The Persian Dependency Treebank Made Universal
Mohammad Sadegh Rasooli1, Pegah Safari2, Amirsaeid Moloodi3, Alireza Nourian4

1 Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
2 Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
3 Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran
4 Sobhe, Tehran, Iran
1 rasooli@seas.upenn.edu, 2 p_safari@sbu.ac.ir, 3 amirsaeid.moloodi@shirazu.ac.ir, 4 nourian@sobhe.ir

Abstract
We describe an automatic method for converting the Persian Dependency Treebank (Rasooli et al., 2013) to Universal Dependencies. This treebank contains 29107 sentences. Our experiments along with manual linguistic analysis show that our data is more compatible with Universal Dependencies than the Uppsala Persian Universal Dependency Treebank (Seraji et al., 2016), and is larger in size and more diverse in vocabulary. Our treebank brings in a labeled attachment F-score of 85.2 in supervised parsing. Our delexicalized Persian-to-English parser transfer experiments show that a parsing model trained on our data is ≈2% absolutely more accurate than that of Seraji et al. (2016) in terms of labeled attachment score.1

1 Introduction
In recent years, there has been a great deal of interest in developing universal dependency treebanks (McDonald et al., 2013; Rosa et al., 2014; Nivre et al., 2020). The main goal of the Universal Dependencies project (Nivre et al., 2020) is to develop a consistent linguistic annotation scheme in different levels from tokenization to syntactic dependency relations. As a result, the majority of annotation discrepancies disappear, and the resulting dataset facilitates several cross-lingual natural language processing tasks including part-of-speech transfer (Täckström et al., 2013), syntactic transfer (Naseem et al., 2010; McDonald et al., 2011; Ammar et al., 2016; Zhang et al., 2019), and probing (Tenney et al., 2019; Hewitt and Manning, 2019). Starting with 10 treebanks in 2015, there are 163 treebanks in version 2.6 (May 2020) including the Uppsala Persian Treebank (Seraji et al., 2016).

Persian (aka Farsi) is a pro-drop morphologically rich language with a high degree of free word order and a unique light verb construction (Karimi-Doostan, 2011). Despite its importance, it still suffers from lack of sufficient annotated data. The Uppsala Universal treebank (Seraji et al., 2016) is currently the only publicly available universal treebank for Persian. It is a valuable resource based on news genre, and has been used as a testbed in previous work (Zeman et al., 2018; Chi et al., 2020). Among other non-universal treebanks, the Persian dependency treebank (PerDT) (Rasooli et al., 2013) is significantly larger than (Seraji et al., 2016) (29K vs. 6K sentences), and its sentences are sampled from contemporary Persian texts in different genres (as opposed to only news genre).

In this paper, we propose an automatic method for converting PerDT (Rasooli et al., 2013) to Universal Dependencies (An example of such conversion is shown in Figure 1). After a thorough analysis of dependency relations in the treebank, we design different mapping rules to generate trees with universal relations. This process involves a series of steps including unifying tokenization, part-of-

Figure 1: An example of our automatic conversion. The universal labels are shown at the bottom of words.
speech tags, named-entity recognition, and finally mapping dependencies. The mapping for many relations are not necessarily one-to-one, and we have to deal with peculiar cases that are specific to certain structures in modern Persian. Therefore, our approach is neither a blind one-to-one mapping, nor an expensive and time-consuming manual process. We empirically show that our annotations are more compatible with the Universal guidelines via learning a delexicalized transfer model with more than 2% absolute difference in labeled attachment score. The summary of our contributions is as following:

- We propose an automatic annotation conversion process with manual care of special cases. We develop a new Persian Universal Treebank with 29107 sentences. This is in contrast to the treebank of Seraji et al. (2016) that contains 5997 sentences.
- We develop a modified and corrected version of PerDT with the Universal tokenization scheme. Moreover, the new release resolves various tagging errors in the original dataset. Most of these corrections are made by manually fixing annotation errors flagged by our mapping pipeline.

2 Related Work

There has been a great deal of interest in designing and developing Persian dependency treebanks (Pouramini and Mozayani, 2007; Seraji et al., 2012, 2014, 2016; Rasooli et al., 2011b, 2013; Ghayoomi and Kuhn, 2014). Among them, the Uppsala UD treebank (Seraji et al., 2016) is the only treebank with Universal Dependencies. We have found some caveats in the Uppsala Universal Treebank (Seraji et al., 2016). This causes annotation discrepancies in some frequently used dependency relations such as compound:lvc, cop, csubj, fixed, obl, and xcomp (see §A for more details).

