot. The case of self-repair, which is far more frequent in spontaneous spoken language, has been discussed in the context of UD treebanks for spoken language material (see Caron et al., 2019; Dobrovolic & Nivre, 2016; Lacheret et al., 2014; Leung et al., 2016; Øvrelid & Hohle, 2016; Tyers & Mishchenkova, 2020; Wong et al., 2017; Zeldes, 2017, amongst others) where solutions for syntactic analysis have been presented. The UD guidelines propose the use of the reparandum relation for disfluency repairs (universaldependencies.org, 2019f). This is illustrated in the German example in Fig. 21.

In this example, the tweet author starts writing the phrase *Das Wort zum Sonntag* ‘The word for Sunday’, a reference to a well-known German religious TV program featuring a brief homily, then abandons the word “Sunday” and repairs it to “Tuesday”. The disfluency marker “äh” (uh) is used to indicate the repair.

However, the treatment displayed above (Fig. 21) loses information whenever the reparandum does not carry the same grammatical function as the repair, as illustrated in Fig. 22 (left). In this German example from Twitter, the user plays with the homonymic forms of the noun *Hengst* (stallion) and the verb *hängst* (*hang*_{2.Ps.Sg}). The repair changes the grammatical function from *vocative* to *nsubj*, which cannot be encoded in the core UD schema. The missing information, however, could be easily added, based on the enhanced UD scheme (universaldependencies.org, 2019b), following the treatment of conjoined subjects and objects. Similarly, we could add an edge from the reparandum to the first word in the sentence that specifies the missing relation type (Fig. 22, right).²⁷

Other open questions concern the use of hesitation markers in UGC. We propose to consider them as multi-functional discourse structuring devices (Fischer, 2006) and annotate them as discourse markers. However, it remains unclear whether they should be attached to the root node as for repair markers, this would often result in non-projective trees. We therefore recommend to attach them to the reparandum (see Fig. 21). When no reparandum is present, for example when functioning as markers of humorous intent (Examples 18 and 19) or to mimic spontaneous speech (Example 20), we recommend to attach them to the root node (see Fig. 23).

²⁷ Please note that the enhanced dependencies do not provide such a treatment at the moment.

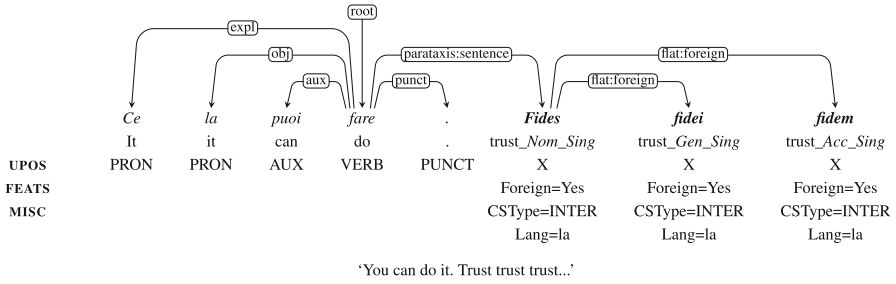


Fig. 19 Italian-Latin code-switching example tree. Adapted from POSTWITA, 2011

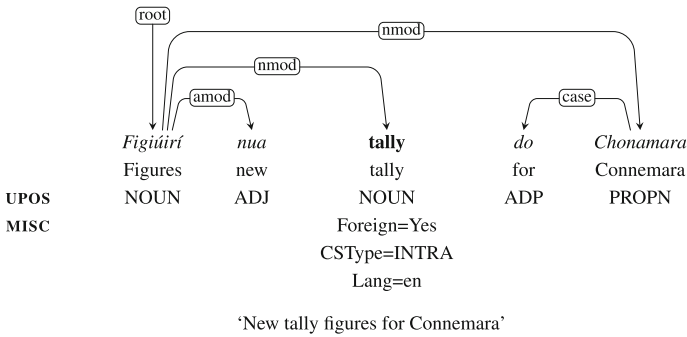


Fig. 20 Irish-English code-switching example from Twitter, 2014

(20) „Also ich glaube **ähm**...das es die schlimmste Krankheit ist...weil **ähm**“ Til Schweiger. Über #alzheimer
 ‘So I think **uhm**...that it is the worst disease...because **uhm**...’ Til Schweiger. On #alzheimer
 (from Twitter, 2014)

5 Discussion

In this final section, we discuss some open questions in which the nature of the phenomena described makes their encoding difficult by means of the current UD scheme.

5.1 Elliptical structures and missing elements

In constituency-based treebanks that contain canonical texts, such as the Penn Treebank (Marcus et al., 1993), the annotation of empty elements results from the need to keep traces of movement and long-distance dependencies, usually marked with trace tokens and co-indexing at the lexical level in addition to the actual nodes dominating such empty elements. The dependency syntax framework usually does

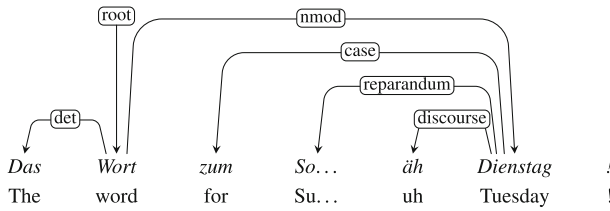


Fig. 21 Example tree for the use of disfluencies in UGC, illustrating the use of the *reparandum* relation (from Twitter, 2021)

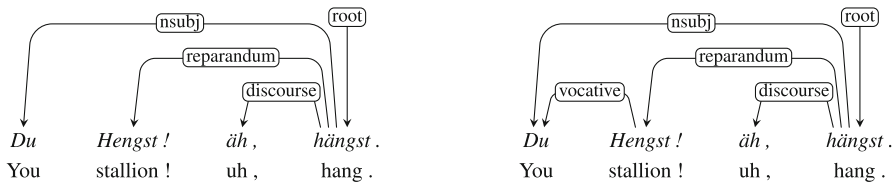


Fig. 22 Example tree for a *reparandum* relation where the core UD annotations (left) lose information as the repair changes the grammatical function of the “*Du*” (You) from *vocative* to *nsubj*. In the right tree, the missing information is added by means of an additional edge

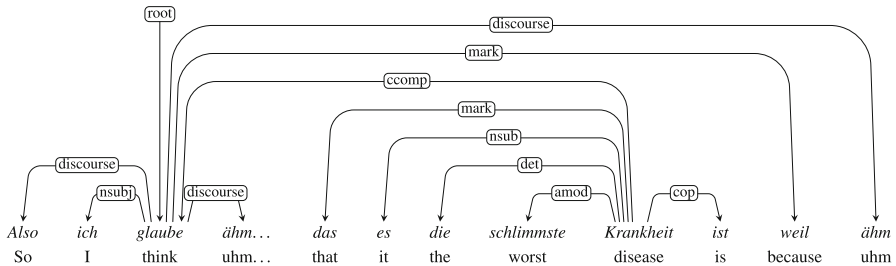


Fig. 23 Example tree for the use of disfluencies in UGC, illustrating the annotation of hesitation markers in an aborted utterance (from Twitter, 2014)

not use such devices as these syntactic phenomena can be represented with crossing branches resulting in non-projective trees.

