Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5240–5250
Marseille, 11–16 May 2020

Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies

Manuela Sanguinetti[1], Cristina Bosco[1], Lauren Cassidy[2], Özlem Çetinoğlu[3], Alessandra Teresa Cignarella[1,4], Teresa Lynn[2], Ines Rehbein[5] Josef Ruppenhofer[6], Djamé Seddah[7], Amir Zeldes[8]

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

{msanguin|bosco|cigna}@di.unito.it, ozlem.cetinoglu@ims.uni-stuttgart.de, {teresa.lynn|lauren.cassidy}@adaptcentre.ie, ines@informatik.uni-mannheim.de, ruppenhofer@ids-mannheim.de, djame.seddah@inria.fr, amir.zeldes@georgetown.edu

Abstract

The paper presents a discussion on the main linguistic phenomena of user-generated texts found in web and social media, and proposes a set of annotation guidelines for their treatment within the Universal Dependencies (UD) framework. 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 paper is twofold: (1) to provide a short, 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 main goal of this paper is to provide a common framework for those teams interested in developing similar resources in UD, thus enabling cross-linguistic consistency, which is a principle that has always been in 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 eligible source of data for a large number of research fields and applications, especially for sentiment analysis and opinion mining. In order to successfully process the data available from such sources, linguistic analysis is often helpful, 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 (Gimpel et al., 2011; Owoputi et al., 2013; Lynn et al., 2015; Bosco et al., 2016; Çetinoğlu and Çöltekin, 2016; Proisl, 2018) and parsing (Foster, 2010; Petrov and McDonald, 2012; Kong et al., 2014; Liu et al., 2018), automatic processing of user-generated content (UGC) still represents a challenging task, as is shown by the workshop series on noisy user-generated text (W-NUT)[1]. UGC is a continuum of text sub-genres that may considerably vary according to the specific conventions and limitations posed by the medium used (blog, discussion forum, online chat, microblog, etc.), its degree of "canonicalness" with respect to a more standard language, as well as the linguistic devices[2] adopted to convey a message. Overall, however, there are some well-recognized phenomena that characterize UGC as a whole (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013), and that continue to make its treatment a difficult task.

The availability of *ad hoc* training resources remaining an essential factor for the analysis of these texts, in the last decade, numerous resources of this type have been developed. A good proportion of these have been annotated according to the UD scheme (Nivre et al., 2016), a dependency-based format which has achieved great popularity, becoming a popular reference for treebank annotation also because of its adaptability to different domains and genres.

On the one hand, this flexibility opens up the possibility of adopting the UD scheme for user-generated texts too; on the other hand, the UD guidelines did not fully account for some specificity of this domain, thus leaving it at the discretion of the individual annotator (or teams of annotators) to interpret the guidelines and identify the most appropriate representation. This paper therefore sets itself the goal of drawing attention to the annotation issues of UGC found especially on social media, and to the main problems encountered while attempting to find a cross-linguistically consistent representation, all within a single coherent framework.

The paper is structured such as to provide 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 propose a discussion on some of these phenomena within the context of the framework of UD and propose, wherever possible, an annotation choice aimed at overcoming the inconsistencies found among the existing resources. Given the nature of the

[1] https://noisy-text.github.io/

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

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; this paper therefore has no prescriptive intent. Instead, the objective is to establish a starting point for discussion, eventually arriving at a set of guidelines that allow for a more uniform treatment of UGC-specific phenomena across different languages.

2. Linguistics of UGC

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

- **Encoding simplification:** This axis covers ergographic phenomena, *i.e* phenomena aiming at reducing the writing effort, such as diachritic or vowel removals (*ppl* → *people*).
- **Transverse Phenomenon**: Some phenomena affect the number of tokens, compared to standard languages, either by replacing several standard language tokens by only one, which we shall call a *contraction* (*iyakşamlar* → *iyi akşamlar*, "good evening"), or conversely by splitting one standard language token into several tokens, called *over-splitting* (*c t* → *c'était*, "it was"). Such phenomena are frequent in our corpora, and they need specific annotation guidelines.
- **Sentiment expression:** This concerns markers of expressiveness, e.g., graphical stretching (*yesssss* → *yes*), replication of punctuation marks (*????* → *?*), emoticons, sometimes used as a verb (*Je t'<3* → *Je t'aime*, "I love you"). These phenomena aim at emulating sentiment expressed through prosody and gesture in direct interaction. Many of these symbols contain punctuation which can lead to spurious tokenization.
- **Foreign Language Influence**: UGC is often produced in highly multilingual settings and we often find evidence for the influence of foreign language(s) on the users' text productions, especially in code-switching scenarios, in domain-specific conversations (video games chat log) or in the productions of L2 speakers. A good example would be the Irish term coined by one user to mean 'awkward', *áicbheaird* (instead of the Irish term *amscaí*), whose pronunciation mimics the English word.
- **Context dependency** Given the conversational nature of most social media, not unlike dialogue-based interaction, speaker turns in UGC are often marked by the thread structure and can provide a context rich enough to allow for varying levels of ellipsis and anaphora. 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. This makes the annotation task more difficult, prone to interpretation errors if the actual context is not available.