We primarily focus on converting the Persian dependency treebank (PerDT) (Rasooli et al., 2013). PerDT has been used in previous studies for Persian dependency parsing (Khallash et al., 2013; Feely et al., 2014; Nourian et al., 2015; Pakzad and Minaei-Bidgoli, 2016). It has been extended to other representations including semantic roles (Mirzaei and Moloodi, 2016) and discourse (Mirzaei and Safari, 2018). It is also included in the HamleDT collection (Rosa et al., 2014).

3 Approach

In the conversion process, we have noticed several key differences between PerDT and UD. We decompose the conversion process into 3 steps: 1) tokenization, 2) part-of-speech mapping, 3) systematic changes to PerDT, and 4) dependency relation mapping. In this section, we briefly describe the mentioned steps.

3.1 Tokenization

There are two key differences in PerDT tokenization from UD: 1) Multiword inflections of simple verbs in Persian are grouped as one word with spaces in between parts following the deterministic rules from (Rasooli et al., 2011a). We follow the guidelines in (Rasooli et al., 2013, Table 3) to find out the main verb and make other parts an “aux” dependent of the main verb. We introduce the “AUX part-of-speech tag and “aux:pass” for passive verbs) in this UD tokenization scheme. 2) Clitics are only detached from words in cases for which they play an object or verbal role. Other clitics are pronominal clitics attached to nouns, prepositions, pronouns and adjectives. By looking at the word lemma, we recover those pronouns, and detach them, and assign their heads to the closest nominal word with the “MOZ” (Ezafe) dependency label.

3.2 POS Mapping

This is the most straightforward step except for proper nouns. We could only discover a small portion of them by finding noun phrases with an identifier (IDEN POS tag for words such as “Dr.” or “Mr.”). In addition to mapping the IDEN POS to PROPN, we use a recent BERT-based Persian named-entity tagger (Taher et al., 2020) to recover additional proper nouns. The tagger can find 7 different entities including date, location, money, organization, percent, person and time. We only consider the person and location entities, and manually revise the results to add missing entities, foreign words and the name of months. Table 1 shows the mappings.

3.3 Systematic Changes to PerDT

Before starting to convert the treebank (Dadegan Research Group, 2012), we have made the following systematic changes to PerDT:
Table 1: Mapping rules for part-of-speech tags.

Table 2: Mapping rules for dependencies. PerDT labels are described in Rasooli et al. (2013, Table 2). First Preconditions (2nd column) should satisfy. Afterwards, Preactions (3rd column) are applied before applying the UD conversions (4th column). These preactions are depicted in Figure 3.

Figure 2: The result of applying rotations of conjunctions for the sentence Why didn’t you defend in the court and acted like that? in this example, case rotation for preposition is also shown.

We convert the order of verbal conjunctions in the original data. In PerDT, verbal conjunctions are conventionally attached from the end to the beginning (Dadegan Research Group, 2012). We find this convention unintuitive and reverted the order of conjunctions. Figure 2 shows an example of such rotation.

Words such as “billion”, “million”, “thousand” are tagged as nouns. This might be due to the fact that these words can be inflected as plurals while number should not be inflected in Persian. We believe that a better tagging decision for these words is number since their inflection as plurals is due to a special kind

2 Examples in https://bit.ly/2Mfz1iH
3 Examples in https://bit.ly/2Y105Yv
of zero derivation or conversion numbers to nouns in particular contexts (Booij, 2012).

- PerDT assumes that all inflections of “کدن” [Jodæn] is passive and its lemma is “گدن” [kædæn]. We have changed this assumption and use the superficial lemma for those instances. The decision makes our data similar to the annotations of Seraji et al. (2016).

Table 3 shows the statistics of changes that we have made to the data including systematic changes and fixes to incorrect annotations.

### 3.4 Dependency Relation Mapping

PerDT contains 43 syntactic relations for which many of them cannot easily map to UD. Moreover, conjunctions in PerDT are arranged from the beginning of the sentence to the end in chain-style manner. More importantly, compared to UD scheme for which content words are considered as heads, PerDT assigns prepositions as the head of prepositional phrases and auxiliary verbs as the head of sentences.

