Development of a Persian Syntactic Dependency Treebank
Mohammad Sadegh Rasooli¹, Manouchehr Kouhestani², and Amirsaeid Moloodi³
¹rasooli@cs.columbia.edu ²m.kouhestani@modares.ac.ir ³a.moloodi@ut.ac.ir
¹Columbia University ²Tarbiat Modares University ³University of Tehran

Objectives
There was a lack of syntactically annotated data. We tried to create a valuable linguistic data set for the Persian language.
- Second linguistic product by Dadegan research group after valency lexicon of Persian verbs [1]
- 30,000 manually annotated sentences.
- The largest syntactic treebank for Persian.
- Extendable to semantic treebank.
- Persian is
  - An Indo-European language.
  - Spoken by more than 100 million speakers.
  - Rich morphology and free word order.
  - An under-resourced language.

Why Dependency Trees?
- Dependency representation is useful for showing
  - Non-projective trees.
  - Compound verbs in Persian.
  - Convertible to phrase-structure trees.

Annnotation Process
We used a bootstrapping approach to annotate the data.

Two Different Representations
There are two possible representations for objects accompanied by the case marker.
- Case marker as a post-position is the head of the object phrase.
  - Creates more non-projective trees.
  - Simplifies the search for objects (closer to the verb than the object and the object should come before it).
- Object is the head of the case marker.
  - Closer to the human interpretation.
  - This representation is provided by automatic conversion from the first representation.

Why Dependency Trees?
- Dependency representation is useful for showing
  - Non-projective trees.
  - Compound verbs in Persian.
  - Convertible to phrase-structure trees.

Statistics
- 29,982 sentences; 498K tokens, and 37K types.
- Avg. sentence length: 16.6
- Number of distinct verbs: 4.7K
- 44 dependency relations
- 17 coarse-grained part of speech tags
- 1.8% non-projective sentences (0.02% non-projective arcs)

Contact Information
• Web: http://www.dadegan.ir/en/perdt
• Email: info@dadegan.ir

Acknowledgements
The project is funded by Iran Supreme Council of Information and Communication Technology (SCICT).

References

Figure 1: An example of free-word order in Persian.
Figure 2: Representation of Persian verbs in dependency trees.
Figure 3: Diagram of bootstrapping approach in the development of the dependency treebank.
Figure 4: A sample sentence with two kinds of representations of object-verb relation.
Figure 5: Statistics about changes in the treebank after the manual correction of the potential errors.
Figure 6: Statistics about agreements among the annotators.
Figure 7: Dadegan dependency treebank search tool.

Online Treebank Search
An online tool for searching dependency relations
http://search.dadegan.ir/advance/

Future Direction
- Create other resources such as SRL treebank.