

EMAD: A Bridge Tagset for Unifying Arabic POS Annotations

Omar Kallas,¹ Go Inoue,^{1,2} Nizar Habash¹

¹Computational Approaches to Modeling Language (CAMEL) Lab, New York University Abu Dhabi

²Mohamed bin Zayed University of Artificial Intelligence

{ok764, nizar.habash}@nyu.edu

go.inoue@mbzuai.ac.ae

Abstract

There have been many attempts to model the morphological richness and complexity of Arabic, leading to numerous Part-of-Speech (POS) tagsets that differ in terms of (a) which morphological features they represent, (b) how they represent them, and (c) the degree of specification of said features. Tagset granularity plays an important role in determining how annotated data can be used and for what applications. Due to the diversity among existing tagsets, many annotated corpora for Arabic cannot be easily combined, which exacerbates the Arabic resource poverty situation. In this work, we propose an intermediate tagset designed to facilitate the conversion and unification of different tagsets used to annotate Arabic corpora. This new tagset acts as a bridge between different annotation schemes, simplifying the integration of annotated corpora and promoting collaboration across the projects using them.

Keywords: Arabic, morphology, tagset, annotation

1. Introduction

Arabic is a morphologically rich language with numerous Part-of-Speech (POS) tagsets that were developed over a large period of time, by different groups of researchers with different goals in mind (Habash, 2010). The tagsets vary in size and hence the amount of information they encode: the type of morphological features and the degree of specification of said features. A small tagset size makes annotation more efficient, more accurate, and less costly; but naturally less informative. On the other hand, large tagsets that are harder to annotate, are more informative, and have been shown to be effective, when accurate, in increasing the accuracy of dependency parsing (Marton et al., 2010).

In this paper, we introduce **EMAD**^{1,2} (The Arabic **E**xtended **M**orphological **A**nalysis and **D**isambiguation Tagset), a fine-grained morphological representation and its associated mapping system, which can be used for optimal tagset conversion. To map a tagset x from/to EMAD, we require configurable map drivers for tagset x . The EMAD mapping process helps identify ambiguity and inconsistency in the original tag. Furthermore, the mapping process supports automatically enriching features in POS-annotated data sets, which also facilitates combining various annotated data sets to build larger resources in a controlled way.

Next we present some linguistic background (Section 2) and related work (Section 3). Then we present the EMAD representation (Section 4), how to use it to map across tagsets (Section 5), and evaluation results (Section 6).

2. The Arabic POS Tagset Zoo: Linguistic Background

Arabic morphological word features can be divided into three categories: lexical, inflectional, and cliticizational (Habash, 2010). Lexical features include the LEMMA, ROOT, and PATTERN, which contribute to deriving the core word meaning. Inflectional features, do not change the core meaning of the word, but generally regularly vary within a constrained space, e.g., GENDER, NUMBER, PERSON, ASPECT, VOICE, MOOD, CASE, and STATE. Cliticization refers to syntactically independent, but orthographically and phonological dependent morphemes that are written attached to the word form, e.g. single letter prepositions, or possessive pronouns. Due to Arabic's use of optional diacritics, some words may be ambiguous in complex ways that span all three types of features. For example, the word **بَاسِمٌ** *bAsm*³ can be the adjective **بَاسِمٌ** *baAsimū* 'smiling' (masculine, singular, nominative, indefinite), or the phrase **بِاسْمِ** *biAs.mi* 'in [the] name of' (masculine, singular, genitive, construct).

Table 1 compares some of the most commonly-used Arabic POS tagsets for the same example. Each tagset uses a relatively small number of sub-tags that are combined to form word tags. The number of sub-tags ranges from 6 in CATiB6 (Habash et al., 2009) and 17 in UD (Nivre, 2014; Taji et al., 2017) to around 170 for Buckwalter (Buckwalter, 2004) and MADA (Pasha et al., 2014). The combined word-level tags are naturally much larger, reach as small as 38 in CATiB6, and 300 in UD, 35,682 for Buckwalter, and 243,720 in MADA.

¹EMAD (إمداد) in Arabic means *supportive pillars*.