Before applying the conversion rules, we label words that are not well-edited and typed as more than one token as goeswith. We then label proper noun phrases that are not syntactically compositional as flat:name. We also analyze complex numbers as flat:num and their coordinating conjunctions as cc dependent of each following word. Afterwards we follow the rules in Table 2. As depicted in the Table, there are conditions that should be satisfied before applying a conversion, and some actions such as flipping a head with its dependent are needed before certain mappings. Finally, we label the few remaining undecided dependencies as dep.

| Correction Type   | #    | %   |
|-------------------|------|-----|
| Lemma             | 3694 | 0.762 |
| POS               | 529  | 0.109 |
| FPOS              | 3693 | 0.762 |
| Dependency head   | 27407 | 5.658 |
| Dependency label  | 18516 | 3.823 |
| Word Form         | 39   | 0.008 |

Table 3: Statistics of PerDT corrections. By systematic, we mean deterministic corrections such as verbal conjunctions (see §3.3 for details).

### 4 Experiments and Analysis

The general statistics of our data vs. the Uppsala treebank (Seraji et al., 2016) are shown in Table 4. We observe that our data is superior in many aspects including size and diversity compared to the Uppsala Treebank (Seraji et al., 2016). The most important fact about PerDT is that its sentences are intentionally sampled in order to cover almost all verbs from the Verb Valency Lexicon (Rasooli et al., 2011b) leading to 3.9 times more verb lemmas than the Uppsala Treebank. Table 5 shows the counts of each dependency label in the converted Data.

### Supervised Parsing

We evaluate the resulting data by training UDpipe V.2 (Straka and Straková, 2017) along with the pre-trained fastText (Grave et al., 2018) embeddings on our data. We also evaluate our models on the Uppsala treebank (Seraji et al., 2016). Table 6 shows the parsing results using a trained model on our data and the Uppsala Treebank evaluated by the CoNLL 2018 shared task evaluation scripts (Zeman et al., 2018). It is worth noting that the goal of this evaluation is not to show which dataset brings in better parsing accuracy: it is clear that the bigger the dataset is, the higher the accuracy can be. Our goal is to show that there is a significant performance difference between the models trained on the two datasets by using the exact same training pipeline. As shown in Table 6, we see that there is a huge tagging and parsing performance difference when we move across the datasets. There are two possible reasons: domain mismatch, and annotation discrepancy. Our analysis show that annotation discrepancy plays an important role here. As described in §A, there are some core incompatibilities between the Uppsala treebank (Seraji et al., 2016) and Universal Dependencies guidelines. Our detailed analysis
Figure 3: A graphical depiction of rotation rules used in this work (see Table 2 for their use cases).

| Label     | Frequency | %  |
|-----------|-----------|----|
| case      | 71118     | 14.1|
| conj      | 23739     | 4.7 |
| acl       | 10034     | 1.9 |
| obl       | 30737     | 6.1 |
| punct     | 44336     | 8.8 |
| cop       | 6366      | 1.2 |
| det       | 10273     | 2   |
| advmod    | 9158      | 1.8 |
| aux/pass  | 822       | 0.1 |
| nmod      | 59442     | 11.6|
| appos     | 1059      | 0.2 |
| aux       | 12886     | 0.16|
| amod      | 22576     | 4.4 |
| compound:lvc | 32339 | 6.4 |
| nsubj:pas | 822       | 0.1 |
| nsubj     | 27181     | 5.4 |
| name:flat | 7899      | 1.5 |
| dep       | 2035      | 0.4 |
| cc        | 21300     | 4.2 |
| root      | 29107     | 5.8 |
| advcl     | 4228      | 0.8 |
| obj       | 19999     | 3.9 |
| xcomp     | 4920      | 0.9 |
| parataxis | 82        | 0.01|
| comp      | 6945      | 1.3 |
| obl:arg   | 21510     | 4.2 |
| flat:num  | 607       | 0.1 |
| nummod    | 5459      | 1   |
| mark      | 11982     | 2.3 |
| fixed     | 144       | 0.02|
| compound:lvc | 439 | 0.08 |
| csubj     | 682       | 0.1 |
| vocative  | 174       | 0.03|
| compound  | 42        | 0.008|
| iobj      | 6         | 0.001|
| dislocated| 1         | 0.0001|

Table 5: Frequency of each universal label the converted dataset.

shows that most of cross-dataset errors come from errors in nmod, obl, fixed, and xcomp. This is in fact consistent with our manual analysis in §A.

