CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing

Amir More,[1,*] Özlem Çetinoğlu,[2] Çağrı Çöltekin,[3] Nizar Habash,[4] Benoît Sagot,[5] Djamé Seddah,[5,6] Dima Taji,[4] Reut Tsarfaty[1]

[1]Open University of Israel, [2]Institut fur Maschinelle Sprachverarbeitung, [3]University of Tübingen, [4]New York University Abu Dhabi, [5]INRIA, [6]Université Paris Sorbonne

habeanf@gmail.com, ozlem@ims.uni-stuttgart.de, ccoltekin@sfs.uni-tuebingen.de, nizar.habash@nyu.edu benoit.sagot@inria.fr, djame.seddah@inria.fr, dima.taji@nyu.edu, reutts@openu.ac.il

Abstract

Following the development of the *universal dependencies* (UD) framework and the CoNLL 2017 Shared Task on end-to-end UD parsing, we address the need for a universal representation of morphological analysis which on the one hand can capture a range of different alternative morphological analyses of surface tokens, and on the other hand is compatible with the segmentation and morphological annotation guidelines prescribed for UD treebanks. We propose the *CoNLL universal lattices* (CoNLL-UL) format, a new annotation format for word lattices that represent morphological analyses, and provide resources that obey this format for a range of typologically different languages. The resources we provide are harmonized with the two-level representation and morphological annotation in their respective UD v2 treebanks, thus enabling research on universal models for morphological and syntactic parsing, in both pipeline and joint settings, and presenting new opportunities in the development of UD resources for low-resource languages.

Keywords: Morphology, Universal Dependencies, Morphological Analysis, Morphological Ambiguity

1. Introduction

The development of the *universal dependencies* (UD) framework and its treebank collection (Nivre et al., 2016; Nivre et al., 2017) follows many shared tasks and multilingual evaluation campaigns in which the linguistic representation schemes across different languages vary (Buchholz and Marsi, 2006; Nivre et al., 2007; Seddah et al., 2013; Butt et al., 2002; Zeman et al., 2012). The UD treebanks collection, in contrast, obeys a single set of annotation guidelines, and respects the discrepancies between surface input tokens and the output nodes in the syntax trees (a.k.a., the *two-level representation* principle.)[1]

The UD initiative has paved the way to the development of cross-lingual models for word segmentation, part-of-speech tagging and dependency parsing (Straka and Straková, 2017), as well as cross-linguistic typological investigations (Futrell et al., 2015). Recently, the CoNLL 2017 Shared Task on Multilingual UD Parsing (Zeman et al., 2017), which used a variant of the UD datasets from the UD v2.0 release, introduced a truly end-to-end parsing setting: participants had to parse raw texts into dependency trees, implying the initial phases of sentence tokenisation, word segmentation, part-of-speech tagging and morphological annotation (if their parsing models required morphological information in their input).

The UD annotation scheme provides guidelines for unambiguously annotating the morphological and syntactic levels of representation of natural language sentences, but it does not provide means to formally capture a range of potential analyses, and in particular the *ambiguous morphological analyses*, of raw surface tokens. Lexical resources that capture this ambiguity, in the form of morphological analyzers or existing lexicons, are available for many of the participating treebanks,[2] but they are far from useful for UD parsing, for two main reasons: (a) existing lexical resources rely on variety of formats and underlying theories that are incompatible with the UD morphological scheme, and (b) the morpho-syntactic interface assumed by these tools' representation is often incompatible with the respective trees in the UD treebanks.

To fill this gap, we propose an annotation format which allows for capturing the full range of potentially-ambiguous morphological analyses of raw surface tokens, and at the same time is compatible with UD treebanks in form and function and respects its two-level representation principle. We name this format "CoNLL-UL", in which UL stands for a *universal lattice* structure meant for formally capturing morphological ambiguity. In addition, we provide adaptations of existing morphological analyzers and lexica to our proposed format, making a wide range of CoNLL-UL resources freely available for the community.

