# Initial Efforts in Creating a Persian-English Parallel TreeBank

Razieh Ehsani and Olcay Taner Yıldız

Işık University, Istanbul, Turkey

**Abstract.** In this paper, we introduce our preliminary efforts in constructing Persian-English parallel treebank corpus. We extract 1,500 sentences from Penn Treebank, where a sentence contains 15 tokens at maximum including punctuation. In our approach, we replace English words with Persian equivalents and reorder subtrees without changing Penn tags.

## 1 Introduction

Not all languages have well developed NLP resources as English. Persian is a language that is spoken widely in Iran, Afghanistan, Tajikestan. Although there are some works on Persian; it still remains mostly unexplored. One of the available works is the Bijankhan Corpus that contains nearly 2,000,000 words with their morphological tags.

In recent years, many efforts have been made to annotate parallel corpora with syntactic structure to build parallel treebanks. Well-known parallel treebank efforts are

- Prague Czech-English dependency treebank annotated with dependency structure [1]
- English-German parallel treebank, annotated with POS, constituent structures, functional relations, and predicate-argument structures [2]
- Linköping English-Swedish parallel treebank that contains 1,200 sentences annotated with POS and dependency structures [3]
- Stockholm multilingual treebank that contains 1,000 sentences in English, German and Swedish annotated with constituent structure [4].

Persian Treebank creation efforts started with the HPSG-based treebank [5]. They use rules in the CLaRK system, then complete the annotation of sentences manually. The treebank contains 1000 sentences.

Uppsala Persian Dependency Treebank (UPDT) is another corpus based on dependency annotation [6]. The treebank contains 6000 sentences which has been developed through a bootstrapping procedure involving the open source data-driven dependency parser MaltParser [7].This treebank is based on Stanford typed dependencies (STD) [8]. They also adopted other syntactical relations which do not exist in STD categories. These relations are mostly related to different verb forms specific to Persian and some additional tags such as “foreign word”.

This work follows the same approach which is used to construct a Turkish-English parallel treebank [9]. This paper reports the results of an initial effort in preparing a Persian-English parallel treebank. We convert English parse trees to equivalent Persian

parse trees in two steps: (i) The tree permutation, where we change order of children of a node, (ii) gloss replacement, where we change English word in a leaf to Persian word(s).

This paper is organized as follows: In Section 2, we give a very brief overview on Persian syntax. We give the details of our corpus construction strategy in Section 3 and explain our transformation heuristics in Section 4. Finally, we conclude in Section 5.

## 2 Persian syntax

Persian language is a member of Indo-European family that uses a modified Arabic script and is written from right to left. As a result of this, processing Persian language becomes difficult. Beside this, tokenization is also difficult since delineating spaces are not consistently used. Persian has a flexible word order. Most commonly, sentences follow a subject-object-verb ordering. However, the ordering may change according to the emphasis. For example, if the emphasis is on the subject, one can use object-subject-verb ordering.

Forming question sentences is also different than English, in the sense that the structure of the sentence does not necessarily change. In Persian, affixes can come before or after word roots. Like English, verb form changes according to tense.

## 3 Corpus construction strategy

During corpus construction we selected 9,560 trees from Penn Treebank II which have maximum of 15 tokens. We translated 1,500 trees from this subset.

### 3.1 Tools

Manual annotation is a hard and painful task, and one that is also error-prone. These errors may relate to disagreement among annotators or occur due to inconsistent documentation. To reduce these errors and make annotation easier for the annotator, we designed visual tools. These tools display, manipulate and save annotated trees in treebank. The annotated trees are stored as a text file which uses Treebank II style of syntactic bracketing.

In our tool, we display possible glosses in Persian for each English word. These glosses are presented according to their usage counts in the previous cases. Annotators can choose one of the glosses in the list or they can write an alternative gloss. As a result of this, as the corpus grows in size, the annotators leverage their previous choices. This simple statistical function helps annotators to translate easier and speedy.

### 3.2 Tree permutation

The main step in our project is to convert trees such that they follow Persian syntactic structure. This step contain two operations, the first one is the permutation of children of a node and the other one is replacing the English word token at a leaf node. In fact,

we use same set of tags and predicate labels in the non-leaf nodes without using new tags for Persian trees.

This transformation is not exactly a restriction in translation. Indeed, it is very easy to construct pairs of translated sentences which involve operations outside our restricted set when transformed into each other.