In the specific case of gapping coordination, which can be analyzed as the results of the deletion of a verbal predicate (e.g. John loves_i Mary and Paul (e_i) Virginia), the subject of the right-hand side conjunct (Paul) is promoted to the head position and is attached to the verb of the left-hand side conjunct (loves) via the *conj* relation. The object (Virginia) of the elided verb is attached to the subject (Paul) via the *orphan* relation (Schuster et al., 2017).

Even though the Enhanced UD scheme proposes to include a *ghost*-token (Schuster & Manning, 2016a) which will be the actual governor of the right hand-side conjuncts, nothing is prescribed regarding the treatment of ellipsis without an antecedent. Given the contextual nature of most UGC sources and their space constraints, those cases are very frequent. The problem lies in the interpretation

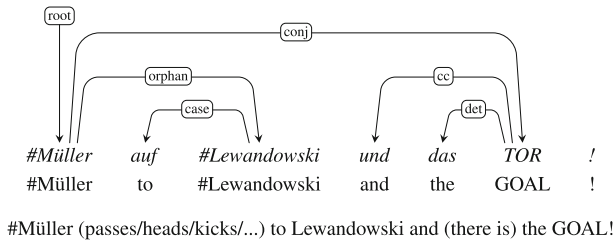


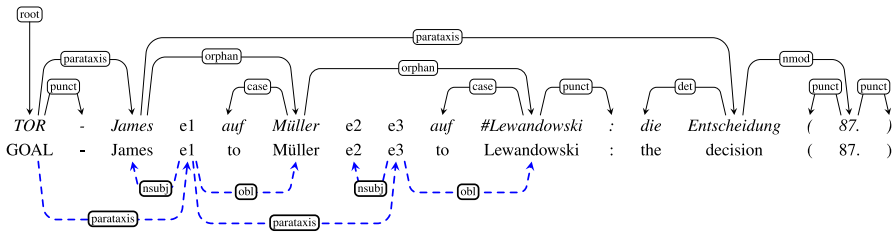
Fig. 24 German example of ellipsis from Twitter, 2020

underlying some annotation scenarios. Martínez Alonso et al. (2016) analyzed an example from a French video game chat log where all verbs were elided. Depending on contextual interpretation of a modifier, a potential analysis could result in two concurrent trees. Such an analysis is not allowed in the current UD scheme, unless the trees are duplicated and one analysis is provided for each of them.

The German example in Fig. 24 below illustrates a type of antecedent-less ellipsis that occurs in the spoken commentary of sportscasters but is also used on Twitter by users who mimic such play-by-play commentary on football games they are watching in real time. As with the French video chat example in Fig. 26, it is not clear which verb should be reconstructed in the first elliptical clause as there is no antecedent in the prior discourse. Given the context—Müller and Lewandowski are well known football players (in Germany) and the preposition *auf* signals a motion event, it is clear that the first conjunct reports an event where one player passes the ball to another. But the specific manner in which the ball is moved, whether it is ‘headed’ or ‘kicked’, could only be determined by watching footage of the game. The example also illustrates a second verb-less clause that is potentially difficult to recognize: ‘TOR’ heads its own clause and is coordinated with #Müller rather than being conjoined to #Lewandowski. The relevant clue is the capitalization that evokes the loudness and emphasis of the goal cheer. Again, one cannot be fully confident which verb to reconstruct here: several existential-type verbs are conceivable.²⁸

The example in Fig. 25 below illustrates a further variation of the above case in German: the PPs can be iterated to iconically capture a series of passes. Thus, in the example below, Müller is not only the recipient of the ball from James but also the one that passes it on to Lewandowski. However, it is not clear what structure to assume in an enhanced UD analysis that would explicate this. One could assume (i) the use of a relative clause for the second clause, (ii) two explicitly coordinated

²⁸ An anonymous reviewer asks why “das TOR” could not be treated as an interjection. The answer is the presence of the article: what the fans shout in a football stadium is simply bare *Tor!* ‘Goal!’ or *Foul!* but not *‘Das Tor!’ or *‘Das Foul’. Another reviewer suggests that the construction *PROPN auf PROPN* is not clausal but may just involve *nmod* modification of the first noun. What argues against this is the semantics: one understands this as reporting an event (*Müller passes to Lewandowski*) rather than a reference to an entity (*Müller, who passes to Lewandowski*). Making the preposition *auf* the head would go against the usual policy of UD to treat prepositions as dependents. If we thus forego that analysis, the best option seems to assume an elliptical clause. Note that this keeps the analogy with more clearly elliptical cases such as the famous (in Germany) *Manni Banane, ich Kopf - Tor* ‘Manni (kicks a) curving cross, I (hit the ball with the) head - Goal!’.



GOAL - James (passes/heads/kicks/...) to Müller who (passes/heads/kicks/...) to Lewandowski: the clinching goal (87.)

Fig. 25 German example of ellipsis from Twitter, 2020

clauses, or (iii) two clauses related by parataxis. None of these analyses would be obviously right or wrong; all would require the use of empty nodes without linguistic antecedents. For reasons of space, the bottom part of the figure shows only an enhanced UD analysis for the paratactic option (iii) using two empty nodes for predicates (e1, e3) and one for a missing subject (e2).²⁹

In any event, it is very likely that the growing number of UD treebanks containing user-generated content (and/or spoken language) will be found to feature many constructions that cannot readily be handled based on the existing guidelines for written language.

Following from a more complex French example taken from Martínez Alonso et al. (2016), Fig. 26 shows an attachment ambiguity caused by part-of-speech ambiguity and verb ellipsis. A natural ellipsis recovery of the example shown in Fig. 26 would read as “Every time **there are** 3VS1, and suddenly **I have** -2 P4”. The token “3VS1” stands for “3 versus 1”³⁰, namely an uneven combat setting, and ‘P4’ refers to a Minecraft character’s protection armor. The token “-2” allows for more than one analysis. The first analysis is the simple reading as number, complementing the noun “P4”, in blue in the graph below. A second analysis, in red, treats “-2” as a transcription of *moins de* (less of), which would be the preferred analysis given the P4 as an armor level interpretation. Needless to say, a standard UD treebank needs to commit to one analysis or the other (in this case the second), but this example shows the interplay between frequent ellipses, ergographic phenomena and the need for domain knowledge in user-generated data. It also highlights the importance of annotators’ choices when facing elided content as it would have been perfectly acceptable to use the *orphan* relation to mark the absence of, in this a case, a verbal predicate (e.g. *orphan*(3VS1, fois) and *orphan*(p4, cou ^)).