Table 1 presents some cross-language examples of the strands presented above.

3. Web 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. 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. 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). The results of our search are summarized in Table 2.[3]

Based on the selection criteria mentioned above, we found 19 papers describing a resource featuring web/social media texts; most of them are freely available, either from a GitHub repository, a dedicated web page or upon request.

Their sizes vary, ranging from 500 (DWT) to approximately 6,700 tweets (Pst) for the Twitter treebanks, and from 974 (xUGC) to more than 200 million sentences (TDT) for the other datasets.

Languages English is still the most represented language, however, some of these resources focus on different language variants such as Hindi-English code switching data (Hi-En-CS), African-American English (TAAE) and Singaporean English (SDT). Three resources are in French (Frb, xUGC, FSMB) and two in Italian (TWRO, Pst); the remaining ones are in Arabic (ATDT), Chinese (CWT), Finnish (TDT), German (tweeDe) and Turkish (ITU). While the current Irish Twitter corpus has not yet been converted to treebank format (and as such is not listed in Table 2), its annotation presented most of the challenges that make up this discussion (Lynn et al., 2015; Lynn and Scannell, 2019).

Data sources 12 out of 19 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 as regards the use of data for academic and non-commercial purposes[4]. Only two resources include data from social media other than Twitter, i.e. Facebook (FSMB) and Sina Weibo (CWT), and, overall, most of the remaining resources comprise texts from discussion forums of any kind. Only two treebanks consist of texts from different sub-domains, i.e. blogs, reviews, emails, newsgroups and question answers (EWT), and Wikinews, Wikivoyage, wikiHow, Wikipedia, interviews, Creative Commons fiction and Reddit (GUM), and one is made up of generic data automatically crawled from the web (TDT).

Syntactic frameworks As regards the formalism adopted to represent the syntactic structure, dependencies are by far the most used paradigm, especially among the treebanks created from 2014 onward. As also pointed out by Martínez Alonso et al. (2016), a dependency-based annotation lends

[3] A more complete table with additional information on the surveyed treebanks can be found here: `http://di.unito.it/webtreebanks`.

[4] `https://developer.twitter.com/en/developer-terms/agreement-and-policy#c-respect-users-control-and-privacy`

Phenomenon	Lang	Attested example	Standard form	Gloss
Ergographic phenomena (encoding simplification)				
Diacritic removal	GA	**Leigh aris***!*	*Léigh arís!*	'Read again!'
	TR	**Istanbuldaki agaclar**	*İstanbul'daki ağaçlar*	'trees in Istanbul'
Vowel removal	EN	**ppl**	*people*	'people'
	TR	**slm**	*selam*	'hi'
Phonetization	EN	*Happy Birthday* **2** *me*	*Happy Birthday to me*	'Happy Birthday to me'
	TR	**n** *zmn*	*ne zaman*	'when'
Simplification	FR	*je* **sé**	*je sais*	'I know'
	GA	**gura** *míle*	*go raibh míle*	'thank you very much'
Spelling errors	FR	*tous mes* **examen**	*tous mes examens*	'All my examinations
	FR	**son** *normaux*	*sont normaux*	are normal'
	IT	**anno** *mangiato*	*hanno mangiato*	'(they) have eaten'
Transverse phenomena				
Contraction	FR	**nimp**	*n'importe quoi*	'rubbish'
	EN	**govt**	*government*	'government'
Oversplitting	FR	**c a dire**	*c'est-à-dire*	'namely'
	TR	**gele bilirim**	*gelebilirim*	'I can come'
Marks of expressiveness				
Punct. transgression	FR	*Joli !!!!!!*	*Joli !*	'nice!'
	IT	*chi?!?!?!*	*chi?*	'who?'
Graphemic stretching	EN	*superrrrrrrrr*	*super*	'great'
	IT	*siiiiiiiiiiii*	*sì*	'yes'
Self-censorship	IT	**caxxo**	*cazzo*	'fuck'
	TR	**mok / b.k / b*k**	*bok*	'shit'
Emoticons/smileys	-	*:-)* <3	–	–
	GA	*<3 mór*	*Grá mór*	'Lots of love'
Foreign Language Influence				
Transliteration	GA	*áicbheaird*	*amscaí*	'awkward'
	TR	*taymlayn*	*zaman akışı*	'timeline'
Verb Formation	IT	*tuittare*	*twittare*	'to tweet'
	EN	feel free to *PM*	*personal message*	'to send a message'
Autocorrection	GA	*concise*	*coicíse*	'fortnight'

Table 1: Multi-lingual examples of UGC phenomena.

itself well to noisy texts, since it is easier to deal with disfluencies and fragmented text breaking prescriptive linguistic rules that prohibit discontinuous constituents.

