Universal Dependencies for Persian

***Mojgan Seraji, **Filip Ginter, *Joakim Nivre**
*Uppsala University, Department of Linguistics and Philology, Sweden
**University of Turku, Department of Information Technology, Finland
*firstname.lastname@lingfil.uu.se
**figint@utu.fi

Abstract

The Persian Universal Dependency Treebank (Persian UD) is a recent effort of treebanking Persian with Universal Dependencies (UD), an ongoing project that designs unified and cross-linguistically valid grammatical representations including part-of-speech tags, morphological features, and dependency relations. The Persian UD is the converted version of the Uppsala Persian Dependency Treebank (UPDT) to the universal dependencies framework and consists of nearly 6,000 sentences and 152,871 word tokens with an average sentence length of 25 words. In addition to the universal dependencies syntactic annotation guidelines, the two treebanks differ in tokenization. All words containing unsegmented clitics (pronominal and copula clitics) annotated with complex labels in the UPDT have been separated from the clitics and appear with distinct labels in the Persian UD. The treebank has its original syntactic annotation scheme based on Stanford Typed Dependencies. In this paper, we present the approaches taken in the development of the Persian UD.

Keywords: Universal Dependencies, Persian, Treebank

1. Introduction

In the past decade, the development of numerous dependency parsers for different languages has frequently been benefited by the use of syntactically annotated resources, or treebanks (Böhmová et al., 2003; Haverinen et al., 2010; Kromann, 2003; Foth et al., 2014; Seraji et al., 2015; Vincze et al., 2010). However, treebanks only exist for a small number of languages, and considering the number of 7,000+ languages in the world,[1] a large number of languages still lack treebanks.

Due to the diverse typologies and grammatical structures that exist across languages, treebanks are created with different annotation schemes. These annotation variations can further be explained by different linguistic theories and the syntactic annotations that treebank developers select based on their own preferences (Nivre, 2015). These dissimilarities in annotation schemes often have an impact on syntactic parsers, which means that the results are not comparable across languages. McDonald et al. (2013) enumerate several issues for natural language parsing when treebanks are labeled with different annotation schemes.

This has brought many researchers and developers in natural language processing to the conclusion that having a common standard and cross-linguistically valid annotation scheme would favor parsing research. Recently, there have been a number of initiatives for developing data sets with cross-linguistically consistent annotation scheme for morphological and syntactic structures. These efforts have resulted in the emergence of the Stanford Typed Dependencies Representation (de Marneffe et al., 2006; de Marneffe and Manning, 2008), the Google Universal Part-of-Speech Tagset (Petrov et al., 2012), and Interset interlingua for morphosyntactic features used in the HamleDT treebank collection (Zeman, 2008; Zeman et al., 2012). The most recent effort is the Universal Dependencies (UD), which more or less combine all the earlier efforts in this regard. In this paper, we present how we adapt the Universal Dependencies to Persian by converting the Uppsala Persian Dependency Treebank (UPDT) (Seraji, 2015) to the Persian Universal Dependencies (Persian UD). First, we briefly describe the Universal Dependencies and then we present the morphosyntactic annotations used in the extended version of the Persian UD.

[1] http://www.bbc.co.uk/languages/guide/languages.shtml

2. Universal Dependencies

Universal Dependencies (UD) is an ongoing project that aims to facilitate multilingual parser development, cross-lingual learning, and research on parsing with a perspective of language structure and usage. In pursuing this goal, the project focuses on the development of cross-linguistically consistent syntactic annotations in the treebanks, for a large number of languages. The fundamental idea underlying this project is to discover the common and universal categories across languages as well as the language-specific ones. In other words, while the annotation scheme in UD stipulates a consistent annotation of similar constructions across languages, it also allows language-specific extensions or omissions when necessary. This means that when some language-specific constructions cannot be covered by the scheme introduced in the UD, the scheme can be extended by new dependency relations for that specific language.