Our contribution is hence many-fold. We first introduce a UD-compatible annotation format that is suitable for representing competing morphological analyses — each of which consisting of word segmentation, POS tagging and morphological features — for tokens or token sequences. Secondly, we provide a set of lexical resources for broad-coverage morphological analysis obeying this CoNLL-UL format, based on different sources: (i) For Arabic, Hebrew, and Turkish, morphologically rich-and-heavily-ambiguous languages, we developed or adapted morphological analyzers such that their output is in the CoNLL-UL format and subscribe to the word segmentation and morphological annotation theories of their respective UD v2 treebanks (Section 3.1). (ii) For several languages with less-complex morphology, we converted the output of existing freely-

* Corresponding author.

[1] http://universaldependencies.org/

[2] universaldependencies.org/conll17/data.html

FROM	TO	FORM	LEMMA	UPOS	CPOS	FEATURES	MISC	ANCHORS
0	1	*her*	her	DET	_	Definite=Def	_	goldid=1
1	2	*şey*	şey	NOUN	_	Case=Nom|Number=...	_	goldid=2
2-4	güzeldi	_						
2	3	*güzel*	güzel	ADJ	_	Case=Nom|Number=...	_	goldid=3
3	4	*di*	i-	VERB	_	Aspect=Perf...	_	goldid=4

(a) Morphological analysis of the Turkish phrase `her şey güzeldi` (everything was beautiful) in the CoNLL-UL format.

ID	FORM	LEMMA	UPOS	CPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	*her*	her	DET	_	Definite=Def	2	det	_	_
2	*şey*	şey	NOUN	_	Case=Nom|Numbe...	3	nsubj	_	_
3-4	güzeldi								
3	*güzel*	güzel	ADJ	_	Case=Nom|Numbe...	0	root	_	_
4	*di*	i-	VERB	_	Aspect=Perf...	3	cop	_	_

(b) The CoNLL-U representation of the Turkish phrase `her şey güzeldi .`

Table 1: The relationship between CoNLL-UL and CoNLL-U for a linear (unambiguous) lattice.

available morphological lexicons to the CoNLL-UL format. In that way, these lexica can be used for generating all possible analyses of a given token in the required lattice structure, provided it is known to the lexicon (see Section 3.2). (iii) For languages that do not have freely-available (if any) lexical resources, we provide a rudimentary tool to induce a CoNLL-UL lexicon from a UD treebank, which can be used as a baseline broad-coverage morphological analyzer.

We propose that CoNLL-UL will serve as a complement to the CoNLL-U format, and likewise, that conforming lexical resources will complement the respective treebanks in the UD treebank collection. CoNLL-UL will help researchers exchange language resources and tools at the morphological level, therefore improving their systems and allowing for proper cross-lingual comparison. Moreover, universal access to broad-coverage lexical resources harmonized with the UD treebanks scheme will pave the way for (otherwise infeasible) research on joint models for universal morpho-syntactic parsing.

We detail our proposal, motivated by the CoNLL-2017 shared task, in Section 2, and describe the resources we make available in Section 3. We cover related work and contrast particularities of broad-coverage morphological analyses with the morphology specifically annotated in the UD treebanks, and discuss the limitations of UD morpho-syntax — to be potentially addressed by the UD community in the future — in Section 4, and we conclude with a summary of our contributions in Section 5.

2. The CoNLL-UL Proposal

In this section we detail our proposed universal annotation format for morphological ambiguity, which represents the competing analyses of a given source token.

Throughout this paper, we use the following terminology. A *source token* is a sequence of characters in the raw input text, segmented from surrounding characters by conventional, typographic (non linguistic) criteria. It corresponds to "tokens" in the UD model. A *tree token* is a lexical unit output by a morphological analyzer that is meant to serve as a leaf node in syntactic structures. It corresponds to "words" in the UD model. The CoNLL-UL annotation format scheme is similar to the SPMRL lattice format, where every lattice represents a surface token, and each edge in the lattice represents a tree token with the following properties:

FROM: Index of the outgoing vertex of the edge;

TO: Index of the incoming vertex of the edge;

FORM: Tree token (word form or punctuation mark);

LEMMA: Lemma or stem of the word form; underscore if not available;

UPOSTAG: Universal POS tag;