5.2 Limitations

In focusing only on user-generated content at the sentence level, our proposal does not cover phenomena that spread over multiple sentences, which would be relevant

²⁹ A coordination analysis would require an additional empty node for the conjunction; the relative clause analysis would assume a different word order: e3 would follow #Lewandowski.

³⁰ We analyze the 3VS1 token as we would analyze the word KO (knocked out) an acronym whose usage is now more predominant.

at a discourse annotation level and seen for example in cases of extra-sentential references. In the case of threaded discussions, similar to dialogue interaction, cases of gapping and more generally syntactic ellipsis can occur. These are not covered by our proposal, nor are they permitted in the UD framework, as they would require a more elaborate token indexing scheme spanning over sentences.

Like any content expressed in digital (i.e. non-handwritten) media, any conceivable variation of *ASCII art* can be used and carry meaning (Fig. 27). Formatting variations, such as a recent trend of two-column tweets, as shown in Fig. 28, are observed where some graphical layout recognition is needed to interpret the two columns as two consecutive sentences. This is similar to challenges in standard text corpora acquired from visual media, such as literary corpora from multi-column pages digitized by Optical Character Recognition (OCR). Cases such as this do not require any specific annotation as this proposal refers to the processing of (mostly) text-based interactions.

Another difficult phenomenon to annotate lies in the multi-modal nature of most user-generated content platforms that enable the inclusion of various media contents (picture, video, etc.) that often provide context to a tweet or provide meta-linguistic information that changes the whole interpretation of this content. While those phenomena do not change the core syntactic annotation *per se*, they can change the way that tokens such as proper noun strings or URLs can be interpreted.

5.3 Interoperability and other frameworks

UD enhancements and modifications The UD framework is an evolving, ongoing community effort, informed by contributors from wide and varied linguistic backgrounds. One line of changes being discussed among treebank developers concerns the inventory and design of the UD dependency relations. Croft et al. (2017) proposed a redesign of the relations-inventory that would reflect four principles from linguistic typology while essentially not changing the topology of UD trees. By contrast, Gerdes et al. (2018) proposed a surface-syntactic annotation scheme called SUD that would follow distributional criteria for defining the dependency tree structure and the naming of the syntactic functions. SUD results in changes to the dependency inventory and to tree topology.

Along another, more consensual line of exploration, proposals have been made regarding how to augment UD annotations so as to explicate additional predicate-argument relations between content words that are not captured by the basic surface-syntactic annotations of UD. Schuster and Manning (2016b) formulated five kinds of enhancements, such as propagating the relation that the head of a coordination bears to the other conjuncts. Candito et al. (2017) added the neutralization of diathesis alternations as another kind of enhancement. Given that some of the enhancements require human annotations, Droганova and Zeman (2019) propose to identify a subset of deep annotations that can be derived semi-automatically from surface trees with acceptable quality. The enhanced representation that results from these various proposals forms a directed graph but not necessarily a tree. It may contain 'null' nodes, multiple incoming edges and even cycles. Note that within the UD community, the enhancements are optional: it is acceptable for a treebank to be annotated with just a subset of possible enhancements.

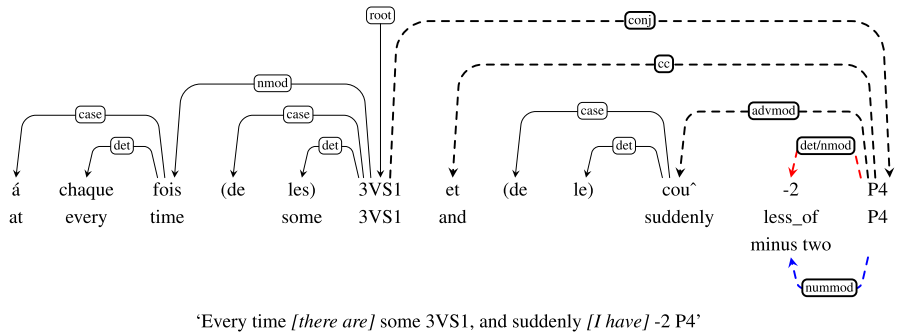


Fig. 26 Problematic example with two contesting structures from two different readings of the token “- 2” surrounded by at least 2 elided elements. (Adapted to UD v2.5 from Martínez Alonso et al. (2016). The red and blue dashed edges denote two parallel analysis depending on how the token - 2 is analyzed.). (Color figure online)

Fig. 27 ASCII art example (NO), adapted from Twitter, 2018

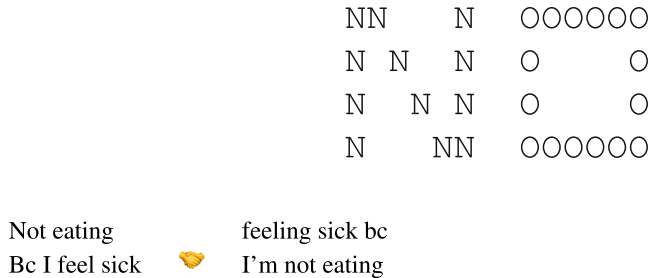


Fig. 28 Two-column tweet example from Twitter, 2020

In order to highlight the interoperability of our proposed guidelines with other frameworks (such as SUD, Gerdes et al. 2018), in Figs. 29 and 30 we display the same Italian tweet, represented in UD framework—on the left—and in SUD framework—on the right. As can be seen, in the text, a copula ellipsis occurs, as is fairly common in news headlines and other kinds of user-generated content. The different approaches of UD and SUD, with regard to the election of syntactic heads and their dependents, or the different naming of syntactic functions, both pose no problems in the syntactic representation of such a case, therefore demonstrating the interoperability of our proposal.

We also note explicitly in response to reviewer feedback that our proposal is conservative and widely harmonizes with most treebanks with respect to sentence splitting, as outlined in Sect. 4.1. This is a necessity, if we wish to enable multi-genre treebanks containing both UGC and other genres, as well as to prevent underanalysis of the syntax and overuse of paratactic ‘escape hatch’ labels (see Sect. 3.4.1).