²<https://github.com/CAMEL-Lab/emad>

³HSB Arabic transliteration (Habash et al., 2007).

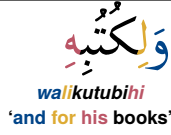
Tagset	
Penn (Bies/Kulick) (Kulick et al., 2006)	CC+IN+NN+PRP
CATiB6 (Habash et al., 2009)	PRT+PRT+NOM+NOM
Khoja (Khoja et al., 2001)	PC+PPr+NCPIMGI+NPrPSg3M
CAMEL (Khalifa et al., 2018)	CONJ+PREP+NOUN.MP.CG+PRON
ElixirFM (Smrž, 2007)	C-----+P-----+N-----P2R+SP---3MS2-
Buckwalter (Buckwalter, 2004)	CONJ+PREP+NOUN+CASE_DEF_GEN+POSS_PRON_3MS
UD (Nivre, 2014; Taji et al., 2017)	CONJ+ADP+NOUN'+PRON'' *[Case=Gen Definite=Cons Gender=Masc Number=Sing] **[Case=Gen Definite=Def Gender=Masc Number=Sing Person=3 PronType=Prs]
MADA (Pasha et al., 2014)	pos:noun prc3:0 prc2:wa_conj prc1:li_prep prc0:0 asp:na vox:na mod:na gen:m num:p stt:c cas:g per:na enc0:3ms_poss

Table 1: An example of one Arabic word annotated in different tagsets. The word and the tags have been color-coded so that the features match their corresponding tokens in the original word. The two Arabic Penn POS variants (Bies and Kulick) look the same for this example.

The different POS tagsets that have been proposed for Arabic vary in how they represent it in a number of dimensions. First is the dimension of **tokenization**: *what is the basic unit represented?* Some tagsets treat the full word including its clitics to be the base unit (e.g., MADA). Others separate clitics with different degrees, e.g., tags designed for treebanking tokenize all clitics except for the definite article, e.g., Penn/Bies/Kulick (Kulick et al., 2006), CATiB6, UD, ElixirFM (Smrž, 2007).

Second is the dimension of **depth**: *does the tag represent the form of the tagged morpheme or its underlying features?* Some tags are surface oriented, e.g., Buckwalter’s tags mark the suffix *h* as feminine singular even when its part of a broken masculine plural word such as *كاتبه katabah* ‘scribes’. Other tagsets dig deeper but then lose the connection the surface form, e.g., MADA represents gender and number as independent features even when they realize in a singular morpheme.

Third is the dimension of **granularity**: *how much details to specify?* Some tags specify fine details, e.g., MADA, others abstract away to general categories, e.g., CATiB6, and the extreme being the traditional Arabic grammar three-way split of nouns, verbs, and particles. In the middle space we find many variants such as the Arabic Penn POS (or Reduced Tag Set, RTS), which includes a couple of versions, Bies and Kulick (Kulick et al., 2006). The Arabic Penn POS tags reduce from the Buckwalter POS tagset and are inspired by the English Penn POS (Marcus et al., 1993). Other variants include the different versions of the Extended RTS (ERTS) (Diab, 2007; Aldarmaki and Diab, 2015). The SALMA tagset (Sawalha and Atwell, 2013) even goes further than other sets by indicating details such as declension and conjugation cate-

gories, but merges verbal mood and nominal case (as is done in traditional Arabic grammar).

Finally is the dimension of **representation**: *how are features and values indicated?* Some tags are more verbose (MADA, Buckwalter) and others less so (ElixirFM) even when they represent comparable levels of detail, e.g., the feature values of case, nominative, genitive, and accusative are represented as NOM, GEN, ACC in Buckwalter, cas:n, cas:a, cas:g in MADA, and 1, 2, 4 in ElixirFM. The representation is sometimes inconsistent across tokens, e.g., MADA treats enclitics such as *enc0:pron_3fs* as a single non-decomposable feature (3rd-person feminine singular pronoun) instead of explicitly indicating its person, gender, and number. An explicit representation is useful for controlled generation tasks such as gender rewriting over basewords and enclitics (Alhafni et al., 2022) as it gives direct access to needed information.

