Development of a Persian Syntactic Dependency Treebank

Mohammad Sadegh Rasooli[1], Manouchehr Kouhestani[2], and Amirsaeid Moloodi[3]

[1]rasooli@cs.columbia.edu [2]m.kouhestani@modares.ac.ir [3]a.moloodi@ut.ac.ir
[1]Columbia University [2]Tarbiat Modares University [3]University of Tehran

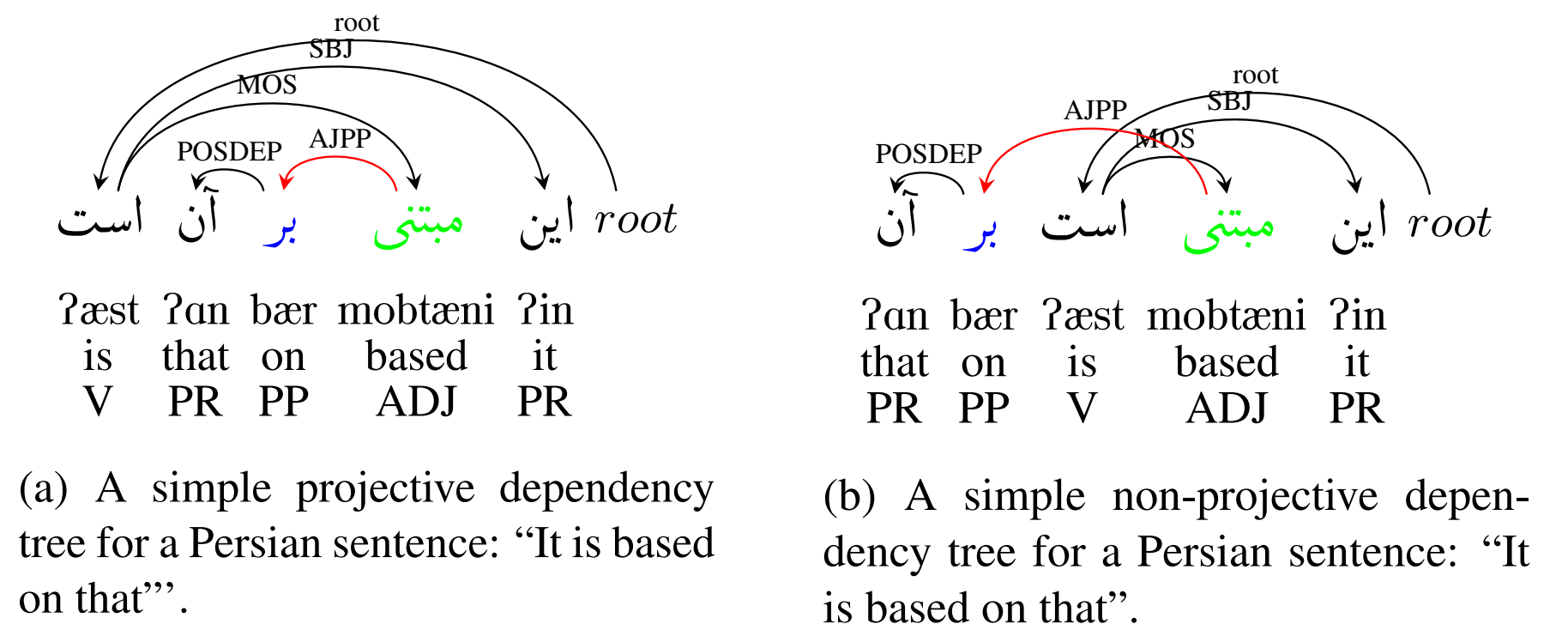
Objectives

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 speaker.
  - Rich morphology and free word order.
  - *An under-resourced language.*

Why Dependency Trees?