Treatment of morphology: alignment with UniMorph Like the UD group, the collaborative UniMorph project (Kirov et al., 2016, 2018; McCarthy et al., 2020) is developing a cross-lingual schema for annotating the morphosyntactic details of language. While UD’s main focus is on the annotation of dependency syntactic

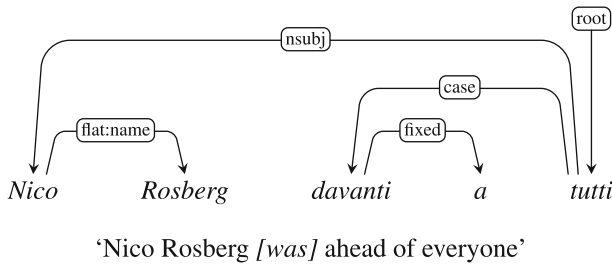


Fig. 29 Italian example of copula ellipsis represented in UD. Adapted from POSTWITA, 2011

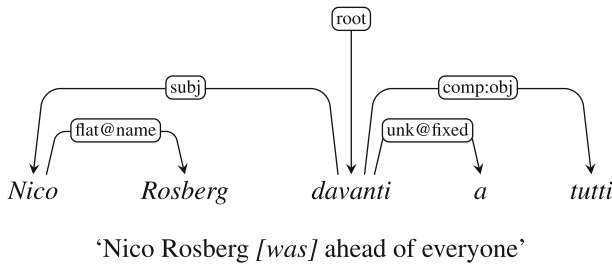


Fig. 30 Same example of ellipsis from Fig. 29, but represented through the SUD framework

relations between tokens in corpora of running text, UniMorph focuses on the analysis of morphological features at the word type level. Nevertheless, UD corpora also include more detailed annotations of lexical and grammatical properties of tokens beyond POS.

As reported by Kirov et al. (2018), a preliminary survey of UD annotations shows that approximately 68% of UD features have direct UniMorph schema equivalents, with these feature sets covering 97.04% of the complete UD tags.³¹ As the authors note, some UD features are outside the scope of UniMorph, which marks primarily morphosyntactic and morphosemantic distinctions. Conversely, some UniMorph features are not represented in UD due to its bottom-up approach. However it seems likely that many features could be mapped automatically with high accuracy using rules based on both the features in each framework and the dependency tree itself, which UD/CoNLL-U-based tools such as Udapi (Popel et al., 2017) or DepEdit (Peng & Zeldes, 2018) would facilitate.

While we present recommendations on user-generated content in this article, we neither extend the UPOS tagset nor the morphological features of the UD scheme. In that sense, existing mappings between UD and UniMorph are applicable to social media corpora. Nevertheless, our recommendations for the annotation of UGC go beyond the scope of UniMorph’s top-down approach with new types of tokens, e.g. the morphological features of pictograms. As McCarthy et al. (2020) note, UniMorph now recognizes that lemmas and word forms can be segmented, hence

³¹ McCarthy et al. (2018) reports on experiments in which UD feature annotation is deterministically converted to UniMorph format for multiple languages.

in case of clitics or agglutinative formations, morphological features can be mapped onto segments of a word form. The resulting policy should be taken into account in aligning the current UD and UniMorph.

6 Conclusion

In this article we addressed the challenges of annotating user-generated texts from the web and social media, proposing, in the context of UD, a unified scheme for their coherent treatment across different languages. Due to the variety and complexity of UGC, adequate representation of the linguistic phenomena that occur in this domain by means of an already existing scheme, such as UD, is a non-trivial task. The guidelines we outline to address this issue are relevant to all treebanks containing UGC such as those we listed in Table 1, and all those which we do not refer to here, due to their being released after our survey was carried out or being outside the scope of this article. The Irish Twitter Treebank, TwittIrish, released in UD version 2.8 is the first UGC dataset annotated according to the approach that we describe.

We hope that this proposal will trigger discussions throughout the treebanking community and will pave the way for a uniform handling of user-generated content in a dependency framework.

Acknowledgements We warmly thank Nathan Schneider for his comments on a previous version of this paper. We also thank the anonymous reviewers for their insightful comments. All remaining errors are ours. The work of C. Bosco and A. T. Cignarella was partially funded by the research project “STudying European Racial Hoaxes and stereOTYPES” (STERHEOTYPES, under the call “Challenges for Europe” of Volkswagen Stiftung). The work of C. Bosco is also partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01). The work of M. Sanguinetti is funded by PRIN 2017 (2019–2022) project *HOPE - High quality Open data Publishing and Enrichment*. Ö. Çetinoğlu is funded by DFG via project CE 326/11 *Computational Structural Analysis of German Turkish Code-Switching (SAGT)*. D. Seddah is partially funded by the ANR projects ParSiTi (ANR-16-CE33-0021) and SoSweet (ANR15-CE38-0011-01). T. Lynn and L. Cassidy are funded by the Irish Government Department of Culture, Heritage and the Gaeltacht under the GaelTech Project, and also supported by Science Foundation Ireland in the ADAPT Centre (Grant No. 13/RC/2106) at Dublin City University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

See Fig. 31 and Tables 2, 3, 4, 5, 6, 7.

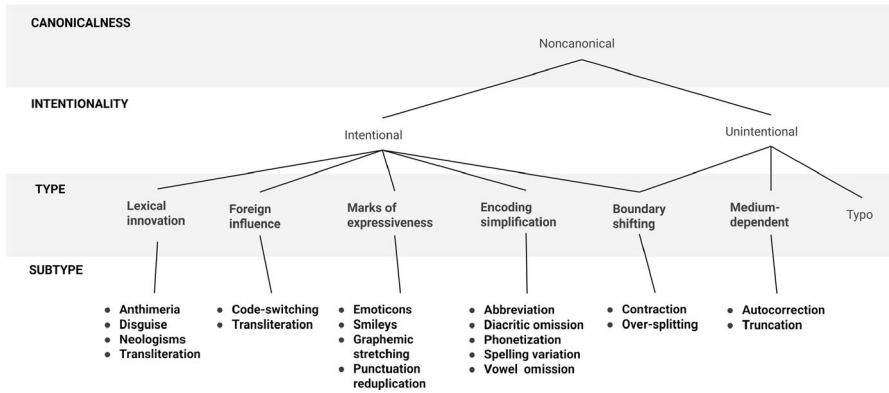


Fig. 31 Diagram of UGC phenomena (the elements in boldface are also exemplified in Table 2). “Canonicalness” refers to whether a phenomenon is also common in standard text. “Intentionality” refers to whether its production was deliberate. “Type” refers to the variety of the phenomenon, while “Subtype” provides sub-categorization of each type. Our focus is on the noncanonical linguistic phenomena prevalent in UGC which do not yet have standardized annotation guidelines within the UD framework. These elements in boldface are exemplified in Table 2. It is finally worth pointing out that the categorization of intentionality may only be guessed at by the annotator as it is unknowable by observing the surface text alone