The increasing popularity of UD may also have a role in this trend, considering that 12 out of the 15 dependency treebanks are based on the UD scheme. Although not all of them have been released in the official repository, and some of them do not strictly comply with the format specifications, this highlights the need to converge into a single annotation framework, to allow for better comparability of the resources.

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

3.1. Annotation Comparison

To explore the similarities and divergences among the resources summarized in Table 2, we carried out a comparative analysis of the 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 (SANCL-2014)[5]. 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 2, and, whenever possible, by searching for the given phenomena within the resources themselves.

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. However, this is frequently not the case with UGC content on social media.[6] Often,

[5] http://www.spmrl.org/sancl-posters2014.html

[6] This is not to say that there is no conventional, well-punctuated data on social media. 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

Name	Reference	Source	Language	UD-based
ATDT (UD)	(Albogamy and Ramsay, 2017)	Twitter	Arabic	yes
Hi-En-CS	(Bhat et al., 2018)	Twitter	Hindi, English (code-switch)	yes
TwitterAAE (TAAE)	(Blodgett et al., 2018)	Twitter	African-American English Mainstream American English	yes
TWITTIRÒ-UD (TWRO)	(Cignarella et al., 2019)	Twitter	Italian	yes
DWT	(Daiber and Van Der Goot, 2016)	Twitter	English	no*
W2.0	(Foster et al., 2011)	Twitter, sport forums	English	no‡
Foreebank (Frb)	(Kaljahi et al., 2015)	technical forums	English, French	no‡
Tweebank (Twb)	(Kong et al., 2014)	Twitter	English	no*
Tweebank2 (Twb2)	(Liu et al., 2018)	Twitter	English	yes
TDT	(Luotolahti et al., 2015)	various	Finnish	yes
xUGC	(Martínez Alonso et al., 2016)	various	French	yes
ITU	(Pamay et al., 2015)	n.a.	Turkish	no*
tweeDe	(Rehbein et al., 2019)	Twitter	German	yes
Postwita-UD (Pst)	(Sanguinetti et al., 2018)	Twitter	Italian	yes
FSMB	(Seddah et al., 2012)	various	French	no‡
EWT	(Silveira et al., 2014)	various	English	yes
SDT	(Wang et al., 2017)	discussion forum	Singaporean English	yes
CWT	(Wang et al., 2014)	Twitter, Sina Weibo	Chinese	no*
GUM	(Zeldes, 2017)	various	English	yes

Table 2: Overview of treebanks featuring user-generated content from the web, 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 syntactic representation.

punctuation marks may be missing, mis-applied relative to the norms of written language, or used for other communicative needs altogether (e.g. emoticons such as :-|). In some cases, no punctuation is used whatsoever.

Against this background, it is a non-trivial task to segment social media text manually, let alone automatically. Given that many social media posts by private users tend to consist of sequences of short phrases, clauses and fragments, it is understandable that many resources consider the entire tweet as a basic unit. Further, certain types of annotations deem retaining the tweet as one segment as 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 intra-sentential code-switching (see Section 4.) can be considered more appropriate at tweet level. Finally, keeping tweets as single units saves the effort needed to develop, maintain, adapt or do post-processing on an automatic sentence segmenter[7].

On the other hand, there are counterbalancing considerations that motivate performing segmentation on UGC data, among these a possible overuse of syntactic relations that define side-by-side (or run-on) sentences (e.g. `parataxis`); second, as mentioned previously, at least for some UGC data collections punctuation is found frequently enough and can be used (e.g. blog posts). Third, given that Twitter doubled its character limit for posts from 140 to 280 at the end of 2017, treating tweets as a whole might pose a usability problem for manual annotation. 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. The CoNLL-U format used in the UD project provides the means to implement this in a straight-forward manner.

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

The above segmentation policies notwithstanding, tweeDe still features the use of `parataxis` for juxtaposed clauses that are not separated by punctuation.

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 split their forum data into sentences using NLTK (Bird and Loper, 2004), with no post-corrections. Accordingly, the resource contains instances where multiple sentences are merged into one sentence due to punctuation errors such as a comma being used instead of a full stop, as in Example 1. Conversely, there are cases where a single sentence is split over multiple lines, resulting in multiple sentences (Example 2) that are not rejoined.

(1) Combofix will start, When it is scanning don’t move the mouse cursor inside the box, can cause freezing.

(2) I’m sure the devs.
can give you more details on this

Tokenization Tokenization problems in informal text include cases of various kinds that can sometimes even require

and captions.