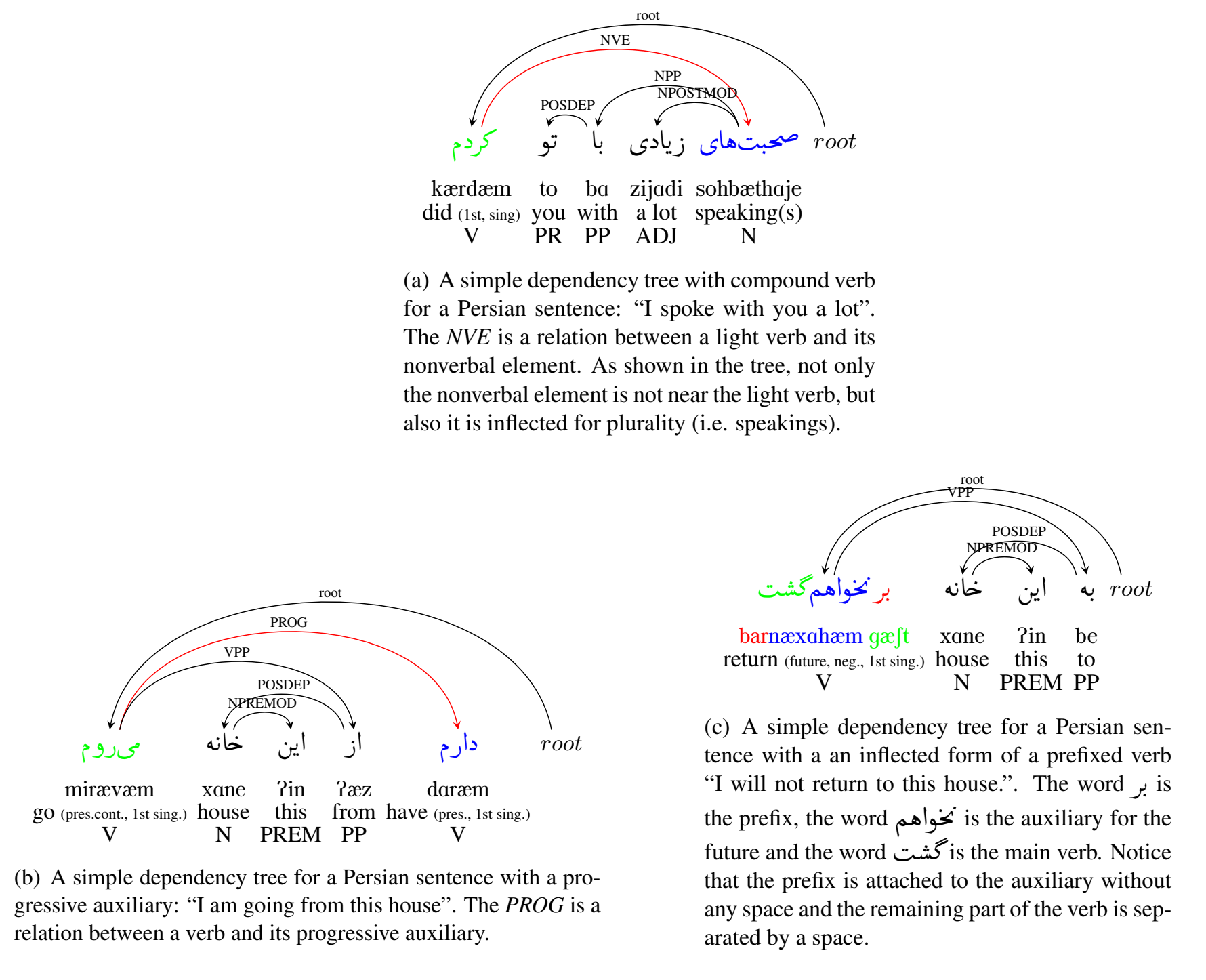
- Dependency representation is useful for showing
  - Non-projective trees.
  - Compound verbs in Persian.
- Convertible to phrase-structure trees.



(a) A simple projective dependency tree for a Persian sentence: "It is based on that"".

(b) A simple non-projective dependency tree for a Persian sentence: "It is based on that".

Figure 1: An example of free-word order in Persian.


root
NVE
NPP
NPOSTMOD
POSDEP
کردم تو با زیادی صحبت‌های root
kærdæm to bɑ zijɑdi sohbæthɑje
did (1st, sing) you with a lot speaking(s)
V PR PP ADJ N
root
PROG
VPP
POSDEP
NPREMOD
می‌روم خانه این از دارم root
miræværm xɑne ʔin ʔæz dɑræm
go (pres.cont., 1st sing.) house this from have (pres., 1st sing.)
V N PREM PP V
root
VPP
POSDEP
NPREMOD
برنخواهم گشت خانه این به root
barnæxɑhæm gæʃt xɑne ʔin be
return (future, neg., 1st sing.) house this to
V N PREM PP

(a) A simple dependency tree with compound verb for a Persian sentence: "I spoke with you a lot". The *NVE* is a relation between a light verb and its nonverbal element. As shown in the tree, not only the nonverbal element is not near the light verb, but also it is inflected for plurality (i.e. speakings).