Table 2 Examples of UGC phenomena in different languages (DE: German, EN: English, FR: French, GA: Irish, IT: Italian, TR: Turkish) based on the categorization proposed in Fig. 31

Phenomenon	Lang	Attested example	Standard form	Gloss
Encoding simplification				
Diacritic omission	GA	<i>Léigh arís!</i>	<i>Léigh arís!</i>	'Read again!'
	TR	<i>İstanbuldaki ağaçlar</i>	<i>İstanbul'daki ağaçlar</i>	'trees in Istanbul'
Vowel omission	EN	<i>ppl</i>	<i>people</i>	'people'
	TR	<i>slm</i>	<i>selam</i>	'hi'
Phonetization	EN	<i>Happy Birthday 2 me</i>	<i>Happy Birthday to me</i>	'Happy Birthday to me'
	TR	<i>1 az</i>	<i>biraz</i>	'some'
	DE	<i>k 1 Mensch hat so</i>	<i>kein Mensch hat so</i>	'nobody has such a
		<i>1 Thailandhass</i>	<i>einen Thailandhass</i>	hatred of 'Thailand'
Spelling variation	FR	<i>je sé</i>	<i>je sais</i>	'I know'
	GA	<i>gura míle</i>	<i>go raibh míle</i>	'thank you very much'
	FR	<i>tous mes examen</i>	<i>tous mes examens</i>	'All my examinations
		<i>son normaux</i>	<i>sont normaux</i>	'are normal'
	IT	<i>anno mangiato</i>	<i>hanno mangiato</i>	'(they) have eaten'
Abbreviation	EN	<i>govt</i>	<i>government</i>	'government'
	DE	<i>zuggm</i>	<i>zugegebenermaßen</i>	'admittedly'
Boundary shifting				
Contraction	FR	<i>nimp quoi</i>	<i>n'importe quoi</i>	'rubbish'
Over-splitting	FR	<i>c a dire</i>	<i>c'est-à-dire</i>	'namely'
	TR	<i>gele bilirim</i>	<i>gelebilirim</i>	'I can come'
Marks of expressiveness				
Punct. reduplication	FR	<i>Joli !!!!!</i>	<i>Joli !</i>	'nice!'
	IT	<i>chi ???!?!?</i>	<i>chi?</i>	'who?'
Case variation	GA	<i>is BREÁ le daoine</i>	<i>is breá le daoine</i>	'people love'

Table 2 continued

Phenomenon	Lang	Attested example	Standard form	Gloss
Graphemic stretching	EN	superTTTTTTT	<i>super</i>	'great'
	IT	stiiiiiiiiiii	<i>sì</i>	'yes'
Emoticons/smileys	-	:-) <3	-	-
	GA	<3 <i>mór</i>	<i>Grá mór</i>	'Lots of love'
Lexical innovation				
Disguise	IT	caxxo	<i>cazzo</i>	'fuck'
	TR	mo<i>k</i> / b<i>k</i> / b*<i>k</i>	<i>bok</i>	'shit'
Anthimeria	DE	Verfick**t lange Reise	<i>Verfickt lange Reise</i>	'fucking long trip'
	IT	tuittare	<i>twittare</i>	'to tweet'
	EN	feel free to PM	<i>personal message</i>	'to send a message'
	DE	achtisch	<i>EN eightish</i>	'about 8 o'clock'
Foreign language influence				
Transliteration	GA	áicbheaird	<i>amscaí</i>	'awkward'
	TR	taymlayn	<i>zaman akış</i>	'timeline'
Medium-dependent phenomena				
Truncation	GA	thart fa' 53 nó...	<i>thart fa' 53 nóiméad</i>	'over 53 mi...(minutes)'
Autocorrection	GA	concise	<i>coicise</i>	'fortnight'

Table 3 Summary of CoNLL-U proposed implementations (part 1)

Annotation issue	Token		Lemma		UPOS	
	No change	Split	No change	Normalize	Standard synt. role	Other
Diacritic omission	✓			✓	✓	
Vowel omission	✓			✓	✓	
Phonetization	✓			✓	✓	
Spelling errors	✓			✓	✓	
Abbreviation	✓			✓	✓	
Contraction						
Canonical		✓		✓	✓	
Noncanonical & unintentional		✓		✓	✓	
Noncanonical & intentional	✓		✓		✓	
Oversplitting	✓		✓ (remaining tok.)	✓ (first token)	✓ (first token)	× (remaining tok.)
Punctuation redup.	✓			✓	✓	
Graphemic stretch.	✓			✓	✓	
Disguise	✓			✓	✓	
Transliteration	✓			✓ (remove inflect.)	✓	
Neologism	✓				✓	
Truncation	✓		✓		✓ (if known)	
Autocorrection	✓				✓ (if known)	
Hashtags				✓		
Synt. integrated Standalone	✓		✓		✓	
At-mentions						
Synt. integrated Standalone	✓		✓			PROPN
URLs						
Synt. integrated Standalone	✓		✓		SYM	

Table 3 continued

Annotation issue	Token		Lemma		UPOS		
	No change	Split	No change		Normalize	Standard synt. role	Other
Pictograms/emoticons							
Synt. integrated standalone RTs	✓ (single)	✓ (multi)	✓ (if not resolvable to word)	✓ (if resolvable to word)	✓ (if resolvable to word)	✓ (if resolvable to word)	SYM
Synt. integrated standalone	✓	✓	✓	✓ (remove inflect.)	✓ (VERE/NOUN)		SYM
Markup symbols	✓	✓	✓		PUNCT		
Code-switching							
INTRA	✓		✓ (if not known)	✓ (if known)	✓ (if known)		X (if not known)
INTER							
MIXED							
Disfluencies							
Repairs hesitation	✓		✓			✓ (if known)	X (if not known)
Sent. boundaries							

Table 4 Summary of CoNLL-U proposed implementations (Part II)

Annotation issue	FEATS	DEPREL			MISC
		Standard	synt. role	Other	
Diacritic omission		✓			NonCan=SpellVar CorrectForm
Vowel omission		✓			NonCan=CharOm CorrectForm
Phonetization		✓			NonCan=Phon CorrectForm
Spelling errors	Typo=Yes	✓			CorrectForm
Abbreviation	Abbr=Yes	✓			FullForm
Contraction					
Canonical	Abbr=Yes	✓			
Noncanonical and unintentional	Typo=Yes	✓			CorrectForm CorrectSpaceAfter SpaceAfter NonCan=Cont
Noncanonical and intentional	Abbr=Yes	✓			NonCan=Cont
Oversplitting		✓	(first token)	goeswith	NonCan=OS
Punctuation redupl.		✓			NonCan=PunctVar
Graphemic stretching		✓			CorrectForm NonCan=Stretch
Disguise		✓			CorrectForm NonCan=SpellVar
Transliteration	Foreign=Yes	✓			CorrectForm NonCan=Transl
Neologism		✓			NonCan=Neo
Truncation		✓			FullForm NonCan=Trunc (if known)
Autocorrection		✓			CorrectForm NonCan=AutoC
Hashtags					
Synt. integrated		✓			
Standalone				parataxis:hashtag	
At-mentions					
Synt. integrated		✓			