Currently, 37 treebanks for 33 languages are annotated based on version 1 of the UD guidelines and were released in November 2015. These treebanks are in the version 1.2 and represent the following languages: Ancient Greek, Arabic, Basque, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Gothic, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Latin, Norwegian, Old Church Slavonic, *Persian*, Polish, Portugese, Romanian, Slovenian, Spanish, Swedish, and Tamil. Furthermore, there are more languages that have recently joint the UD project.

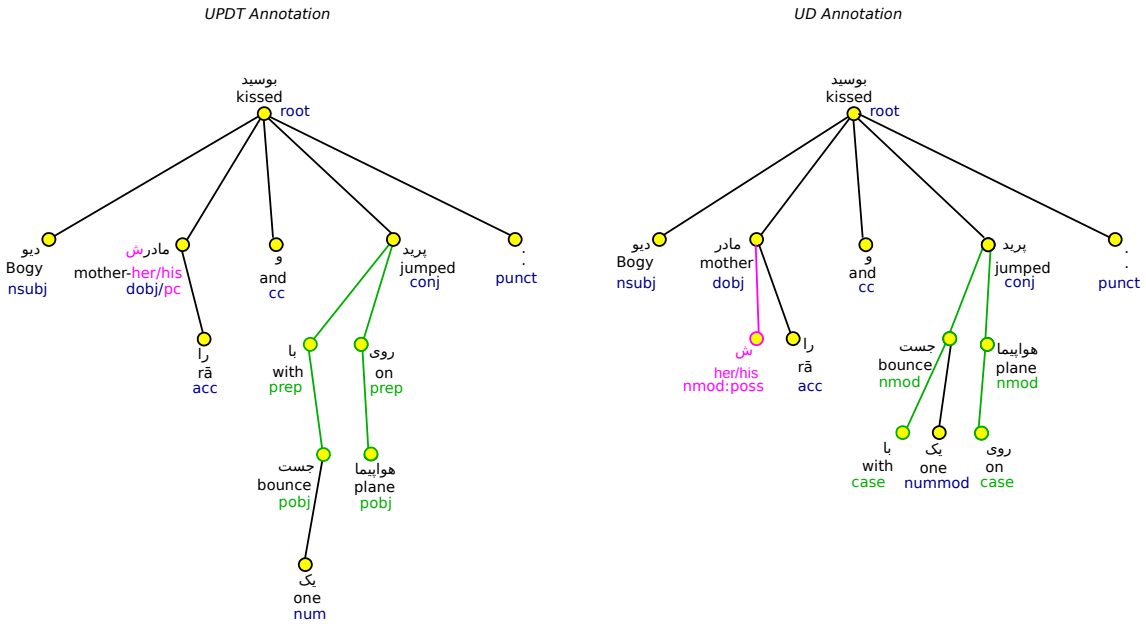

Figure 1: A comparative syntactic annotation for a Persian sentence in the UPDT and UD. To make the figure more readable, glosses have been simplified as follows: mother-her/his = mother-pc.3sg, kissed = kiss.past.3sg, jumped = jump.past.3sg.
Gloss: Bogy mother-pc.3sg rā kiss.past.3sg and with one bounce on plane jump.past.3sg.
Translation: The bogy kissed her/his mother and jumped on the plane with one bounce.

3. Persian Universal Dependencies

The Persian Universal Dependencies (Persian UD) is the converted version of the Uppsala Persian Dependency Treebank (UPDT)[2] (Seraji, 2015) to the UD framework. The treebank has its original annotation scheme based on Stanford Typed Dependencies (de Marneffe et al., 2006; de Marneffe and Manning, 2008). The scheme was extended for Persian to include the language-specific syntactic relations that could not be covered by the primary scheme developed for English. The treebank consists of nearly 6000 sentences from written text with large domain variations, in terms of different genres (containing newspaper articles, fictions, technical descriptions, and documents about culture and art)[3] and tokenization. The variations in the tokenization are due to the orthographic variations of compound words, fixed expressions, and different types of clitics in the language. In the UPDT, only fixed expressions delimited with whitespace are handled as distinct tokens. Fixed expressions in attached forms and different types of unsegmented clitics such as pronominal and copula clitics are not separated from their head word. In other words, these cases have not been treated at morphological level, but annotated with special labels at the syntactic level instead. The syntactic annotation of the UPDT has been done semi-automatically using MaltParser (Nivre et al., 2006) in a bootstrapping scenario. All sentences have been manually validated.