In this paper, we build on the MADA tagset, which is an example of a fine-grained morphological tagset containing the following 13 features: *prc3*, *prc2*, *prc1*, *prc0*, *enc0* (clitics), *asp*, *vox*, *mod*, *gen*, *num*, *stt*, *cas*, *per* (features), in addition to the POS and lemma (Pasha et al., 2014). MADA is used in a couple of popular tools for Arabic disambiguation, MADAMIRA (Pasha et al., 2014), and CamelTools (Obeid et al., 2020). The design decisions of EMAD target extending MADA to make the best of the clitics’ tokenization modeling in ElixirFM, Khoja, UD, and CATiB6 but maintain the advantages of MADA’s depth, granularity, and representation. We leave other morphological and lexical information such as declension classes and derivational details (as is done in the SALMA tagset) to the lexicon and morphological analyzer and do not model them.

3. Related Work

There are many efforts that summarize the different tagsets designed for Arabic (Habash, 2010; Sawalha and Atwell, 2013; Zeroual et al., 2017). In this project, we have investigated and referenced some of the major tagsets as part of our design decisions; these include Khoja (Khoja et al., 2001), Buckwalter (Buckwalter, 2004), ElixirFM (Smrž, 2007), Penn/Bies/Kulick (Kulick et al., 2006), ERTS (Diab, 2007; Aldarmaki and Diab, 2015), CATiB (Habash and Roth, 2009), MADA (Pasha et al., 2014), UD (Taji et al., 2017), SALMA (Sawalha and Atwell, 2013), and CAMEL POS (Khalifa et al., 2018).

The idea of this project was inspired by the Inter-set project (Zeman, 2008), which suggests a similar idea for converting between different tagsets by defining a universal intermediate tagset. We chose to create our own intermediate space instead of using Inter-set to tailor it specifically for Arabic tagsets, as our intermediate tagset is based on the MADA-style features which are already defined and used for Arabic. We also take inspiration from previous work on enrichment mapping across tagsets such as the work of Alkuhlani et al. (2013) who showed that under-specified morphological features can be predicted with an accuracy of 94%-95% when mapping from CATiB dependency trees to Buckwalter tags.

4. The EMAD Representation

The EMAD tag is a *matrix of feature values* organized around columns of tokens in the D3 tokenization scheme, which splits off all clitics including the definite article (Sadat and Habash, 2006). For each token column, the rows indicate specific lexical and inflectional features: **diac** (diacritization), **pos** (part-of-speech), **per** (person), **asp** (aspect), **cas** (case), **stt** (state), **mod** (mood), **vox** (voice), **gen** (functional gender), **form_gen** (form gender), **num** (functional number) and **form_num** (form number). The feature names follow the same naming convention used in the MADA tagset.

The token columns follow the order of the proclitic, base word, and clitic as they appear in the word. EMAD organizes these morphemes along eight categories that must appear in the EMAD tag in this specific order: **PRC_QUEST** (interrogative proclitic), **PRC_CONJ** (conjunction proclitic), **PRC_PREP** (prepositional proclitic), **PRC_VPAR** (verbal particle proclitic), **PRC_DET** (determiner proclitic), **BASE** (baseword), **ENC_PRON** (pronominal enclitic), and **ENC_PART** (particle enclitic).

Since a word may have multiple analyses out of context, these would be represented in EMAD as an array of alternative matrices that can be ranked

	PRC CONJ	PRC PREP	BASE	ENC PRON
diac	wa	li	kutubi	hi
pos	conj	prep	noun	pron
per	na	na	na	3
asp	na	na	na	na
cas	na	na	g	g
stt	na	na	c	d
mod	na	na	na	na
vox	na	na	na	na
gen	na	na	m	m
form_gen	na	na	m	m
num	na	na	p	s
form_num	na	na	s	s

Table 2: The EMAD version of Table 1’s example, **وَالِكُتُبِ** *walikutubihi* ‘and for his books.’

in context or according to some independent criteria (Pasha et al., 2014; Inoue et al., 2022).