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

a mapping effort to identify the correspondence between syntactic words and tokens. We thus may find multiple words that are merged into a single token, as in contractions[8] (Example 3) and acronyms (Example 4), or, conversely, a single syntactic word split up into more than one token (Examples 5 and 6).

(3) gonna → going to

(4) *tvtb → 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 be the one that entails their decomposition (Twb2 xUGC, tweeDe, FSMB, EWT[9]), although a few inconsistencies are found in the resulting lemmatization. Consider the contraction in Example 3. 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 decomposed and also normalized by mapping their components onto their standard form (DWT, ITU[10]), or rather leaving them unsplit (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), 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 3 would thus be annotated according to the main role played by the verb "go".
As stated above, acronyms 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.
In the opposite case, 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 as main element of the compound. 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 5) and simple spelling errors (Example 6):

(5) anti vir program → antivir program

(6) i t → it

In the first case, both tokens are tagged based on the corresponding category of the intended word, while in the second one the two tokens are treated as a spelling error and an extraneous token, respectively.
In the remaining resources, neither explicit information nor regular/consistent patterns have been found concerning the morpho-syntactic treatment of these units. For syntactic annotation, especially in the framework of dependency grammar, common practice is to attach all the remaining tokens to the one that it has been promoted as head. Finally, a distinctive tokenization strategy is adopted in ATDT with respect to at-mentions that are always split by separating the @ symbol from the username.

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 are 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 putting the short message into context. On the syntactic level, these tokens are usually not integrated, as illustrated in Example 7.

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

It is, however, possible for those tokens to fill a syntactic slot in the tweet, as shown in Example 8.

(8) *#kahvaltı zamanı*
time for #breakfast

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). Some resources adopt different strategies that do not fit into this pattern: in tweeDe, for example, at-mentions referring to user names are always considered as proper nouns while hashtags are tagged according to their respective part-of-speech, except for multi-word hashtags that are annotated as 'other' (e.g. *#WirSindHandball* "We are handball"). In Twb2, a different POS tag is assigned to at-mentions when they are used in retweets.
Similar to hashtags and mentions, links can either be annotated as symbols (TWRO, Pst), nouns (W2.0, ITU, FSMB) or 'other' (tweeDe, EWT).
Emoticons and emojis, on the other hand, are mostly classified as symbols (TWRO, Twb2, tweeDe, Pst, EWT), 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) or 'other' (Twb2[11]).
On the syntactic level, these 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 units, while

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

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

[10] In ITU, however, institutionalized and formal abbreviations are not expanded.

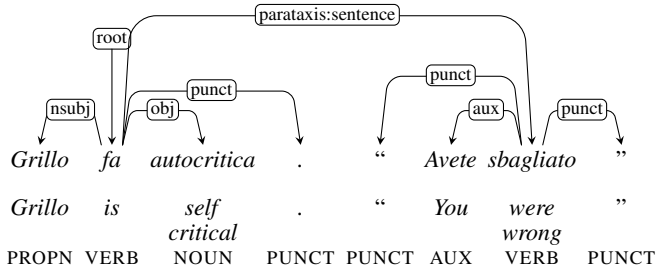
[11] Unless when considered an abbreviation of the verb "retweet", thus being annotated accordingly.

in Twb, they are not included in the syntactic analysis.
Finally, in cases where meta-tokens are syntactically integrated in the tweet, the recurring practice is to annotate them according to this role (TAAE, TWRO, DWT, Twb2 tweeDe, Pst).
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 syntactic function to these tokens.
Based on what was briefly outlined in this section, in the next section, we define an extended inventory of possible annotation issues while proposing a set of tentative guidelines for their proper representation within the UD framework.

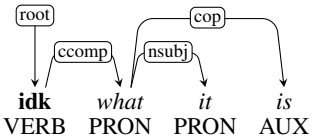
4. Towards a Unified Representation

Along with the challenges outlined in Section 3.1., here we also discuss other phenomena that can be found in user-generated text, such as code switching and disfluencies.

Sentential unit of analysis In the interest of maintaining compatibility with treebanks of standard written language, we propose splitting UGC data 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`, could be used temporarily during annotation and later serve as a pointer to identify where the tweet should be split into sentences.