Some tokens in English do not have a meaning in Persian. For these cases, we use *NONE* tag. For example, we replace “The” with *NONE* tag in translation. As a result of this, one can reduce a subtree in English to a token in Persian.

We see an example case in Figure 1. First, the noun order is reversed, “west countries” become “countries (of) west” in Persian. Second, “the” has no corresponding word in Persian, therefore it is replaced with *NONE*. Third, “their” under the PRP tag becomes a postfix “rā” with a NNS tag in Persian. Fourth, verb group “will redouble” is translated as “afzāyesh khāhand dād” under the VBG tag.

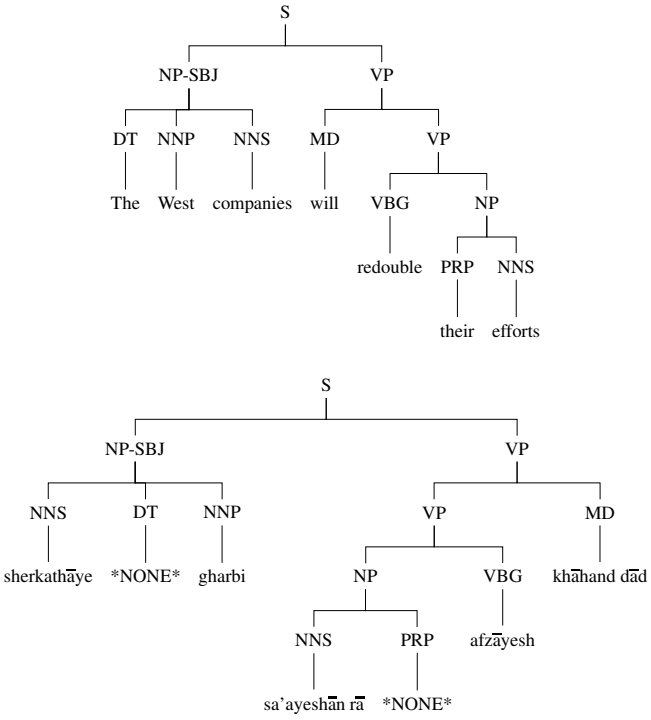


**Fig. 1.** The permutation of the nodes and the replacement of the leaves by the glosses or *NONE*. sherkathāye gharbi sa’ayeshān rā afzāyesh khāhand dād

## 4 Transformation heuristics

When we have a sufficiently rich corpus of parallel trees, our next step is to train a SMT learner to imitate the human translator who operates under our restricted set of operations. Naturally, human translators often base their transformation decisions on the whole tree. Still, having a common set of rules and heuristics helps the translators in both consistency and speed. In the following, we illustrate these heuristics.

### 4.1 Constituent order and *NONE* cases

In Section 2 we mentioned that majority of Persian sentences have the SOV order. Based on this fact, when translating English trees, we permute its subtrees to reflect change of constituent order in Persian.

Some suffixes in Persian correspond to an exact word in English. For example: "Ali//NNP did/VBP Not/RB read/VB his/PRP book/NN" is translated as Ali na-khand ketab-ash rā .
Ali not-read book-his ... .

It is obvious that some English words will not have a corresponding form in the Persian side, so in these cases, we replace the English constituent leaf with *NONE*. In some cases, the personal pronouns acting as subjects (sometimes also objects) are naturally embedded in the verb. In those cases, pronoun in the original tree is replaced with *NONE* and its subtree is moved after the verb phrase.

### 4.2 Case markers

Persian uses case marker "rā" to denote the syntactic functions of nouns and noun groups. For example, accusative case may be used to mark the direct object of a transitive verb and locative case may be used to mark the head of a prepositional phrase. In English, there are no tokens corresponding to "rā". When the direct object is definite it is always followed by "rā"; when the direct object is indefinite but individuated it may or may not be followed by "rā" under certain conditions [10]. We add this token after direct object. For example:
Sara book-ez poem-ez Hafez rā read.
Sara read the poem book of Hafez.
accusative(ketāb, -rā) accusative(book,-rā)

Also, in translation from English to Persian, the prepositions are sometimes replaced with *NONE*. See Figure 2 as an example.

### 4.3 Plural in nouns and verb inflection

Subject-Verb number agreement is optional in Persian. In translation, we used the case that sounds more natural. Also, plural nouns under NNS tag in the English tree are sometimes translated as singular. See Figure 3 as an example.