Table 2 shows the EMAD matrix representation for the example word in Table 1.

5. Mapping across Tagsets

We discuss next how we approach the process of mapping other Arabic POS tagsets from and into EMAD.

5.1. Approach Overview

For each Arabic POS tagset, we create a driver that defines the tagset. This driver ideally contains five components:

1. The tagset features, e.g., POS, number, gender, case, state, etc.
2. The tag format, i.e., how the features are arranged into the tag.
3. The possible values of each feature.
4. The possible combinations of features and values.
5. The map of the feature-values of the POS tagset to EMAD feature-values (many-to-many mapping, indicated as a set of one item to many mappings).

The first four components in the list are information about the tagset which is independent from EMAD, while the 5th component, is defined with respect to it. The drivers we create for mapping minimally require the first three components as well as the 5th one. The 4th component is not needed for the mapping but is required for well-formedness checking – the process of validating that a specific POS tag from a given tagset does not violate the rules and guidelines of that tagset’s valid

feat	value	EMAD_cat	pos	per	gen	num
enc0	2fp_dobj	ENC_PRON	pron	2	f	p
enc0	2ms_dobj	ENC_PRON	pron	2	m	s
pos	noun	BASE	noun	-	-	-
per	2	BASE	-	2	-	-

Table 3: Sample rows taken from the map of MADA to EMAD that show how a few feature-value pairs for enclitics map into the EMAD tagset.

forms and valid sub-tag combinations, e.g., flagging NUON as an invalid form of NOUN, or flagging VERB+CASE:NOMINATIVE as an impossible combination. This step has not been implemented; we leave it for future work.

5.2. Parsing the Input Tag

The first step to convert a tag from tagset A into EMAD is to parse it, i.e., to convert the input tag into a feature-value pair representation. This is done using the format of the tag that is included in the map driver as discussed above. For example, parsing the CAMEL POS tag (Khalifa et al., 2018) CONJ+NOUN.MS+PRON would give the feature-value pairs: PRC2:CONJ, POS:NOUN, GEN:M, NUM:S, ENC0:PRON.

While this step might be easy for a tagset like MADA, where the tag is already in a feature-value pair representation, it could be difficult in other tagsets, especially when the tag is ambiguous. Take for example the CATiB6 tag PRT+NOM. The ambiguity of this tag comes from the fact that it can be used for a word like *لِكِتَابٍ* *likitAbī* ‘for the book’ or a word like *مِنْهُ* *min.hu* ‘from him’. For the first word, the parsing output should be (PRC1:PRT, POS:NOM) because the nominal here is a noun and is considered to be the main token of the word. For the second word, however, the parsing output should be (POS:PRT, ENC0:NOM) since the preposition is the main part of the word and it is followed by a nominal which is a pronoun.

To handle the processing of such ambiguity, we use Finite State Transducers (FSTs) which allow us to find all possible ways of parsing a tag provided the format of the tag and the possible values of each feature. For the implementation of the FSTs, we used the Python library Pyfoma (Hulden, 2022).

5.3. Mapping into and out of EMAD

The map from any tagset into EMAD is in the format of a table, where each row contains a feature-value pair from the source tagset, along with the EMAD category to which it would map, and the values for each morphological feature. Table 3 shows how these rows look like using an example from

the MADA map driver. The same map table that is used for converting into EMAD is also used to convert in the reverse direction.

The tool for converting from a tagset into the EMAD tagset has been implemented using the guidelines mentioned above, and the code is publicly available.⁴ So far, four drivers have been implemented for the Buckwalter POS tagset, the CAMEL POS tagset, the CATiB6 POS tagset, and the MADA POS tagset.

6. Evaluation

We evaluate our ability to map into and from EMAD through its use in converting across different types of tagsets: $tagset A \Rightarrow EMAD \Rightarrow tagset B$.

6.1. Data Set and Metric

We evaluate on conversion across three tagsets with very different design considerations: Buckwalter, CATiB6, and MADA. In total we have six conversion pairs: Buckwalter \Rightarrow CATiB6 or MADA, CATiB6 \Rightarrow Buckwalter or MADA, and MADA \Rightarrow Buckwalter or CATiB6. We plan to add more tagsets to the EMAD framework in the future.