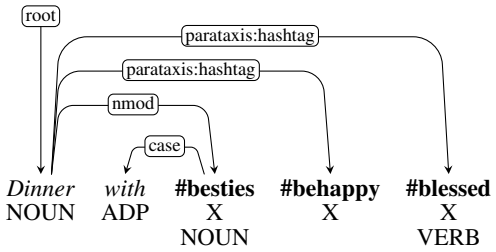


Tokenization As shown in the examples in Table 1, user-generated text can include a number of lexical and orthographic variants whose presence has repercussions on their segmentation in the first place. The basic principle adopted in UD, for which morphological and syntactic annotation is only defined at the word level (universaldependencies.org, 2019d), can sometimes clash with the complexity of these cases, whose treatment in fact has been a matter of debate within the UD community[12].
As regards the special case of contractions, this word-based segmentation principle could be easily applicable to more "traditional" ones, whose tokenization have assumed more standardized criteria over time. However, the ever-changing and dynamic nature of user-generated text makes the use of such standardized criteria mostly inadequate, or at least insufficient to cover the whole host of possible phenomena. Therefore, we propose to leave this kind of contractions unsplit, keeping the same lemma as their word form.
A distinction, however, should be drawn between multiword contractions that have reached a non-compositional status (cf. the English LOL, WTF, etc.), thus mostly functioning as discourse markers, and those phrasal contractions that actually bear a semantic and syntactic role within the sentence. For such cases, our proposal is in line with the principle proposed in Blodgett et al. (2018), where annotation has been carried out according to the root of the sub-tree of the original phrase.

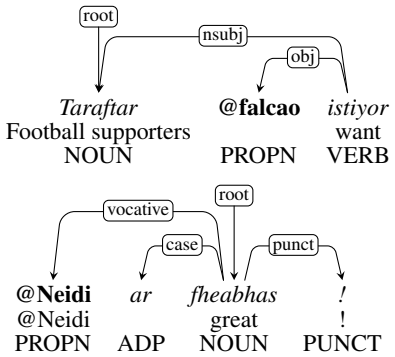


Domain-specific issues UGC includes many words and symbols with domain-specific meanings. We treat the various groups as follows:

- **Hashtags** are labeled with the X tag as their Universal POS tag (UPOS). When they play a syntactic role and are composed of single tokens, their standard UD POS tag is stored in the XPOS column, e.g., *#besties*/X/NOUN. If a hashtag comprises of multiple words, it is kept untokenised, e.g., *#behappy*/X. Syntactically integrated hashtags bear their standard dependencies. Classifying hashtags that are at the end of tweets are attached to the root with subtype `parataxis:hashtag`.



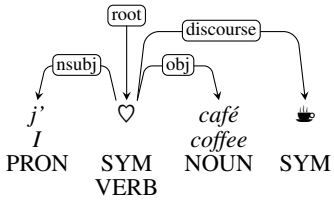
- **At-mentions** are labelled as PROPN. Their syntactic treatment is similar to hashtags: when in context they bear the actual syntactic role, otherwise they are dependent on the main predicate with the `vocative` label.



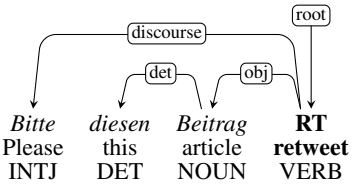
- **URLs** are tagged as SYM as per UD guidelines. They are often appended at the end of the tweet and do not seem to bear any syntactic function; thus they are attached to the root with a `dep` relation. In cases where they are syntactically integrated in the sentence, e.g. “For more information, visit [URL]”, the URL takes XPOS NOUN and is attached to its head with the appropriate dependency label.

[12] `https://github.com/UniversalDependencies/docs/issues/641`

- **Pictograms** are often used at the end of the tweets as discourse markers. In such cases they are POS-tagged as SYM and attached to the root with the `discourse` relation. But there are also cases where they replace an actual word in a syntactic context, in which case they are annotated with the XPOS tag and dependency relation of the word they substitute:



- **RT**s are originally used with at-mentions so that the Twitter interface interprets it as a retweet, as in Example 7. In such cases, their UPOS is SYM with a dependency label `parataxis` attached to the root. However they are now more commonly used as an abbreviation for *retweet* inside a tweet. The UPOS tag is NOUN or VERB depending on the usage. The dependency relation also depends on the function of the full form.



- **Markup** symbols (e.g. < >) have the UPOS SYM similar to e.g., math operators in the UD guidelines, and they are attached to the root with `dep`.

Code switching While capturing code-switching (CS) in tweets is also a motivation for a tweet-based unit of analysis (Çetinoğlu, 2016; Lynn and Scannell, 2019), it is an emerging topic of interest in NLP (Solorio and Liu, 2008; Solorio et al., 2014; Bhat et al., 2018) and thus should be captured in treebank data. CS (switching between languages) can occur on a number of levels. CS that occurs at the sentence or clause level is referred to as inter-sentential switching (INTER) as shown between English and Irish in Example 9:

(9) "*Má tá AON Gaeilge agat, úsáid í!* It's Irish Language Week."
If you have ANY Irish, use it! It's Irish Language Week.

INTER 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 those who have bi/multi-lingual followers.