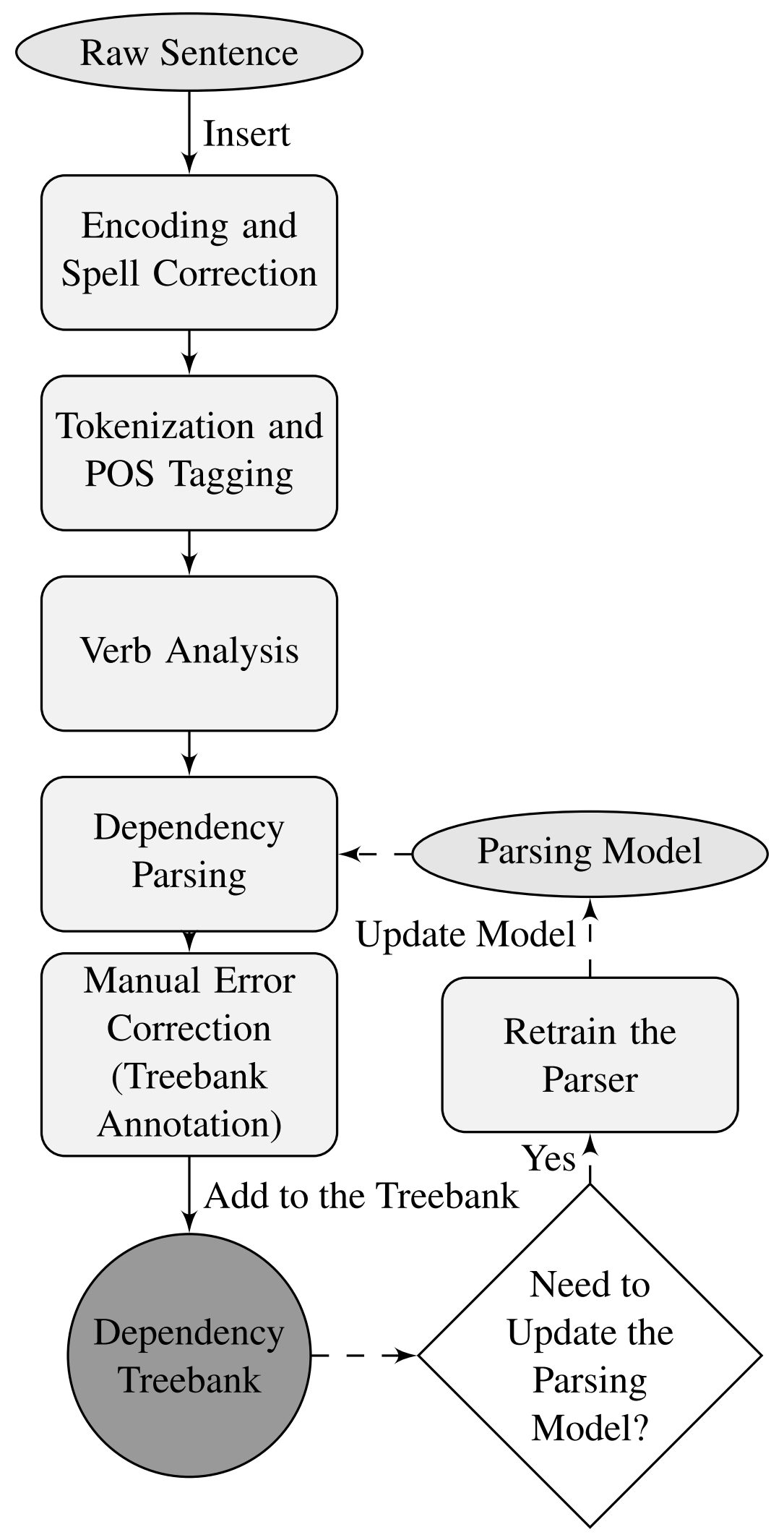
(b) A simple dependency tree for a Persian sentence with a progressive auxiliary: "I am going from this house". The *PROG* is a relation between a verb and its progressive auxiliary.