XPOSTAG: Language-specific POS tag; underscore if not available;

FEATS: List of morphological features from the universal feature inventory or from a pre-defined language-specific extension; underscore if not available;

MISC: Any other annotation related to morphology; underscore if not available;

ANCHORS: Identifiers linking to specific disambiguation if needed; underscore if not available

We also borrow from the UD format properties for specifying a surface token spanning multiple tree tokens, respecting the two-level representation principle:

RANGE: Start and end vertex ids in the tree token lattice;

TOKEN: Source token;

MISC: Any other token-level annotation (e.g. spelling issue or canonical representation); underscore if not available

We generalise this notation for dealing with source tokens that are different from the corresponding tree token in the case of 1-to-1 mappings, using ranges of length one.

Note that as opposed to CONLL-U, integer indices in the FROM, TO and RANGE fields index vertices in a lattice, and not tree tokens. Nevertheless, in the case of a linear lattice (a lattice with only one path) as shown in Tables (1a) and (1b), the CoNLL-U representation can be directly obtained as follows:

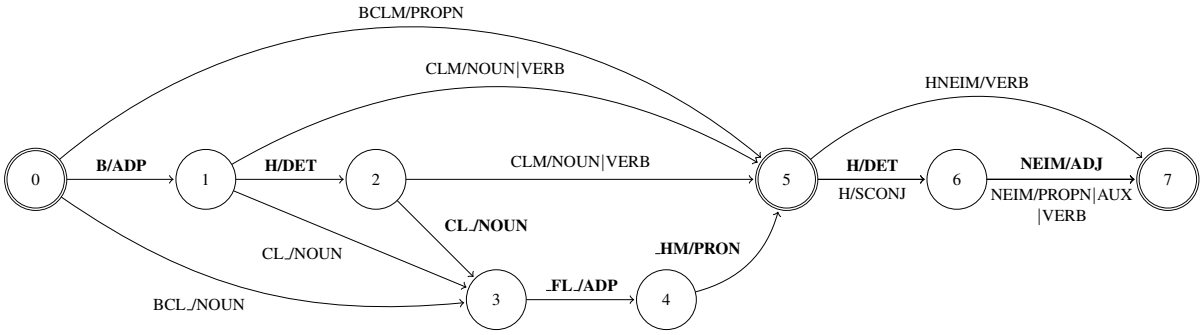

Figure 1: Lattice of possible analyses in terms of transliterated tree token sequences for the first two Hebrew source tokens of the phrase בצלם הנעים של העצים (BCLM HNEIM FL HECIM, meaning "in the pleasant shadow/shade of the trees", using the transliteration scheme of Sima'an et al. (2001)). The correct disambiguation of each source token, in the context of the phrase, is highlighted in bold. Lemma and morphological property ambiguity are not shown for brevity; '|' indicates part-of-speech ambiguity

FROM	TO	FORM	LEMMA	UPOS	CPOS	FEATURES	MISC	ANCHORS
0-5	בצלם	_						
0	5	*BCLM*	BCLM	PROPN	_	_	_	_
0	1	*B*	B	ADP	_	_	_	goldid=1
0	3	*BCL_*	BCL	NOUN	_	Gender=Masc|Number=Sing	_	_
1	2	*H*	H	DET	_	PronType=Art	_	goldid=2
1	3	*CL_*	CL	NOUN	_	Gender=Masc|Number=Sing	_	_
2	3	*CL_*	CL	NOUN	_	Gender=Masc|Number=Sing	_	goldid=3
2	5	*CLM*	CILM	VERB	_	Gender=Masc|Number=Si…	_	_
2	5	*CLM*	CILM	VERB	_	Gender=Masc|Mood=Imp…	_	_
2	5	*CLM*	CLM	NOUN	_	Definite=Cons|Gender=Mas…	_	_
2	5	*CLM*	CLM	NOUN	_	Gender=Masc|Number=Sing	_	_
3	4	*_FL_*	FL	ADP	_	_	_	goldid=4
4	5	*_HM*	ANI	PRON	_	Gender=Masc|Number=Plur|Person=3	_	goldid=5
5-7	הנעים	_						
5	7	*HNEIM*	HNEIM	VERB	_	Gender=Masc…	_	_
5	6	*H*	H	DET	_	PronType=Art	_	goldid=6
5	6	*H*	H	SCONJ	_	_	_	_
6	7	*NEIM*	NEIM	ADJ	_	Gender=Masc|Number=Sing	_	goldid=7
6	7	*NEIM*	NEIM	ADJ	_	Definite=Cons|Gender=…	_	_
6	7	*NEIM*	NEIM	ADV	_	Polarity=Neg	_	_
6	7	*NEIM*	NEIM	AUX	_	Gender=Masc|Number=Sing…	_	_
6	7	*NEIM*	NEIM	PROPN	_	_	_	_
6	7	*NEIM*	NEIM	VERB	_	Gender…|Tense=Part|VerbForm=Part	_	_
6	7	*NEIM*	NEIM	VERB	_	Gender…|VerbForm=Part	_	_