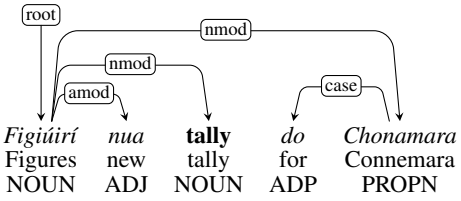
CS occuring within a clause or phrase is referred to as Intrasentential switching (INTRA). Example 10 shows INTRA switching between Italian and English

(10) "*Le proposte per l'**education** di Confindustria*"
'The proposals for the Confinustria's education'

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 11 shows the use of a Turkish verb derived from the German noun *Kopie* 'copy'.

(11) *Adamın 3-4 biyografisi var **Kopie**lenip yapıştırılmış.*
'The guy has 3-4 biographies copied and pasted.'

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 CS, the suggested UPOS is the switched token's POS – if known or meaningful – otherwise X is used (universaldependencies.org, 2019c). 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 of the conllu format, along with an indication of the CS type. As such, in Example 10, *education* would have the MISC values of `CSType=INTRA | LangID=EN`.
In terms of syntactic annotation, the UD `flat` or `flat:foreign` label is used to attach all words in a foreign string to the first token of that string – this would apply to INTER CS (universaldependencies.org, 2019a). In the cases of INTRA CS that are compositional and the grammar of the switched text is known to annotators, the dependency labels should represent the syntactic role each switched token plays.



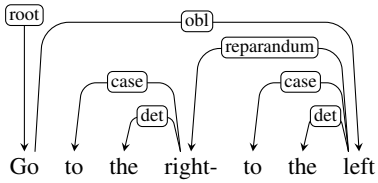
Lemmatization of CS tokens can prove difficult if a corpus contains multiple languages that annotators may not be familiar with. To enable more accurate cross-lingual studies, all switched tokens should be (consistently) lemmatized if the language is known to annotators. Otherwise the surface form should be used, allowing for more comprehensive lemmatization at a later date.

Disfluencies Similar 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 seems to be a performance error has in fact a completely different function (Rehbein, 2015). Here, repetitions, self-repair and hesitation markers are often used with humorous intent (Example 12)

(12) *Du hast den Apple Wahnsinn... äh, Spirit einfach noch nicht verstanden ;)*
'You haven't yet understood the Apple madness... uh spirit ;)'

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. An additional problem is caused by the high ambiguity resulting from fragmented texts where the context needed for determining the grammatical function of each argument is missing.

For UD, some treebanks with spoken language material exist (Lacheret et al., 2014; Dobrovoljc and Nivre, 2016; Leung et al., 2016; Øvrelid and Hohle, 2016; Caron et al., 2019) and the UD guidelines propose the following analysis for disfluency repairs (universaldependencies.org, 2019b).



This treatment, however, loses information whenever the reparandum does not have the same grammatical function as the repair, which is sometimes the case, as illustrated in Example 13. In this example from Twitter, the user plays with the homonymic forms of the German noun *Hengst* (stallion) and the verb *hängst* (hang$_{2.Ps.Sg}$).

(13) Du Hengst! äh, hängst.
You stallion! uh, hang$_{2.Ps.Sg}$.
"You stallion! uh, you're stalled."

Other open questions concern the use of hesitation markers in UGC. We propose to consider them as multi-functional discourse structuring devices and annotate them as discourse markers, attached to the root.

5. Discussion

In this last section, we propose a brief discussion on some open questions in which the nature of the phenomena described makes their encoding difficult by means of the current UD scheme.

Elliptical structures and missing elements In constituency-based treebanks of 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 traces and co-indexations at the lexical level in addition to actual nodes dominating such empty elements. The dependency syntax framework usually does not use such devices as this 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), both the subject and object of the right hand-side conjunct are annotated with the `orphan` or `remnant`[13] relations (Schuster et al., 2017). Even though the Enhanced UD scheme proposes to include a *ghost*-token (Schuster and Manning, 2016) which will be the actual governor of the right hand-side conjuncts , nothing is prescribed regarding treatment of ellipsis without

[13]Not used in UD version 2.

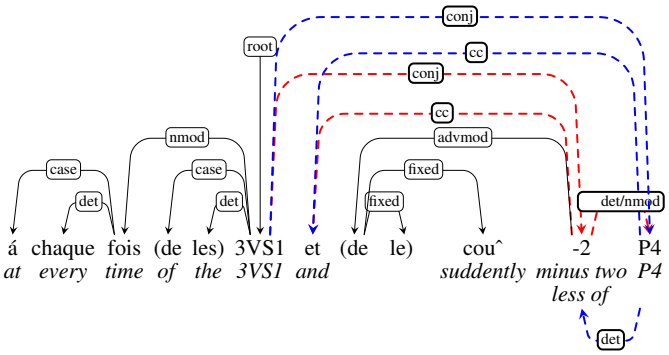


Figure 1: Pathological example with two contesting structures from two different readings of the token "-2" surrounded by at least 2 elided elements. *(Adapted to UD2.5 from (Martínez Alonso et al., 2016))*

an antecedent. Given the contextual nature of most UGC sources and their space constraints, those cases are very frequent. The problem lies in the interpetation 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.