The evaluation dataset is based on the Penn Arabic Treebank (PATB) parts 1v4.1, 2v3.1 and 3v3.2 (Maamouri et al., 2004, Maamouri et al., 2010a,b, 2011). We only use the training portion as specified by Diab et al. (2013) (~500k words). The PATB provides the basic Buckwalter tags. We use the Camel Tools mapper (Obeid et al., 2020) and the synchronized data sets from Inoue et al. (2022) for MADA and CATiB6 tag versions. We remove all lexicalizations unless they appear as part of feature values (e.g., *prc2:wa_conj*). In total, there are 3,162 unique triplets (Buckwalter, CATiB6, MADA) within the dataset. We make the dataset publicly available.⁴

We chose Recall as the metric for assessing mapping success. It accounts for ambiguity in conversions, considering them correct if the expected tag is among the outputted tags. However, this metric does not address over-generation of tags. Future work should explore additional evaluation metrics.

6.2. Results

Table 4 shows the results, where the percentages of correctly recalled conversions are calculated relative to the number of parsed tags, not the total number of tags. Parsed tags are those from the input tagset that were processed by the map and produced a valid output. The results are quite promising, with minor errors, which we discuss next.

⁴<https://github.com/CAMEL-Lab/emad>

Conversion	Parsed	Recalled
Buckwalter to CATiB6	99.75%	100.00%
Buckwalter to MADA	99.75%	98.76%
CATiB6 to Buckwalter	100.00%	99.75%
CATiB6 to MADA	100.00%	99.94%
MADA to Buckwalter	100.00%	98.61%
MADA to CATiB6	100.00%	100.00%

Table 4: Evaluation results of the conversion between Buckwalter, CATiB6, and MADA tagsets using 3,162 unique tag triplets.

6.3. Error Analysis and Discussion

Parsing Errors Eight Buckwalter tags could not be converted into the other tagsets because they were not recognizable by the tagset map. An example is the tag CONJ+PSEUDO_VERB+NOUN+CASE_DEF_ACC for the compound word **ولا بُدّ** *wlAbud~a* ‘and no avoiding.’ In this example, the two subtags (PSEUDO_VERB and NOUN) can only be interpreted as the main POS of the word. This leads to a parsing error, since there can only be one main POS in a BW tag, as defined in our map driver. However, this case is the result of a spelling error in the original data, where this word should have been written and tagged as two separate words: **ولا بُدّ** *wlA bud~a*. In a way, the failure of the conversion process has helped us identify spelling errors in the dataset. This corresponds to the missing 0.25% in the Parsed column in Table 4.

Mapping Errors When mapping into CATiB6, we get perfect recall. However, the largest number of recall failures happen when mapping from **MADA to Buckwalter**: we get a total of 44 errors, resulting in 98.61% recall. Eight of these errors are the same as the above-mentioned spelling errors. 22 errors are because of the difference in a linguistic decision between the Buckwalter and MADA tagsets, such as the case where Buckwalter considers the subjunctive particle **لِ** ‘for’ as a preposition, but MADA considers it as a subjunctive particle since it only proceeds verbs and not nouns. This error can be fixed either by modifying the map table to account for this case, or it can be ignored and flagged during conversion, allowing for such a decision to be made manually. The remaining 14 errors are annotation errors of the data, where the MADA tag incorrectly disagreed with the Buckwalter tag in some features.

When converting from **Buckwalter to MADA**, we get a total of 39 errors. These include the same (22+14=36) errors from above. The remaining three errors are also based on a difference in the linguistic decision made by the two tagsets about

the closed-class word **لَقَدْ** *laqad* ‘already’ which is treated by Buckwalter as a single baseword, but by MADA as consisting of a proclitic and a baseword.

The errors in converting from **CATiB6 to Buckwalter** are the same as the eight parsing error cases mentioned above since our mapping couldn’t produce the faulty tags.