Table 2: The CoNLL-UL representation of the lattice shown in Figure 1. Tree tokens and lemmas are transliterated here for the convenience of the reader. Some morphological features strings are shortened for brevity. Note the reuse of columns for source token span lines as in CoNLL-U, where the FROM column is a range, the TO column is a source token, and the FORM column is a misc field set to underscore when empty

- ignore the "FROM" column and the "MISC" column for source tokens,
- ignore source token specifications with ranges of length 1,
- ignore the "ANCHORS" column, and
- increment by one the starting vertex id for ranges with length greater than 1.[3]

We now illustrate this format on an example. Let us consider the Hebrew source token sequence בצלם הנעים (transliterated as BCLM HNEIM).[4] Figure 1 displays the lattice of possible analyses in terms of tree tokens, as could be output by a non-deterministic morphological analyzer. This lattice illustrates two ambiguity types: (i) *morphological segmentation ambiguity*, which is directly visible in the different path lengths in the lattice, (ii) *morphological tagging ambiguity*, visible in the two analyses provided for the tree token **צלם** (*CLM*). The corresponding CoNLL-UL representation is provided in Table 2.

[3]Only the starting vertex id is incremented since ranges in CoNLL-UL correspond to vertices in lattices, of which there are two (start and end) for each possible tree token, whereas ranges in CoNLL-U correspond to tree tokens themselves. Lattice vertices are 0-indexed, therefore the starting vertex id is incremented, rather than the ending vertex id decremented.

[4]Henceforth, source tokens will be typeset in `typewriter` and tree tokens in *italics* throughout the paper.

FROM	TO	FORM	LEMMA	UPOS	CPOS	FEATURES	MISC
0	1	*encodent*	encoder	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres…	_
0-2	auxquels	_					
0	1	*à*	à	ADP	_	_	_
1	2	*lesquels*	lequel	PRON	_	Gender=Masc|Number=Plur	_

Table 3: Two entries resulting from the conversion of the L*efff* in the CoNLL-UL format.

Let us now consider the correct segmentation ב ה צל של הם (*B H CL FL HM*, meaning in-the-shadow-of-them) of the source token בצלם (BCLM) in context. These segments would form the syntactic words of a CoNLL-U file describing UD trees. We relate the CoNLL-UL morphologically ambiguous files to CoNLL-U unambiguous files by *anchoring* lattice arcs with their syntactic identifiers in the ANCHORS field. This simplifies the process of merging one or more morphological disambiguations, a necessary step for evaluation of the prediction and for joint morpho-syntactic processing.

3. Morphological Analysis with CoNLL-UL

We have developed a set of resources and tools that can perform morphological analysis and output it in the CoNLL-UL format. Morphological analysis can be performed either online, using for instance finite-state or statistical/neural models, and/or based on lexical resources. In this section, we first briefly describe several CoNLL-UL-compatible morphological analyzers we developed for languages such as Arabic, Hebrew and Turkish; we then sketch how we converted existing lexicons into the CoNLL-UL format, which can be used straightforwardly as the basis for simple morphological analyzers.