Table 4 continued

Annotation issue	FEATS	DEPREL		MISC
		Standard synt. role	Other	
Standalone			vocative	
URLs				
Synt. integrated		✓		
Standalone			parataxis:url	
Pictograms/emoticons				
Synt. integrated		✓		
Standalone			discourse	
RTs				
Synt. integrated	Abbr=Yes	✓		FullForm
Standalone			parataxis	punct
Markup symbols				
Code-switching				
INTRA	Foreign=Yes	✓ (if known)	flat:foreign (if not known)	CSType=INTRA Lang=[isocode]
INTER				CSType=INTER Lang=[isocode]
MIXED				CSType=MIXED Lang=[isocode]
Disfluencies				
Repairs			reparandum	
Hesitation markers			discourse	
Sentence boundaries			parataxis:sentence	

Table 5 CoNLL-U example of transliteration and graphemic stretching in Irish

ID	Token	Lemma	UPOS	FEATS	Head	DEPREL	MISC
1	Féar	fair	ADJ	Foreign=Yes	2	amod	Lang=en NonCan=Transl
2	plé	play	NOUN	Foreign=Yes	0	root	Lang=en NonCan=Transl
3	:))	:))	SYM	_	2	discourse	NonCan=Stretch

The English phrase 'fair play' has been transliterated using the Irish spelling system

```
# text = Féar plé :)
# gloss = Fair play :))
```

Table 6 CoNLL-U annotation of an example of character omission in German

ID	Token	Lemma	UPOS	FEATS	Head	DEPREL	MISC
1	Bin	sein	AUX	Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin	4	cop	-
2	zuggm	zugegebenermaßen	ADV	-	4	advmod	CorrectForm=zugegebenermaßen NonCan=CharOm
3	etw	etwas	ADV	-	4	advmod	CorrectForm=etwas NonCan=CharOm
4	naiv	naiv	ADJ	-	0	root	-

text = Bin zuggm etw naiv

gloss = Admittedly, I was somewhat naive

Table 7 CoNLL-U annotation of an example of over-splitting in Turkish

ID	Token	Lemma	UPOS	FEATS	HEAD	DEPREL	MISC
1	Partiye	parti	NOUN	Case=Dat Number=Sing	2	obl	—
2	gele	gel	VERB	Aspect=Hab Mood=Pot Number=Sing Person=1 Polarity=Pos Tense=Pres	0	root	NonCan=OS
3	billirim	—	X	—	3	goeswith	—

text = Partiye gele billirim

gloss = I can come to the party

References

- Albogramy, F., & Ramsay, A. (2017). Universal dependencies for Arabic Tweets. In *International conference recent advances in natural language processing, (RANLP)* (pp. 46–51).
- Aufrant, L., Wisniewski, G., & Yvon, F. (2017). LIMS@CoNLL'17: UD shared task. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 163–173).
- Azzi, A. A., Bouamor, H., & Ferradans, S. (2019). The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the first workshop on financial technology and natural language processing* (pp. 74–80), Macao, China. <https://www.aclweb.org/anthology/W19-5512>
- Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 120–128), Atlanta, GA.
- Behzad, S., & Zeldes, A. (2020). A cross-genre ensemble approach to robust reddit part of speech tagging. In *Proceedings of the 12th web as corpus workshop (WAC-XII)* (pp. 50–56), Marseille, France.
- Bhat, I., Bhat, R. A., Shrivastava, M., & Sharma, D. (2018). Universal dependency parsing for Hindi–English code-switching. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Vol. 1 (Long Papers)* (pp. 987–998).
- Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL interactive poster and demonstration sessions* (pp. 214–217), Barcelona, Spain. Association for Computational Linguistics.
- Björkelund, A., Falenska, A., Yu, X., & Kuhn, J. (2017). IMS at the CoNLL 2017 UD shared task: CRFs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 40–51), Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-3004>
- Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K., Mertens, P., & Willems, D. (1990). *Le français parlé. Études grammaticales*. CNRS Editions.
- Blodgett, S. L., Wei, J. T. Z., & O'Connor, B. (2018). Twitter universal dependency parsing for African-American and mainstream American English. In *Proceedings of the ACL 2018—56th annual Meeting of the Association for Computational Linguistics (Long Papers)* (Vol. 1, pp. 1415–1425). ACL.
- Bosco, C., Tamburini, F., Bolioli, A., & Mazzei, A. (2016). Overview of the EVALITA 2016 part of speech tagging on TTwitter for ITALian Task. In *Proceedings of the fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*. CEUR.
- Candito, M., Guillaume, B., Perrier, G., & Seddah, D. (2017). Enhanced UD dependencies with neutralized diathesis alternation. In *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)* (pp. 42–53), Pisa, Italy. Linköping University Electronic Press. <https://www.aclweb.org/anthology/W17-6507>
- Caron, B., Courtin, M., Gerdes, K., & Kahane, S. (2019). A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th international workshop on treebanks and linguistic theories (TLT'19)* (pp. 13–24). ACL.
- Çetinoğlu, Ö. (2016). A Turkish-German Code-Switching Corpus. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC'16)* (pp. 4215–4220). ELRA.
- Çetinoğlu, Ö., & Çöltekin, Ç. (2016). Part of speech annotation of a Turkish-German code-switching corpus. In *Proceedings of the tenth Linguistic Annotation Workshop (LAW-X)* (pp. 120–130). ACL.
- Cignarella, A. T., Bosco, C., & Rosso, P. (2019). Presenting TWITTURO-UD: An Italian Twitter Treebank in universal dependencies. In *Proceedings of the fifth international conference on dependency linguistics (Depling, SyntaxFest 2019)* (pp. 190–197).