(c) A simple dependency tree for a Persian sentence with a an inflected form of a prefixed verb "I will not return to this house.". The word بر is the prefix, the word نخواهم is the auxiliary for the future and the word گشت is the main verb. Notice that the prefix is attached to the auxiliary without any space and the remaining part of the verb is separated by a space.

Figure 2: Representation of Persian verbs in dependency trees.

Annotation Process

We used a bootstrapping approach to annotate the data.



Raw Sentence
Insert
Encoding and Spell Correction
Tokenization and POS Tagging
Verb Analysis
Dependency Parsing
Parsing Model
Update Model
Manual Error Correction (Treebank Annotation)
Retrain the Parser
Yes
Add to the Treebank
Dependency Treebank
Need to Update the Parsing Model?

Figure 3: Diagram of bootstrapping approach in the development of the dependency treebank.

- **Annotators**
  - Started with 8 annotators and gradually increased them to 12 people.
  - 12 months of annotation.
- **Tools**
  - Virastyar [2] for spell checking, lemmatization and POS tagging.
  - Persian verb analyzer [3] for recognizing and lemmatizing verbs.
  - MST parser [4] for parsing input sentences.

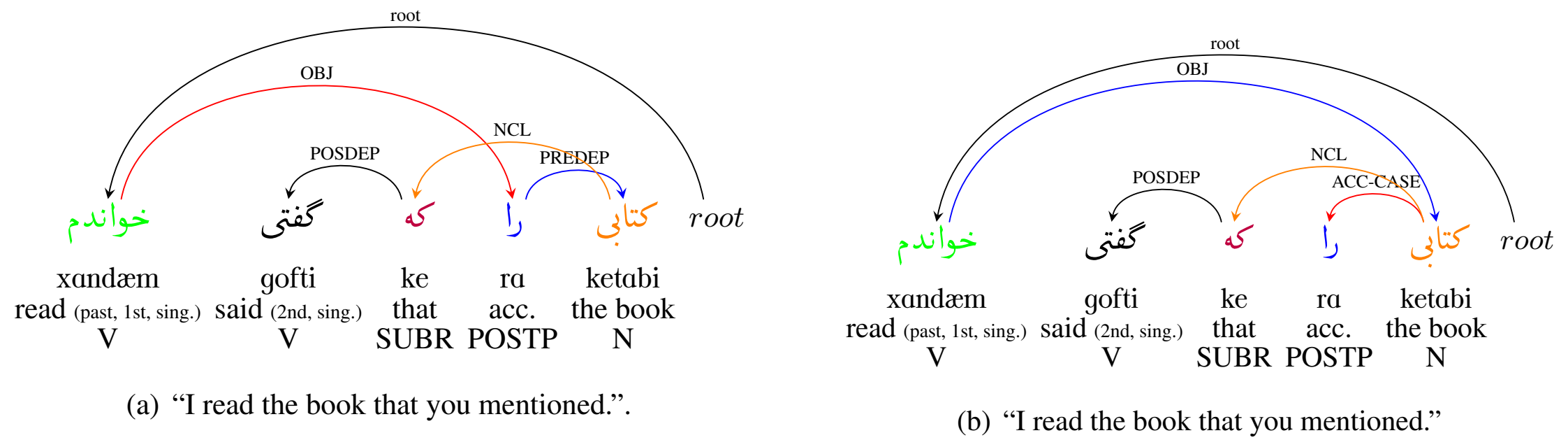
Statistics

- 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)

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


root
OBJ
NCL
POSDEP
PREDEP
ACC-CASE
خواندم گفتی که را کتابی root
xɑndæm gofti ke rɑ ketɑbi
read (past, 1st, sing.) said (2nd, sing.) that acc. the book
V V SUBR POSTP N

(a) "I read the book that you mentioned.".

(b) "I read the book that you mentioned."

Figure 4: A sample sentence with two kinds of representations of object-verb relation.

Correcting Potential Errors

- We provided additional scripts for finding potential errors;
  - E.g. annotation mistakes such as a verb as a subject for a noun or mismatch with valency lexicon information.

Changes to Unlabeled Relations	4.91%
Changes to Labeled Relations	6.29%
Changes to POS Tags	4.23%

Figure 5: Statistics about changes in the treebank after the manual correction of the potential errors.

Annotators' Agreement

5% of the data is doubly annotated.