The Hebrew and Turkish morphological analyzers, and lexicons mentioned in this paper are freely available. The Arabic morphological analyzer requires a license, which can be acquired by following the instructions at the provided link. In addition, we apply our analyzers to existing UD treebanks, and provide the resulting CoNLL-UL analyses to the community.[5]

3.1. Morphological Analyzers for MRLs

As our first contribution to the bootstrapping of universal morphological resources, we provide here adaptations of morphological analyzers for three Morphologically Rich Languages (MRLs): Arabic, Hebrew, and Turkish. For these morphological analyzers, their pre-existing morphological analyses adhere to schemes that differ from those employed in the respective UD treebanks. These discrepancies are due to differences between the morphological theories adopted by the UD treebanks developers and those employed by the developers of the morphological analyzers. Therefore, the adapted resources we provide are non-trivial to obtain, and required careful alignment of the morpho-syntactic analyses with their UD treebank counterparts.

For Arabic, we adapted the morphological analyzer used in MADAMIRA (Pasha et al., 2014), which is built on top of the databases of SAMA (Maamouri et al., 2010) to output morphology that adheres to the UD Arabic treebank (Taji et al., 2017).[6] The Arabic UD treebank, as with other Arabic treebanks, uses the Penn Arabic treebank tokenization scheme (Maamouri et al., 2004) which segments all proclitics and enclitics except for the definite article. It is worth noting that the format we propose here is independent of the specifics of this tokenization scheme and it can be used with a number of other schemes (Habash, 2010).

For Hebrew, we used the HEBLEX morphological analyzer of More and Tsarfaty (2016), based on the BGU Lexicon (Itai and Wintner, 2008), adapted to the UD Hebrew treebank.[7] We only modified the HEBLEX SPMRL lattices format to follow the proposed CoNLL-UL format, as the HEBLEX annotations have already been adapted to the treebank counterpart (More and Tsarfaty, 2017).

For Turkish, we developed a new morphological analyzer based on TRmorph (Çöltekin, 2010).[8] The analyzer follows the segmentation and morphological analysis scheme of the UD Turkish treebank v2.0 (Sulubacak et al., 2016) and Turkish-PUD treebank (Zeman et al., 2017). These treebanks have employed a different segmentation approach compared to the METU-Sabancı Turkish Treebank (Oflazer et al., 2003). In addition, form and lemma representations, POS tags and morphological tag sets have changed. The existing morphological analyzers are not compatible with this new representation. Thus we introduce a finite-state implementation that complies with the UD v2.0 scheme.

On top of that, for languages in the UD treebanks collection that may not have existing lexical resources and/or morphological analyzers publicly available, we adapted a data-driven rudimentary morphological analyzer (More and Tsarfaty, 2017) that induces a morphological lexicon from existing UD treebanks which provide broad-coverage morphological analyses and adhere to the proposed CoNLL-UL format. The analyzer can use the induced lexicon, as well as the converted morphological lexicons below, to provide analyses of input text in the CoNLL-UL format.

3.2. Converted Morphological Lexicons

As a complement to the CoNLL-UL-compatible analyzers described above, we have created a set of 53 CoNLL-UL-compatible morphological lexicons covering 38 languages, based on existing freely available resources.[9] The source lexicons, the conversion processes and the resulting inventory of freely available CoNLL-UL lexicons are described

[5] https://conllul.github.io

[6] https://camel.abudhabi.nyu.edu/calima-star/

[7] https://github.com/habeanf/yap

[8] https://github.com/coltekin/TRmorph/tree/trmorph2

[9] http://pauillac.inria.fr/~sagot/udlexicons.html

in (Sagot, 2018).[10] Here we only provide in Table 3 two examples converted from the Lefff, the Alexina lexicon for French. The first one illustrates the 1-to-1 case, with an entry converted from the following original entry:

encodent encoder v P3p,

which includes the wordform (i.e. the [source and tree] token) *encodent* 'encode$_{\text{3pl.pres.ind}}$', its lemma, its Lefff POS and its Lefff morphosyntactic tag. The other example illustrates the 1-to-m case with the source token `auxquels`, which is analyzable as reflecting the sequence of two tree tokens *à lesquels* 'to which'.