When converting the UPDT to the Persian UD, all words containing unsegmented clitics (pronominal and copula clitics) annotated with complex labels in the UPDT, were separated from the clitics and received distinct labels in the Persian UD. Figure 1 illustrates the differences between the two treebanks for a Persian sentence. In this example, the direct object (*dobj*) in the UPDT consists of the word *mother* and the possessive pronominal clitic *her/his* (colored in pink), marked with the label *dobj/pc*. This label in the UPDT annotation is defined as a complex label (cf. Seraji, 2015) and is used to denote the main function *dobj* and an additional (pronominal clitic) *pc* element. By contrast, the pronominal clitic *her/his* in the UD annotation is separated from its head word *mother* and is placed as its dependent instead, labeled with nominal modifier in possessive construction (*nmod:poss*). The two treebanks further differ in the handling of head nodes in prepositional phrases. The prepositions *with* and *on* in the UPDT functions as heads of prepositional phrases with the dependency relation *prep*, while in the UD annotation they are treated as dependents of the nouns (or the prepositional objects) *bounce* and *plane*, labeled as the relation *case*. The following subsections present the current stage of the Persian UD.

[2]The treebank is open source and freely available in CoNLL-format and can be downloaded from http://stp.lingfil.uu.se/∼mojgan/UPDT.html

[3]The treebank data is extracted from the open source, validated Uppsala Persian Corpus (UPC), which is currently the largest freely available corpus of Persian. The corpus consists of 2,704,028 tokens and annotated with part-of-speech tags and morphological features. For a comprehensive description of the corpus pertaining to the tokenization and morphological annotation see Seraji (2015, Chapter 3)

UPDT	UD
ADJ	ADJ
ADJ_CMPR	ADJ
ADJ_INO	ADJ
ADJ_SUP	ADJ
ADV	ADV
ADV_COMP	ADV
ADV_I	ADV
ADV_LOC	ADV
ADV_NEG	ADV
ADV_TIME	ADV
CLITIC	PART
CON	CONJ and SCONJ
DELM	PUNCT
DET	DET
FW	X
INT	INTJ
N_PL	NOUN
N_SING	NOUN
N_VOC	NOUN
NUM	NUM
P	ADP
PREV	ADP
PRO	PRON
V_AUX	AUX
V_IMP	VERB
V_PA	VERB
V_PP	VERB
V_PRS	VERB
V_SUB	VERB

Table 1: Mapping from the UPDT to the Google Universal Part-of-Speech Tagset in the UD.

3.1. Morphological Representation

Morphological representations in UD provide information about lemmas (lemma is the basic or canonical form of a word), part-of-speech tags (part-of-speech is the grammatical category of a word), and morphological features (a grammatical feature of a word is the characteristic or property of the grammatical category of the word such as gender, number, tense, etc.).

Since lemmas are language dependent, a clear representation of those is not specifically defined in the UD framework. Thus, the lemma field is normally determined by language-specific dictionaries. For Persian this field has not yet been filled and the work is still in progress.

A list of 17 coarse-grained part-of-speech tags, based on the Google Universal Part-of-Speech Tagset (Petrov et al., 2012), is defined in UD to cover the part-of-speech categories across languages. The UPDT is annotated with 29 part-of-speech tags with morphological information. These tags were straightforwardly mapped to 15 part-of-speech tags of the total 17 tags in the UD. The mapped tags are listed in Table 1.