Finally, when converting from **CATiB6 to MADA**, there were only two errors resulting in 99.94% recall. These were two instances of gold errors, where the word **بِلا** *biLA* ‘with+no (without)’ was incorrectly tagged in CATiB6 as PRT+NOM instead of PRT+PRT.

In general, the evaluation and analysis indicate that our mapping effectively handles most tags, with minor exceptions. These exceptions may serve as indicators for identifying errors in the data, or aiding users in making linguistic decisions.

7. Conclusion and Future Work

This paper presented a new extended tagset, EMAD, and a system that used it to support conversion between different Arabic POS tagsets. We evaluated the conversion process and found it useful for detecting errors and inconsistent linguistic decisions between different tagsets in a parallel dataset.

In the future, we plan to use EMAD for well-formedness checking of tags, as well as automatic feature enrichment to support combining multiple annotated datasets together. We plan to study the effect of such improvements on downstream applications.

Bibliographical References

- Hanan Aldarmaki and Mona Diab. 2015. [Robust part-of-speech tagging of Arabic text](#). In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 173–182, Beijing, China. Association for Computational Linguistics.
- Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. [User-centric gender rewriting](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.
- Sarah Alkuhlani, Nizar Habash, and Ryan Roth. 2013. [Automatic morphological enrichment of a morphologically underspecified treebank](#). In *Proceedings of the 2013 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–470, Atlanta, Georgia. Association for Computational Linguistics.
- Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.
- Mona Diab. 2007. Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 89–96, Prague, Czech Republic.
- Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.
- Nizar Habash, Reem Faraj, and Ryan Roth. 2009. Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.
- Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 221–224, Suntec, Singapore.
- Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.
- Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.
- Hulden. 2022. PyFoma <https://github.com/mhulden/pyfoma/>.
- Go Inoue, Salam Khalifa, and Nizar Habash. 2022. [Morphosyntactic tagging with pre-trained language models for Arabic and its dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.
- Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. [A morphologically annotated corpus of emirati Arabic](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Shereen Khoja, Roger Garside, and Gerry Knowles. 2001. A tagset for the morphosyntactic tagging of Arabic. In *Proceedings of the Corpus Linguistics Conference*, pages 341–353, Lancaster, UK.
- Seth Kulick, Ryan Gabbard, and Mitch Marcus. 2006. Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT)*, pages 31–42, Prague, Czech Republic.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Yuval Marton, Nizar Habash, and Owen Rambow. 2010. [Improving Arabic dependency parsing with lexical and inflectional morphological features](#). In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21, Los Angeles, CA, USA. Association for Computational Linguistics.
- Joakim Nivre. 2014. Universal dependencies for Swedish. In *Proceedings of the Swedish Language Technology Conference (SLTC)*.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. [CAMEL tools: An open source python toolkit for Arabic natural language processing](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.
- Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the International Conference on Computational Linguistics and the Conference of the Association for Computational Linguistics (COLING-ACL)*, pages 1–8, Sydney, Australia.

Majdi Sawalha and Eric Atwell. 2013. A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging. *Word Structure*, 6(1):43–99.

Otakar Smrž. 2007. ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 1–8, Prague, Czech Republic. ACL.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. [Universal Dependencies for Arabic](#). In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Daniel Zeman. 2008. [Reusable tagset conversion using tagset drivers](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Imad Zeroual, Abdelhak Lakhouaja, and Rachid Belahbib. 2017. [Towards a standard part of speech tagset for the arabic language](#). *Journal of King Saud University - Computer and Information Sciences*, 29(2):171–178. Arabic Natural Language Processing: Models, Systems and Applications.

8. Language Resource References

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri, and Wajdi Zaghouni. 2010a. Arabic treebank: Part 1 v 4.1. Linguistic Data Consortium (LDC2010T13).

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri, and Wajdi Zaghouni. 2011. Arabic treebank: Part 2 v 3.1. Linguistic Data Consortium (LDC2011T09).

Mohamed Maamouri, Ann Bies, Seth Kulick, Sondos Krouna, Fatma Gaddeche, and Wajdi Zaghouni. 2010b. Arabic treebank: Part 3 v 3.2. Linguistic Data Consortium (LDC2010T08).