4. Related Work and Perspective

Our work overlaps somewhat with previous proposals, in particular the ISO norm "Morphosyntactic Annotation Framework" (hereafter MAF, (Clément and Villemonte de La Clergerie, 2005)).[11] In principle, MAF allows for the representation of any analysis represented in CoNLL-UL, whereas not every analysis in MAF can be represented in CoNLL-UL. This is because CoNLL-UL is intentionally coupled to CoNLL-U in both form and function: first, we use the CoNLL-U flat, tab-delimited file format for consistency and ease of use; second, we intentionally impose the same restrictions on CoNLL-UL morphology that UD itself is restricted to, such that these two resources maintain harmony. As a result of the latter, we can maintain a two-way compatibility promise: every morphological disambiguation in a UD v2 treebank can be represented as a CoNLL-UL lattice, and every possible path in a CoNLL-UL lattice can serve as the syntactic words of a UD-annotated tree. Thus, we ease the burden on morphological and syntax parser research and development, such that they are relieved of adapting lexical resources (or their analyses) to UD-compliant morphology.

The representation scheme for lattices used by the SPMRL shared task datasets (Seddah et al., 2013) and which were introduced by (Tsarfaty et al., 2012; Tsarfaty, 2013),[12] allowed for annotating morphological ambiguity of these same languages. Seeker and Çetinoğlu (2015) extended the SPMRL representation to accommodate marking the gold and optionally a predicted morphological analysis. Our proposal extends the latter with two additions: (i) we use the UD convention of specifying a surface token spanning multiple tree tokens; and (ii) we allow the specification of multiple anchors relating lattice arcs to tree tokens, for possibly grounding more than one syntactic tree (i.e., a forest) in the morphological lattice.

Since we wanted to maintain compatibility with the current version of CoNLL-U, our CoNLL-UL proposal has some limitations. First, although it fully covers 1-to-1 and 1-to-m mappings between source and tree tokens, it only covers n-to-1 mappings via "words with spaces" or special 'GoesWith' dependencies, and does not cover n-to-m mappings. Yet such cases do occur in many languages.[13] It is especially true when taking into account noisy user-generated content and speech productions. Moreover, our proposal does not cover all types of lattices, including cases that cannot be covered with only one set of indices, as used in CoNLL-U and CoNLL-UL. There are lattice shapes which require independent mechanisms for indexing source and tree tokens (or, rather, states in a tree token lattice).[14] However, addressing all these cases would require a format that could not be directly compatible with the (current version of the) UD format/model. We therefore leave open the complete investigation of these issues for future work.

5. Conclusion

Although lexical resources for morphological analysis exist for many languages, they respect varied approaches to morphology. In the context of the CoNLL 2017 Shared Task, the morphological annotation in UD treebanks has not been harmonized with these existing lexical resources, making it hard to develop joint morpho-syntactic parsers.

We propose the CoNLL-UL annotation for morphological ambiguity and provide adapted resources harmonized with existing UD resources. CoNLL-UL addresses the need for a UD/CoNLL-U interface to existing lexical resources, and our adapted resources provide a good starting point, with at least three important MRLs: Arabic, Hebrew, and Turkish. We also adapt morphological lexicons in the Apertium, Giellatekno, and Alexina frameworks, providing lexical resource coverage for numerous languages. For languages without lexical resources, we provide a baseline solution using the UD treebanks to induce data-driven lexica.

We suggest our proposal, together with the adapted resources and tools we provide, form to complement the set of UD treebanks, and facilitate research and development of cross-lingual morphological and syntactic parsing.

6. Acknowledgements

The authors would like to thank Joakim Nivre for his insightful comments. The work of the first and last authors is supported via research grants by the European Research Council (ERC-StG 677352) and the Israeli Science Foundation (ISF 1739/26), for which we are grateful. Özlem Çetinoğlu is supported by the Deutsche Forschungsgemeinschaft (DFG) via the SFB 732, project D2. Benoît Sagot and Djamé Seddah were partly funded by the French ANR projects ParSiTi (ANR-16-CE33-0021 and SoSweet (ANR-15-CE38-0011-01), as well as by the Program "Investissements d'avenir" ANR-10-LABX-0083 (Labex EFL).