- Croft, W., Nordquist, D., Looney, K., & Regan, M. (2017). Linguistic typology meets universal dependencies. In *Proceedings of the 15th international workshop on Treebanks and Linguistic Theories (TLT)* (pp. 63–75).
- Daiber, J., & Van Der Goot, R. (2016). The denoised Web Treebank: Evaluating dependency parsing under noisy input conditions. In *Proceedings of the 10th international conference on Language Resources and Evaluation (LREC 2016)* (pp. 649–653).
- Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies treebank for Slovenian. In *Proceedings of the 6th workshop on Balto-Slavic natural language processing*. Association for Computational Linguistics.
- Dobrovoljc, K., & Nivre, J. (2016). The universal dependencies treebank of spoken Slovenian. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)* (pp. 1566–1573). ELRA.
- Droganova, Kira, & Zeman, Daniel. (2019). Towards deep universal dependencies. In *Proceedings of the fifth international conference on Dependency Linguistics (Depling, SyntaxFest 2019)* (pp. 144–152), Paris, France. ACL.
- Eisenstein, J. (2013). What to do about bad language on the Internet. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 359–369). ACL.
- Fischer, K. (2006). Frames, constructions, and morphemic meanings: The functional polysemy of discourse particles. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 427–447). Elsevier.
- Foster, J. (2010). “cba to check the spelling”: Investigating parser performance on discussion forum posts. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics* (pp. 381–384). ACL.
- Foster, J., Çetinoğlu, Ö., Wagner, J., Le Roux, J., Nivre, J., Hogan, D., & van Genabith, J. (2011). From News to comment: Resources and benchmarks for parsing the Language of Web 2.0. In *Proceedings of 5th international joint conference on Natural Language Processing* (pp. 893–901).
- Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G.. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the second workshop on Universal Dependencies (UDW 2018)* (pp. 66–74), Brussels, Belgium, November. ACL.
- Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 42–47). ACL.
- Kaljahi, R., Foster, J., Roturier, J., Ribeyre, C., Lynn, T., & Le Roux, J. (2015). Foreebank: Syntactic analysis of customer support forums. In *Conference proceedings—EMNLP 2015: Conference on empirical methods in natural language processing* (pp. 1341–1347).
- Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., & Hulden, M. (2018). UniMorph 2.0: Universal morphology. In *Proceedings of the eleventh international conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1293>
- Kirov, C., Sylak-Glassman, J., Que, R., & Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC)*. ELRA.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). A dependency parser for Tweets. In *The conference on Empirical Methods in Natural Language Processing (EMNLP’14)* (pp. 1001–1012).
- Lacheret, A., Kahane, S., Beliaou, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., & Tchobanov, A. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC’14)* (pp. 295–301), Reykjavik, Iceland. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/381_Paper.pdf

- Leung, H., Poiret, R., Wong, T., Chen, X., Gerdes, K., & Lee, J. (2016). Developing universal dependencies for Mandarin Chinese. In *Proceedings of the 12th workshop on Asian Language Resources (ALR12)* (pp. 20–29).
- Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., & Smith, N. A. (2018). Parsing Tweets into universal dependencies. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers)* (pp. 965–975). ACL.
- Luotolahti, J., Kanerva, J., Laippala, V., Pyysalo, S., & Ginter, F. (2015). Towards universal web parsebanks. In *Proceedings of the third international conference on Dependency Linguistics (Depling 2015)* (pp. 211–220). Uppsala University.
- Lynn, T., & Scannell, K. (2019). Code-Switching in Irish Tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology workshop* (pp. 32–40). European Association for Machine Translation.
- Lynn, T., Scannell, K., & Maguire, E. (2015). Minority language Twitter: Part-of-speech tagging and analysis of Irish Tweets. In *Proceedings of the workshop on noisy user-generated text* (pp. 1–8). ACL.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014: System demonstrations* (pp. 55–60). Baltimore, MD.
- Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- Martínez Alonso, H., Seddah, D., & Sagot, B. (2016). From noisy questions to minecraft texts: Annotation challenges in extreme syntax scenarios. In *Proceedings of the 2nd workshop on noisy user-generated text* (pp. 127–137).
- Mataoui, M., Hacine, T. E. B., Tellache, I., Bakhtouchi, A., & Zemat, O. (2018). A new syntax-based aspect detection approach for sentiment analysis in Arabic reviews. In *Proceedings of ICNLPSP 2018* (pp. 1–6), Algiers.
- McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., NatalyKrizhanovsky, A. K., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., & Yarowsky, D. (2020). UniMorph 3.0: Universal morphology. In *Proceedings of The 12th language resources and evaluation conference* (pp. 3922–3931), Marseille, France. European Language Resources Association. ISBN 979-10-95546-34-4. <https://www.aclweb.org/anthology/2020.lrec-1.483>
- McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., & Yarowsky, D. (2018). Marrying universal dependencies and universal morphology. In *Proceedings of the second workshop on Universal Dependencies (UDW 2018)* (pp. 91–101), Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6011>. <https://www.aclweb.org/anthology/W18-6011>
- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F. M., & Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *CoRR*. [arXiv:2004.10643](https://arxiv.org/abs/2004.10643)
- Øvrelid, L., & Hohle, P. (2016). Universal dependencies for Norwegian. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)* (pp. 1579–1585). ELRA.
- Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL, 2013* (pp. 380–390).
- Pamay, T., Sulubacak, U., Torunoğlu-Selamet, D., & Eryiğit, G. (2015). The annotation process of the ITU Web Treebank. In *Proceedings of the 9th linguistic annotation workshop* (pp. 95–101).
- Peng, S., & Zeldes, A. (2018). All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the joint workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)* (pp. 167–177). Santa Fe, NM.
- Petrov, S., & McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the first workshop on syntactic analysis of non-canonical language (SANCL)*.