Following from the example from Martínez Alonso et al. (2016), Figure 1 shows an attachment ambiguity caused by part-of-speech ambiguity and verb ellipsis. A natural ellipsis recovery of the example shown in Figure 1 would read as "Every time **there are** 3VS1, and suddenly **I have** -2 P4". The token "3VS1" stands for "3 versus 1", namely an uneven combat setting, and 'P4" refers to the character's protection armour. The token "-2" allows for more than one analysis. The first analysis is the simple reading as number, complementing the noun "P4". A second analysis treats "-2" as a transcription of *moins de* (less than, less of), which would be the preferred analysis where the verb recovery holds. This example shows the interplay between frequent ellipses, ergographic phenomena and the need for domain knowledge in user-generated data.

Other tokenization issues Another pending issue is the one related to the treatment of over-splitting cases (see the examples in Table 1). The UD scheme already provides for the use of the `goeswith` relation, which was introduced to identify cases of erroneously split words from badly edited texts; nevertheless, their lemmatization and POS still remain a controversial point, for which it is necessary to find a common standard.

6. Conclusion

In this paper we addressed the question of the annotation of user-generated texts from web and social media, proposing, in the context of Universal Dependencies, a unified scheme for their coherent treatment across different languages.

The variety and complexity of the treated phenomena sometimes makes their adequate representation non-trivial by means of an already existing scheme, such as UD. 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

The work of C. Bosco is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01). M. Sanguinetti is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) and by the project "*Studi e Ricerche su Sistemi Conversazionali Intelligenti*" (CENF_CT_RIC_19_01). 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).

7. References

Bhat, Irshad and Bhat, Riyaz A. and Shrivastava, Manish and Sharma, Dipti. (2018). *Universal Dependency Parsing for Hindi-English Code-Switching*.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Blodgett, S. L., Wei, J. T. Z., and 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 (Lng Papers)*, volume 1, pages 1415–1425. ACL.

Bosco, C., Tamburini, F., Bolioli, A., and Mazzei, A. (2016). Overview of the EVALITA 2016 Part Of Speech tagging on TWitter for ITAlian task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR.

Caron, B., Courtin, M., Gerdes, K., and Kahane, S. (2019). A surface-syntactic UD treebank for Naija. In *Proceedings of The 18th International Workshop on Treebanks and Linguistic Theories*, TLT'19, pages 13–24. ACL.

Çetinoğlu, Ö. and Çöltekin, Ç. (2016). Part of speech tagging of a Turkish-German code-switching corpus. In *Proceedings of the tenth Linguistic Annotation Workshop (LAW-X)*, pages 120–130. ACL.

Çetinoğlu, Ö. (2016). A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4215–4220. ELRA.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1566–1573. ELRA.

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*, pages 359–369. ACL.

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*, pages 381–384. ACL.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and 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*, pages 42–47. ACL.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1001–1012.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 295–301. ELRA.

Leung, H., Poiret, R., Wong, T.-s., Chen, X., Gerdes, K., and Lee, J. (2016). Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 20–29.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and 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, Volume 1 (Long Papers)*, pages 965–975. ACL.

Lynn, Teresa and Scannell, Kevin. (2019). *Code-switching in Irish tweets: A preliminary analysis*. European Association for Machine Translation.

Lynn, Teresa and Scannell, Kevin and Maguire, Eimear. (2015). *Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets*. ACL.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. volume 19, pages 313–330.

Martínez Alonso, H., Seddah, D., and 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*, volume 1, pages 127–137.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL 2013*, pages 380–390.

Petrov, S. and McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on . . .* , number C.

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), pages 665–670. ELRA.

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)*, pages 12–21. ACL.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the 10th Language Resource and Evalutation Conference (LREC 2016)*, pages 2371–2378. ELRA.

Schuster, S., Lamm, M., and Manning, C. D. (2017). Gapping constructions in Universal Dependencies v2. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 123–132. ACL.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and 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*, pages 2441–2458.

Solorio, T. and 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)*, pages 1051–1060. ACL.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the CodeSwitch Workshop*.

universaldependencies.org. (2019a). Annotation of foreign strings in the Universal Dependencies guidelines. `https://universaldependencies.org/cs/dep/flat-foreign.html`. Accessed: 2019-11-28.

universaldependencies.org. (2019b). Annotation of speech repair in the Universal Dependencies guidelines. `https://universaldependencies.org/u/dep/reparandum.html`. Accessed: 2019-11-28.

universaldependencies.org. (2019c). Pos-tagging of foreign tokens in the Universal Dependencies guidelines. `https://universaldependencies.org/u/pos/X.html`. Accessed: 2019-11-28.

universaldependencies.org. (2019d). Tokenization and Word Segmentation guidelines. `https://universaldependencies.org/u/overview/tokenization.html`. Accessed: 2019-12-02.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1579–1585. ELRA.