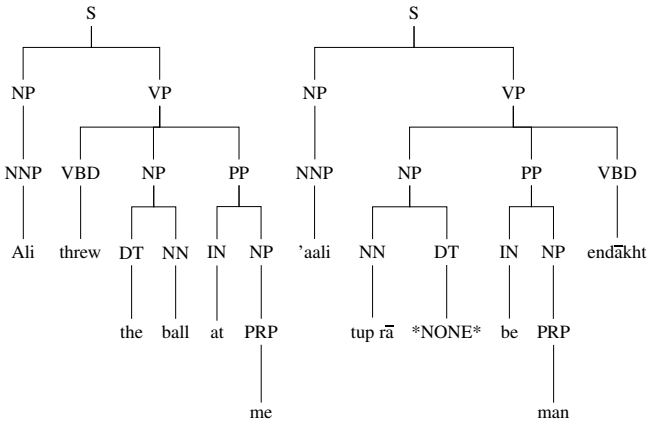


**Fig. 2.** Original and translated trees,Ali ball-ACC DAT-me throw-PAST-3SG'aali tup rā be man endākht.

### 4.4 Tense and Auxiliary Verb ambiguity

In general, tenses in English are different than tenses in Persian. We translate English tenses to corresponding tenses in Persian. If there is no corresponding tense in Persian, we translate in to the closest semantic class in Persian.

When a compound verb comes with an auxiliary verb, another difficult case arises. In that case, auxiliary verb comes between head of compound verb and light verb. Although auxiliary verb(s) have corresponding verb(s) in Persian, we did not translate them and bring them under VB with the main verb.

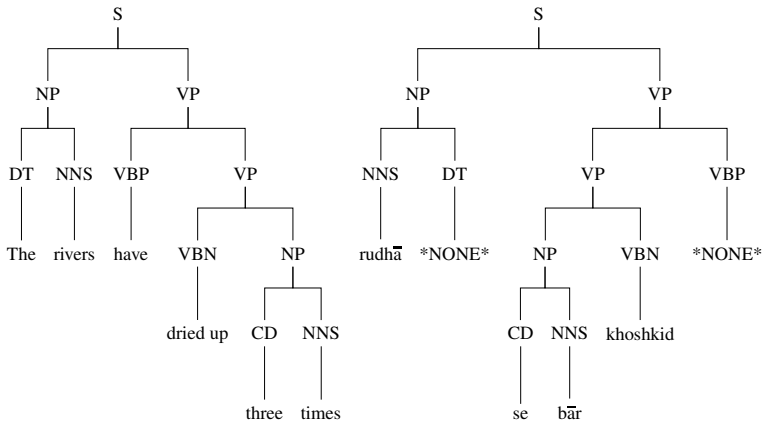


**Fig. 3.** Original and translated trees, river-PL three time dry up-PAST-3SG rudhā se bār khoshkid.

### 4.5 WH- Questions

There are significant differences between English question sentences and Persian question sentences. For this case, we benefit from Penn Treebank II annotation where the movement leaves a trace and is associated with wh- constituent with a numeric marker. For example, “WHNP-1” and “*T*-1” are associated.

During translation, wh- constituent is replaced with *NONE* and its trace is replaced with related question pronoun in Persian. We see an example case in Figure 4. The three words under the -NONE- tag are a translation of “to what thing”. Persian language makes extensive use of composite verbs, here the corresponding word to “believe” is the composite verb “motaged ast”, i.e. “believer is” (to be a believer). Here, “motaged” is tagged under VB (“believer”) and the helper verb of the composite is tagged under VBZ.