- Pietrandrea, P., Kahane, S., Lacheret-Dujour, A., & Sabio, F. (2014). The notion of sentence and other discourse units in corpus annotation. In T. Raso & H. Mello (Eds.), *Spoken corpora and linguistic studies* (pp. 331–364). John Benjamins.
- Popel, M., Zabokrtský, Z., & Vojtek, M. (2017). Udapi: Universal API for universal dependencies. In *Universal dependencies workshop at NoDaLiDa, 2017* (pp. 96–101).
- Proisl, T. (2018). Someweta: A part-of-speech tagger for German Social Media and web texts. In *Proceedings of the 11th international conference on language resources and evaluation (LREC 2018)* (pp. 665–670). ELRA.
- Read, J., Dridan, R., Oepen, S., & Solberg, L. J. (2012a). Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters* (pp. 985–994), Mumbai, India. The COLING 2012 Organizing Committee. <https://www.aclweb.org/anthology/C12-2096>
- Read, J., Flickinger, D., Dridan, R., Oepen, S., & Øvrelid, L. (2012b). The wesearch corpus, treebank, and treecache. a comprehensive sample of user-generated content. In *Proceedings of the 8th international conference on language resources and evaluation*.
- Rehbein, I. (2015). Filled pauses in user-generated content are words with extra-propositional meaning. In *Proceedings of the second workshop on extra-propositional aspects of meaning in computational semantics (ExProM 2015)* (pp. 12–21). ACL.
- Rehbein, I., Ruppenhofer, J., & Do, B.-N. (2019). tweeDe—A universal dependencies Treebank for German tweets. In *Proceedings of the 17th workshop on Treebanks and Linguistic Theories (TLT 2019)*.
- Rehbein, I., Ruppenhofer, J., & Zimmermann, V. (2018). A harmonised testsuite for pos tagging of german social media data. In *Proceedings of the 27th international conference on computational linguistics, KONVENS 2018* (pp. 18–28), Wien, Österreich.
- Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 101–124). John Benjamins.
- Sanchez, G. (2019). Sentence boundary detection in legal text. In *Proceedings of the natural legal language processing workshop 2019* (pp. 31–38), Minneapolis, Minnesota. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-2204>. <https://www.aclweb.org/anthology/W19-2204>.
- Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., & Zeldes, A. (2020). Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the 12th language resources and evaluation conference* (pp. 5240–5250), Marseille, France. European Language Resources Association.
- Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., & Tamburini, F. (2018). PoSTWITA-UD: An Italian Twitter Treebank in universal dependencies. In *LREC 2018—11th international conference on language resources and evaluation* (pp. 1768–1775).
- Schuster, S., Lamm, M., & Manning, C. D. (2017). Gapping constructions in universal dependencies v2. In *Proceedings of the NoDaLiDa 2017 workshop on Universal Dependencies (UDW 2017)* (pp. 123–132). ACL.
- Schuster, S., & Manning, C. D. (2016a). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the 10th language resource and evaluation conference (LREC 2016)* (pp. 2371–2378). ELRA.
- Schuster, S., & Manning, C. D. (2016b). Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 2371–2378), Portorož, Slovenia. European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L16-1376>
- Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Suárez, P. J. O., Sagot, B., & Srivastava, A. (2020). Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 1139–1150). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.107>. <https://www.aclweb.org/anthology/2020.acl-main.107>

- Seddah, D., Sagot, B., Candito, M., Mouilleron, V., & Combet, V. (2012). The French Social Media Bank: A treebank of noisy user generated content. In *24th International conference on computational linguistics—Proceedings of COLING 2012: Technical papers* (pp. 2441–2458). ACL.
- Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on YouTube. *Information Processing & Management*, 52(1), 46–60.
- Silveira, N. Dozat, T., De Marneffe, M. C., Bowman, S. R., Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the 9th international conference on language resources and evaluation, LREC 2014* (pp. 2897–2904). ELRA.
- Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., Alghamdi, F., Hirschberg, J., Chang, A., & Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the CodeSwitch workshop*.
- Solorio, T., & Liu, Y. (2008). Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP '08)* (pp. 1051–1060). ACL.
- Stevenson, M., & Gaizauskas, R. (2000). Experiments on sentence boundary detection. In *Proceedings of the sixth conference on applied natural language processing, ANLC '00* (pp. 84–89), USA. Association for Computational Linguistics. <https://doi.org/10.3115/974147.974159>
- Taulé, M., Martí, M. A., Bies, A., Nofre, M., Garí, A., Song, Z., Strassel, S., & Ellis, J. (2015). Spanish treebank annotation of informal non-standard web text. In F. Daniel & O. Diaz (Eds.), *Current trends in web engineering* (pp. 15–27). Springer.
- Tyers, F. M., & Mishchenkova, K. (2020). Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the fourth workshop on Universal Dependencies (UDW 2020)* (pp. 195–204).
- universaldependencies.org. (2019a). Typos and other errors in underlying text: Wrongly split word. <https://universaldependencies.org/u/overview/typos.html#wrongly-split-word>. Accessed: 2020-07-06.
- universaldependencies.org. (2019b). Enhanced dependencies. Retrieved August 3, 2020, from <https://universaldependencies.org/u/overview/enhanced-syntax.html>
- universaldependencies.org. (2019c). Annotation of foreign strings in the Universal Dependencies guidelines. Retrieved November 28, 2019, from <https://universaldependencies.org/cs/dep/flat-foreign.html>
- universaldependencies.org. (2019d). Pos-tagging of foreign tokens in the Universal Dependencies guidelines. Retrieved November 28, 2019, from <https://universaldependencies.org/u/pos/X.html>
- universaldependencies.org. (2019e). Morphology: General principles. Retrieved July 15, 2019, from <https://universaldependencies.org/u/overview/morphology.html>
- universaldependencies.org. (2019f). Annotation of speech repair in the Universal Dependencies guidelines. Retrieved November 28, 2019, from <https://universaldependencies.org/u/dep/reparandum.html>
- universaldependencies.org. (2019g). Tokenization and Word Segmentation guidelines. Retrieved December 2, 2019, from <https://universaldependencies.org/u/overview/tokenization.html>
- universaldependencies.org. (2021). Annotation of style or sublanguage to which a word form belongs. Retrieved December 15, 2021, from <https://universaldependencies.org/u/feat/Style.html>
- Van Der Goot, R., & van Noord, G. (2018). Modeling input uncertainty in neural network dependency parsing. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4984–4991).
- Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the slovene corpus gos. *Language Resources and Evaluation*, 47, 12. <https://doi.org/10.1007/s10579-013-9216-5>
- Vilares, D., Gómez-Rodríguez, C., & Alonso, M. A. (2017). Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118, 45–55.
- Wang, H., Zhang, Y., Chan, G. Y. L., Yang, J., & Chieu, H. L. (2017). Universal dependencies parsing for colloquial Singaporean English. In *ACL 2017—55th annual meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* (Vol. 1, pp. 1732–1744).

- Wang, W. Y., Kong, L., Mazaitis, K., & Cohen, W. W. (2014). Dependency parsing for Weibo: An efficient probabilistic logic programming approach. In *EMNLP 2014—2014 conference on empirical methods in Natural Language Processing, Proceedings of the Conference* (pp. 1152–1158).
- Westpfahl, S., & Gorisch, J. (2018). A syntax-based scheme for the annotation and segmentation of German spoken language interactions. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)* (pp. 109–120), Santa Fe, New Mexico, USA. Association for Computational Linguistics. <https://www.aclweb.org/anthology/W18-4913>
- Wong, T., Gerdes, K., Leung, H., & Lee, J.. (2017). Quantitative comparative syntax on the Cantonese–Mandarin parallel dependency treebank. In *Proceedings of the fourth international conference on Dependency Linguistics (Depling 2017)* (pp. 266–275), Pisa, Italy. Linköping University Electronic Press. <https://www.aclweb.org/anthology/W17-6530>
- Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3), 581–612.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.