8. Language Resource References

Albogamy, Fahad and Ramsay, Allan. (2017). *Universal dependencies for Arabic tweets*.

Bhat, Irshad and Bhat, Riyaz A. and Shrivastava, Manish and Sharma, Dipti. (2018). *Universal Dependency Parsing for Hindi-English Code-Switching*.

Blodgett, Su Lin and Wei, Johnny Tian Zheng and O'Connor, Brendan. (2018). *Twitter universal dependency parsing for African-American and mainstream American English*.

Cignarella, Alessandra Teresa and Bosco, Cristina and Rosso, Paolo. (2019). *Presenting TWITTIRO-UD : An Italian Twitter Treebank in Universal Dependencies*.

Daiber, Joachim and Van Der Goot, Rob. (2016). *The Denoised Web Treebank: Evaluating dependency parsing under noisy input conditions*.

Foster, Jennifer and Çetinoğlu, Özlem and Wagner, Joachim and Roux, Joseph Le and Nivre, Joakim and Hogan, Deirdre and van Genabith, Josef. (2011). *From News to Comment: Resources and Benchmarks for Parsing the Language of Web 2.0*.

Kaljahi, Rasoul and Foster, Jennifer and Roturier, Johann and Ribeyre, Corentin and Lynn, Teresa and Le Roux, Joseph. (2015). *Foreebank: Syntactic analysis of customer support forums*. Number September.

Kong, Lingpeng and Schneider, Nathan and Swayamdipta, Swabha and Bhatia, Archna and Dyer, Chris and Smith, Noah A. (2014). *A dependency parser for tweets*.

Liu, Yijia and Zhu, Yi and Che, Wanxiang and Qin, Bing and Schneider, Nathan and Smith, Noah A. (2018). *Parsing Tweets into Universal Dependencies*.

Luotolahti, Juhani and Kanerva, Jenna and Laippala, Veronika and Pyysalo, Sampo and Ginter, Filip and Studies, Translation. (2015). *Towards Universal Web Parsebanks*. Number Depling.

Lynn, Teresa and Scannell, Kevin. (2019). *Code-switching in Irish tweets: A preliminary analysis*. European Association for Machine Translation.

Lynn, Teresa and Scannell, Kevin and Maguire, Eimear. (2015). *Minority Language Twitter: Part-of-Speech Tagging and Analysis of Irish Tweets*. ACL.

Martínez Alonso, Héctor and Seddah, Djamé and Sagot, Benoît. (2016). *From Noisy Questions to Minecraft Texts: Annotation Challenges in Extreme Syntax Scenarios*.

Nivre, Joakim and Marie-Catherine de Marneffe and Filip Ginter and Yoav Goldberg and Jan Hajič and Christopher D. Manning and Ryan T. McDonald and Slav Petrov and Sampo Pyysalo and Natalia Silveira and Reut Tsarfaty and Daniel Zeman. (2016). *Universal Dependencies v1: A Multilingual Treebank Collection*.

Pamay, Tuğba and Sulubacak, Umut and Torunoğlu-Selamet, Dilara and Eryiğit, Gülşen. (2015). *The Annotation Process of the ITU Web Treebank*.

Rehbein, Ines and Ruppenhofer, Joseph and Do, Bich-Ngoc. (2019). *tweeDe – A Universal Dependencies treebank for German tweets*.

Sanguinetti, Manuela and Bosco, Cristina and Lavelli, Alberto and Mazzei, Alessandro and Antonelli, Oronzo and Tamburini, Fabio. (2018). *Postwita-UD: An Italian twitter treebank in universal dependencies*.

Seddah, Djamé and Sagot, Benoit and Candito, Marie and Mouilleron, Virginie and Combet, Vanessa. (2012). *The French Social Media Bank: A Treebank of Noisy User Generated Content*. ACL.

Silveira, Natalia and Dozat, Timothy and De Marneffe, Marie Catherine and Bowman, Samuel R. and Connor,

Miriam and Bauer, John and Manning, Christopher D. (2014). *A gold standard dependency corpus for English*. ELRA.

Wang, William Yang and Kong, Lingpeng and Mazaitis, Kathryn and Cohen, William W. (2014). *Dependency parsing for weibo: An efficient probabilistic logic programming approach*.

Wang, Hongmin and Zhang, Yue and Leonard Chan, Guang Yong and Yang, Jie and Chieu, Hai Leong. (2017). *Universal dependencies parsing for colloquial Singaporean English*.

Amir Zeldes. (2017). *The GUM Corpus: Creating Multilayer Resources in the Classroom*.