It is worth noting that the universal part-of-speech tags in the UD, contrary to the UPDT, have distinct tags for coordinating conjunctions and subordinating conjunctions. These tags are presented as *CONJ* for coordinating conjunctions, and *SCONJ* for subordinating conjunctions. In the conversion of the UPDT to the Persian UD all the tags *CON* were automatically traced and received the corresponding UD tags based on their dependency relations. For instance, the tag *CON*, for tokens with the dependency relations *cc* (coordination) and *mwe* (multi-word expressions), was converted to the tag *CONJ* and the rest received the tag *SCONJ*. Furthermore, in the automatic processing, we added multiple rules for rewriting the rest of the coarse-grained part-of-speech tags.

In the current release of the Persian UD, all the morphological features have been included based on those introduced in the UD's universal features to further distinguish lexical and grammatical characteristics of words that could not have been covered by the part-of-speech tags. Adding lemma is still a work in progress.

3.2. Syntactic Representation

The syntactic annotation scheme in UD is based on the Universal Stanford Dependencies (de Marneffe et al., 2014) and consists of 40 dependency relations that are intended to broadly capture various dependency relations between words. The underlying principle of the scheme is that dependencies hold between content words, while function words attach to the content word they further specify.

Since UPDT has its original annotation scheme based on Stanford Typed Dependencies with a language-specific variant for Persian, it already follows this assumption. However, there are two exceptions where the old scheme chooses function words as heads: prepositions in prepositional phrases, and copula verbs that have a prepositional phrase as their complement. These exceptions have been revised in the Universal Stanford Dependencies and now the scheme consistently keeps content words as heads.

The UPDT consists of a total of 96 dependency relations, of which 48 are used for basic relations (including 10 new additions to the STD) and 48 for complex relations. Complex relations are assigned to words containing unsegmented clitics. The conversion of the UPDT to the Persian UD has been carried out semi-automatically. In this process, we have used scripts tailored for Persian to separate different types of clitics from their host. Furthermore, we used another conversion script for reversing the head and dependent relations in the prepositional modifier (prep) and object of a preposition (pobj). Subsequently we added different rules for rewriting the dependency labels. Those sentences that have in any case been involved in the conversion process, have manually been checked and validated. This basically covers all sentences containing any form of clitics and prepositional phrases.

The 10 extended language-specific dependency labels to the STD for Persian (the UPDT), are renewed in the UD framework with new relation names. These relations are displayed in italic in Table 2. Three of the dependency relations that are referred to as universal grammatical relations in the Universal Dependencies initiative, are taken from the relations that are introduced as extended relations to the STD for Persian. These dependency relations in the UPDT are introduced as fw (foreign words), dep-top (dependent topicalization), and dep-voc (dependent vocative). The three relations are generalized in the Universal Dependencies as foreign, dislocated, and vocative respectively.

UPDT	UD
acc	*case*
acomp	xcomp
acomp-lvc	*compound:lvc*
advcl	advcl
advmod	advmod
amod	amod
appos	appos
aux	aux
auxpass	auxpass
cc	cc
ccomp	ccomp
complm	mark
conj	conj
cop	cop
cpobj	*nmod*
cprep	*case*
dep	dep
dep-top	*dislocated*
dep-voc	*vocative*
det	det
dobj	dobj
dobj-lvc	*compound:lvc*
fw	*foreign*
mark	mark
mwe	mwe
neg	neg
nn	name
npadvmod	nmod
nsubj	nsubj
nsubj-lvc	*compound:lvc*
-	nsubj-nc
nsubjpass	nsubjpass
num	nummod
number	nmod
parataxis	parataxis
pobj	nmod
poss	nmod:poss
preconj	conj:preconj
predet	det:predet
prep	case
prep-lvc	*compound:lvc*
prt	compound:prt
punct	punct
quantmod	advmod
rcmod	acl:relcl
rel	mark
root	root
tmod	advmod
xcomp	xcomp

Table 2: Mapping from the UPDT to the dependency relations in the UD.