[10] The lexical information is represented in the CoNLL-UL format, with a minor adaptation; as a lexicon is not a collection of sentences, but a collection of entries, each entry is annotated as a separate sentence, but "sentence" boundaries are not included.

[11] Official page on the ISO website: `https://www.iso.org/fr/standard/51934.html?browse=tc`. Freely available earlier working draft: `http://atoll.inria.fr/~clerger/MAF/html/index.html`.

[12] `http://www.spmrl.org`.

[13] See for instance in French the sequence of source tokens `près du Mans` 'near Le Mans' and its analysis in terms of tree tokens *près_de Le_Mans*, where tree token *près_de* "corresponds" to source token `près` and the "first half" of the source token `du`, and tree token *Le_Mans* "corresponds" to the "second half" of source token `du` and source token `Mans`.

[14] See for instance the non canonical French source token sequence `c t`, which can be analysed (at least) in two ways, namely as the tree token sequence *c' était* 'it was' and the tree token sequence *c' est tes* 'it is your$_{\text{plur.}}$'.

7. Bibliographical References

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, New York City, USA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. In *Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation - Volume 15*, COLING-GEE '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 820–827.

Clément, L. and Villemonte de La Clergerie, E. (2005). MAF: a morphosyntactic annotation framework. In *proc. of the 2nd Language & Technology Conference (LT'05)*, pages 90–94, Poznan, Poland.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Itai, A. and Wintner, S. (2008). Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, March.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Standard arabic morphological analyzer (sama) version 3.1. *Linguistic Data Consortium, Catalog No.: LDC2010L01*.

More, A. and Tsarfaty, R. (2016). Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 337–348, Osaka, Japan, December. The COLING 2016 Organizing Committee.

More, A. and Tsarfaty, R. (2017). Universal joint morph-syntactic processing: The open university of israel's submission to the conll 2017 shared task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 253–264.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Candito, M., Cebiroğlu Eryiğit, G., Celano, G. G. A., Chalub, F., Choi, J., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Erjavec, T., Farkas, R., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hà Mỹ, L., Haug, D., Hladká, B., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kotsyba, N., Krek, S., Laippala, V., Lê Hồng, P., Lenci, A., Ljubešić, N., Lyashevskaya, O., Lynn, T., Makazhanov, A., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Mori, S., Moskalevskyi, B., Muischnek, K., Mustafina, N., Müürisep, K., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Real, L., Rituma, L., Rosa, R., Saleh, S., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Tsarfaty, R., Tyers, F., Uematsu, S., Uria, L., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Oflazer, K., Say, B., Hakkani-Tür, D. Z., and Tür, G. (2003). Building a Turkish Treebank. In Anne Abeille, editor, *Building and Exploiting Syntactically-annotated Corpora.* Kluwer Academic Publishers, Dordrecht.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Sagot, B. (2018). A Multilingual Collection of CoNLL-U-compatible Morphological Lexicons. In *In proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and Villemonte de La Clergerie, E. (2013). Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth*

Workshop on Statistical Parsing of Morphologically-Rich Languages, pages 146–182, Seattle, Washington, USA.

Seeker, W. and Çetinoğlu, O. (2015). A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *TACL*, 3:359–373.

Sima'an, K., Itai, A., Winter, Y., Altman, A., and Nativ, N. (2001). Building a tree-bank of Modern Hebrew text. *Traitment Automatique des Langues*, 42(2).

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Sulubacak, U., Gokirmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., and Eryiğit, G. (2016). Universal dependencies for turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Taji, D., Habash, N., and Zeman, D. (2017). Universal dependencies for arabic. *WANLP 2017 (co-located with EACL 2017)*, page 166.

Tsarfaty, R., Nivre, J., and Andersson, E. (2012). Joint evaluation of morphological segmentation and syntactic parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 6–10, Jeju Island, Korea, July. Association for Computational Linguistics.

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of stanford dependencies. In *ACL (2)*, pages 578–584.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). Hamledt: To parse or not to parse? In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.