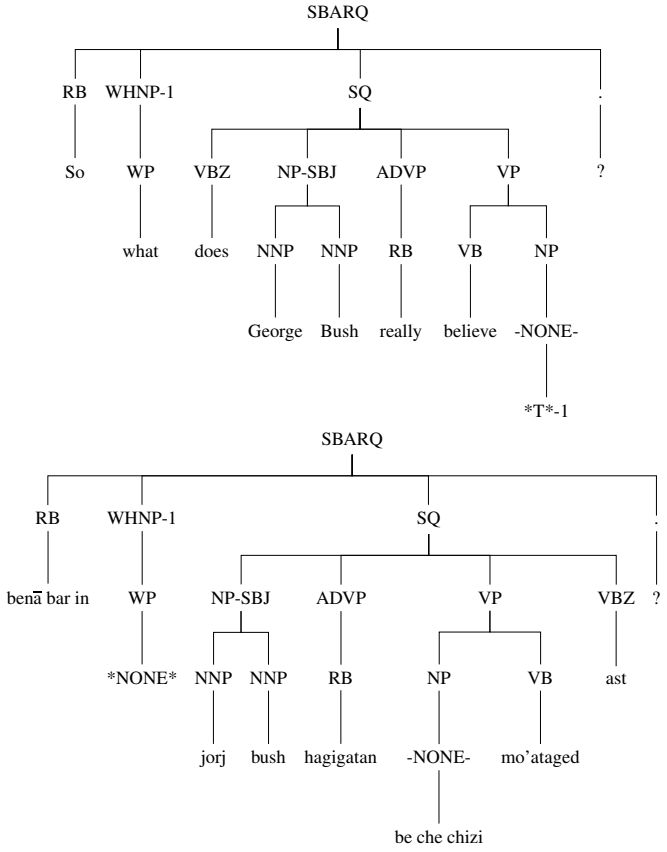


**Fig. 4.** Original and translated trees, So George Bush really what-DAT believe-PRES-3SG? benā ber in jorj bush hagigatan be che chizi mo’ataged ast?

## 5 Conclusion

Parallel treebank construction efforts increased significantly in the recent years. Many parallel treebanks are produced to build statistically strong language models for different languages.

Although Persian language have many speakers in the world, there are still many gaps in NLP for Persian. In order to help to fill this gap, we give our preliminary efforts to build such a parallel corpus for Persian-English pair. We translated and transformed a subset of parse trees of Penn Treebank to Persian.

This work constitutes the preliminary step of parallel treebank generation. As a next step, we will focus on pos tagging of Persian words. As a future work, we plan to expand the dataset to include all Penn Treebank sentences.

## Acknowledgments

This project consumed huge amount of work in translation and annotation. Still, implementation would not have been possible if we did not have a support of Yashar Hosseini Baghanam and Soghra Mehdinejad Gargari. Therefore we would like to extend our sincere gratitude to them. This work was supported by Işık University BAP project 15B201.

## References

1. Cmejrek, M., Haji, J., Kubo, V.: Prague czech-english dependency treebank: Syntactically annotated resources for machine translation. In: In Proceedings of EAMT 10th Annual Conference. (2004) 04
2. Cyrus, L., Feddes, H., Schumacher, F.: FuSe – a multi-layered parallel treebank. In Nivre, J., Hinrichs, E., eds.: Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 14–15 November 2003, Växjö, Sweden (TLT 2003). Volume 9 of Mathematical Modelling in Physics, Engineering and Cognitive Sciences., Växjö, Växjö University Press (2003) 213–216
3. Ahrenberg, L.: Lines: An english-swedish parallel treebank (2007)
4. Gustafson-Capkova, S., Samuelsson, Y., Volk, M.: Smultron (version 1.0) - the stockholm multilingual parallel treebank. an english-german-swedish parallel treebank with subsentential alignments. (2007)
5. Ghayoomi, M., Ghayoomi, M., Berlin, F.U.: Bootstrapping the development of an hpsg-based treebank for persian. In: Linguistic Issues in Language Technology: LiLT. (2012)
6. Seraji, M., Jahani, C., Megyesi, B., Nivre, J.: A persian treebank with stanford typed dependencies. In: The 9th International Conference on Language Resources and Evaluation (LREC). (2014) 796–801
7. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering **13** (2007) 95–135
8. De Marneffe, M.C., Manning, C.D.: The stanford typed dependencies representation. In: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, Association for Computational Linguistics (2008) 1–8

9. Yıldız, O.T., Solak, E., Görgün, O., Ehsani, R.: Constructing a turkish-english parallel treebank. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Volume 2. (2014) 112–117
10. Meisami, J.S.: A grammar of contemporary persian. by gilbert lazard, translated by shirley a. lyon.(persian studies series, 14.) pp. xii, 301. costa mesa, calif, and new york, mazda publishers in assoc. with bibliotheca persica, 1992. us 19.95. an introduction to persian. revised third edition. by wm thackston. pp. xxxvi, 287. bethesda, maryland, iranbooks inc., 1993. us 25.00. Journal of the Royal Asiatic Society (Third Series) **5** (1995) 117–119