Unlabeled Relations	97.06%
Labeled Relations	95.32%
POS Tags	98.93%

Figure 6: Statistics about agreements among the annotators.

Future Direction

- Create other resources such as SRL treebank.

Online Treebank Search

An online tool for searching dependency relations

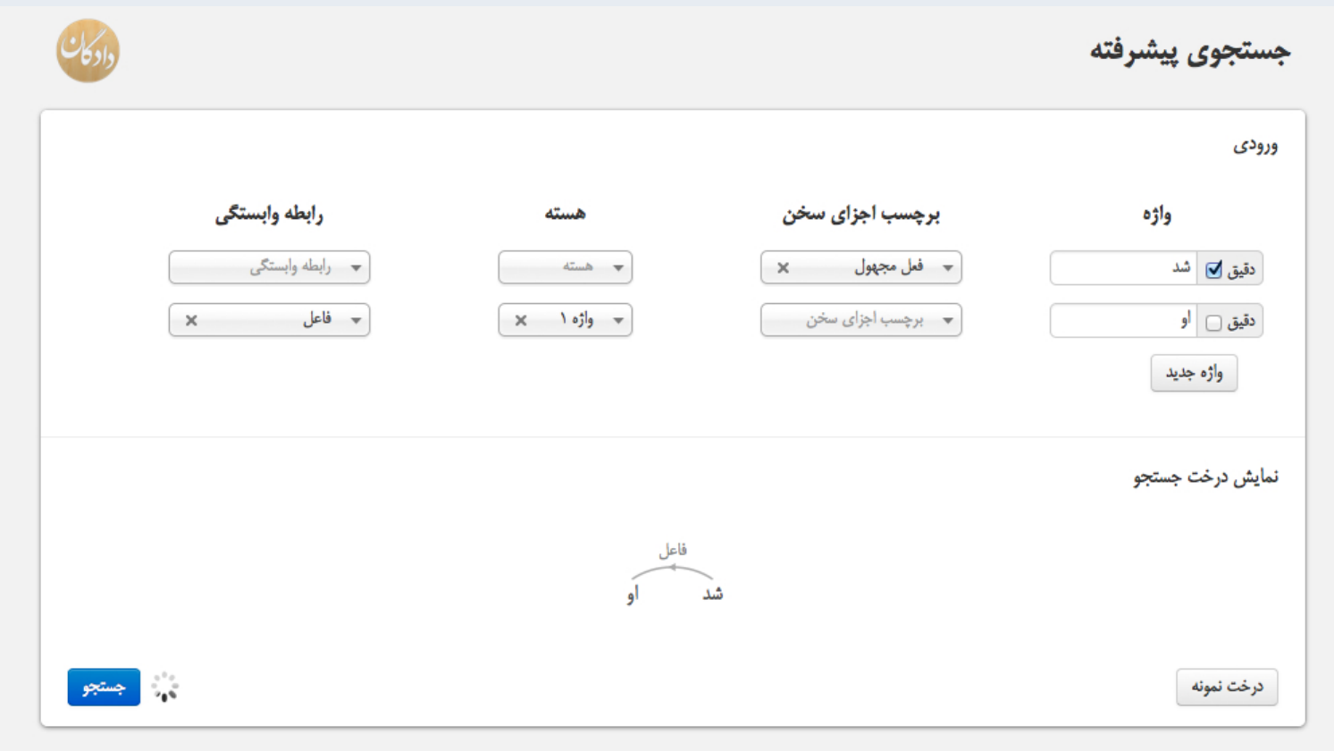


Figure 7: Dadegan dependency treebank search tool.

http://search.dadegan.ir/advance/

References

[1] M. S. Rasooli, A. Moloodi, M. Kouhestani, and B. Minaei-Bidgoli.
A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank.
In *LTC*, pages 227–231, Poznań, Poland, 2011.

[2] O. Kashefi, M. Nasri, and K. Kanani.
Automatic Spell Checking in Persian Language.
SCICT, 2010.

[3] M. S. Rasooli, H. Faili, and B. Minaei-Bidgoli.
Unsupervised identification of Persian compound verbs.
In *MICAI*, pages 394–406, Puebla, Mexico, 2011.

[4] R. McDonald, K. Crammer, and F. Pereira.
Online large-margin training of dependency parsers.
In *ACL*, pages 91–98, Sydney, Australia, 2005.

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).