The total number of dependency relations in the Persian UD is 44, consisting of 37 univeral dependencies and 7 language-specific relations. The language-specific relations include: relative clause modifier *acl:relcl*, predeterminer *det:predet*, light verb construction *compound:lvc*, phrasal particle *compound:prt*, preconjunction *conj:preconj*, the genetive modifier *nmod:poss*, non-canonical subject *nsubj:nc*. Apart from the relation *nsubj:nc*, the rest of the language-specific relations presented in the Persian UD function the same as their UPDT counterparts but under different relation names. The relation *nsubj:nc* was introduced by spliting the pronominal clitics in the Persian UD to mark non-canonical subjects. Non-canonical subjects normally appear in form of pronominal clitics attached to the preverbal elements of the light verbs. For instance the compound verb خوش آمدن (*gloss*: well come-inf, *translation*: to like, to welcome) in the following sentence is used with the non-canonical subject *-am* (pronominal clitic in first person singular):

من از آن خوشم می‌آید.

man az ān xoš-am mi-āy-ad .
I of that well-pc.1sg cont-come.pres-3sg .
I like that.

In the syntactic status of a non-canonical subject, in contrast to canonical subject, the preverbal element in the LVC agrees with person and number instead of the light verb, and the light verb always remains in third singular. In the above example, although the first person singular (I / pc) has no access to canonical verbal agreement, it still functions as the syntactic subject.

4. Conclusion

In this paper we presented the Universal Dependencies for Persian, which still is a work in progress. This is the recent effort of treebanking Persian based on a universal morphosyntactic annotation scheme, called Universal Dependencies. The treebank is available under an open license at *www.universaldependencies.org*.

5. Acknowledgements

We would like to thank Carina Jahani and Forogh Hashabeiky for their continued support and guidance in the Persian grammar. We are extremely thankful to all the UD's contributors for their efforts and collaborations. We would also like to thank the three anonymous reviewers for their comments.

References

Böhmová, Alena, Jan Hajič, Eva Hajičocá, and Barbora Hladká (2003). "The Prague dependency treebank". In: *Treebanks*. Springer Netherlands, pp. 103–127.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford Dependencies: A cross-linguistic typology". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 4585–4592.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning (2006). "Generating Typed Dependency Parses from Phrase Structure Parses". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 449–454.

de Marneffe, Marie-Catherine and Christopher D. Manning (2008). "The Stanford Typed Dependencies Representation". In: *Proceedings of the COLING'08 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8.

Foth, Kilian, Arne Köhn, Niels Beuck, and Wolfgang Menzel (2014). "Because size does matter: The Hamburg dependency treebank". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC '14)*, pp. 2326–2333.

Haverinen, Katri, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski (2010). "Treebanking Finnish". In: *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT '10)*, pp. 79–90.

Kromann, Matthias T. (2003). "The Danish Dependency Treebank and the DTAG Treebank Tool". In: *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pp. 217–220.

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

Nivre, Joakim (2015). "Towards a Universal Grammar for Natural Language Processing". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Springer, pp. 3–16.

Nivre, Joakim, Johan Hall, and Jens Nilsson (2006). "MaltParser: A Data-Driven Parser-Generator for Dependency Parsing". In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 2216–2219.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, pp. 2089–2096.

Seraji, Mojgan (2015). "Morphosyntactic Corpora and Tools for Persian". PhD Thesis. Studia Linguistica Upsaliensia 16.

Seraji, Mojgan, Bernd Bohnet, and Joakim Nivre (2015). "ParsPer: A Dependency Parser for Persian". In: *Proceedings of the 3rd International Conference on Dependency Linguistics (DepLing '15)*, pp. 300–309.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik (2010). "Hungarian Dependency Treebank". In: *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10)*, pp. 1855–1862.

Zeman, Daniel (2008). "Reusable Tagset Conversion Using Tagset Drivers". In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pp. 213–218.

Zeman, Daniel, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič (2012). "HamleDT: To Parse or Not to Parse?" In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2735–2741.