Delexicalized Model Transfer One way to verify our claim about increased consistency of our UD conversion with the UD guidelines is to learn a transfer model. In this setting, we follow the delexicalized parser transfer approach which has been extensively used in previous work (Zeman and Resnik, 2008; McDonald et al., 2011; Täckström et al., 2012). We sample the same number of tokens as of Seraji et al. (2016) from PerDT. Afterwards, we delexicalize both of the treebanks, and learn a parser using the Yara Parser (Rasooli and Tetreault, 2015). We train two models with 15 epochs and evaluate them on the delexicalized test set of the Universal English Web Treebank (Silveira et al., 2014). The model trained on PerDT significantly outperforms the other model by 2% both in unlabeled and labeled attachment score (47.31 vs 45.37 UAS, 38.59 vs. 36.45 LAS). This is a strong indicator that our data is more compatible with the UD annotations.

5 Conclusion
We have introduced our approach in making PerDT (Rasooli et al., 2013) universal. During this process, we have faced different challenges such as annotation errors in the original data, tokenization inconsistencies, lack of named entities, part-of-speech and dependency label mapping. Due to automatic conversions and potential annotation errors in the original treebank, there is always a chance of some annotation incompatibilities between our treebank and the Universal guidelines. Therefore, we cannot claim that our conversion is perfect. How-
ever, our experiments have shown that our data is more compatible with the Universal Dependencies guidelines than the Uppsala treebank (Seraji et al., 2016).

Acknowledgments

We would like to thank Morteza Rezaei-Sharifabadi for his efforts in getting copyright permissions and making our treebank publicly available. We also thank Daniel Zeman for helping us in the process of releasing this treebank.

Appendix

A Problems in the Universal Annotations of the Uppsala Universal Treebank

We briefly mention some of the problems in the Uppsala Universal Treebank (Seraji et al., 2016):

- Seraji et al. (2016) does not determine the csubj label in their analysis. For example, in “lAzem ?ast ?u beresæd” (it is necessary for him to arrive), it is obvious that what comes after “?ast” is the clausal subject of the adjectival sentence predicate “lAzem”. A simple syntactic test supports this viewpoint: one can convert the clausal complement “?u beresæd” to a noun phrase “residan-e u” (his arrival). The new phrase plays the nsubj role of the sentence. Therefore, the clausal complement of the sentence should be csubj. Our converted data contains 682 cases of csubj.

- Seraji et al. (2016) considers prepositional and possessive complements of adjectival heads as nmod and nmod:poss respectively. Their analysis clearly stands in contradiction to UD annotation guideline in which nmod is used just for dependents of a nominal head. obl is much better suited for these cases.

- Seraji et al. (2016) consider “peydA kærdæn” (to become) as copula. What UD asserts under the cop (copula) label is that “the equivalents of to become are not copulas despite the fact that traditional grammar may label them as such.” Instead, it should be deemed as a verbal predicate and its second complement as xcomp.

- “حاصل کردن” (“peydA kærdæn”) and “پیادا کردن” (“høsl kærdæn”) are considered as two-word light verbs (Mołoodi and Kouhestani, 2017). We consider the non-verbal part as the first part of the two-word light verb, and use the compoundlv label for it (439 cases in PerDT). However, Seraji et al. (2016) annotate the non-verbal elements of these complex predicates as obj and considers “peydA” as a nonverbal element.

- iobj label is absent in (Seraji et al., 2016), most likely due to the low frequency of this syntactic relation. Our converted treebank contains 6 cases of iobj.

- Proper nouns are not labeled in (Seraji et al., 2016). Ours covers proper nouns (more than 23K tokens).

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. Transactions of the Association for Computational Linguistics, 4:431–444.

Geert Booij. 2012. The grammar of words: An introduction to linguistic morphology. Oxford University Press.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. arXiv preprint arXiv:2005.04511.

Dadegan Research Group. 2012. Persian dependency treebank, annotation manual and user guide. Supreme Council of Information and Communication Technology (SCICT), Tehran, Iran.

Weston Feely, Mehdi Manshadi, Robert Frederking, and Lori Levin. 2014. The CMU METAL Farsi NLP approach. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 4052–4055, Reykjavik, Iceland. European Language Resources Association (ELRA).

Masood Ghayoomi and Jonas Kuhn. 2014. Converting an HPSG-based treebank into its parallel dependency-based treebank. In LREC, pages 802–809.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan. European Languages Resources Association (ELRA).

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language
Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. 2012. Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7(18).

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2897–2904, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Milan Straka and Jana Straková. 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. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Ehsan Taher, Seyed Abbas Hoseini, and Mehrnoush Shamsfard. 2020. Beheshti-NER: Persian named entity recognition using BERT. *arXiv preprint arXiv:2003.08875*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Jan Hajíč, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. *CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*. Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 997–1006, Hong Kong, China. Association for Computational Linguistics.