

A link grammar parser for Persian

Jon Dehdari & Deryle Lonsdale

Brigham Young University

Department of Linguistics

Provo, UT, USA

{jonsafari, lonz}@byu.edu

- Link grammar syntax
- Design of links and lexicon
- Current status
- Applications & future development

- **Motivation**

- Build fast, robust, freely available syntax parser for Persian
- Investigate theoretical issues of linearity at a morphemic level

- **Previous work** – Shiraz project: unification-based, bidirectional chart parser (Amtrup et al., 2000a)

- Builds simple, explicit relations between pairs of words, rather than constructing constituents in tree-like hierarchy
- Basic parameters:
 - Directionality
 - Distance
- English example:

```

                +---0---+
+---D---+---S---+   +-D--+
|       |       |   |   |
the  student read  a  book
```

- Originally by Sleator and Temperley (1993) for English
- Fast, robust, portable, freely available
- Currently used in speech processing, information extraction, essay grading, A.I.

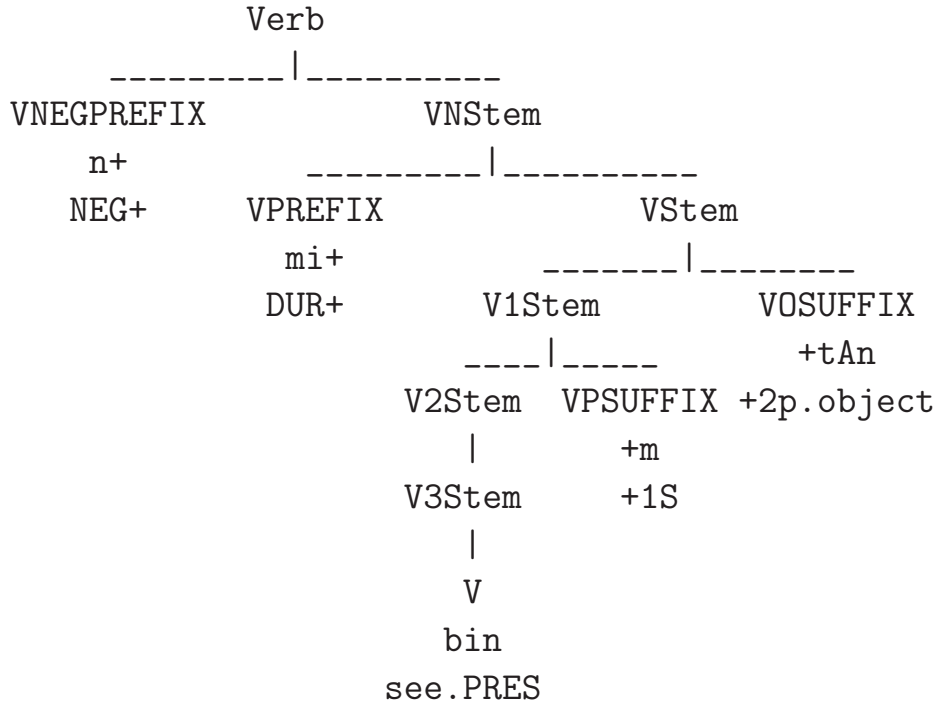
- Input is romanized and morphologically decomposed before syntax
- Accepts UTF-8, CP 1256, ISIRI 3342, Unicode HTML decimal, and romanized text
- Currently two optional morphology parsers:
 - PC-Kimmo (Koskenniemi, 1983; Antworth, 1990)
 - Stemmer

Two-level engine resolves surface form and lexical forms

نميينمتان

PC-KIMMO>recognize nmibinmtAn

n+mi+bin+m+tAn NEG+DUR+see.PRES+1S+2p.object



- Uses Perl regular expressions
- Fast, lightweight, broad coverage

نمییمنتان → nmibinmtAn → n+_mi+_bin+_m_+tAn $\left\langle \begin{matrix} \text{bin} \\ n \text{ mi bin m tAn} \end{matrix} \right.$

nmi-guim → n mi gu m (نمی گویم)

nmiguiim → n mi gu im (نمیگویم)

ktAb hAi → ktAb hA e (کتاب های)

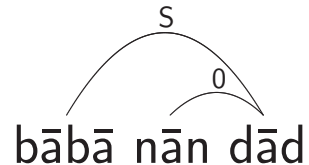
ktAbhAitAn → ktAb hA tAn (کتابهایتان)

- Built using word frequency lists derived from various corpora
- Multipurpose design
- Vowelled lexicon includes categories, glosses, valency, etymology

- General SOV format:

words.n: 0+ or S+;

words.v: {0-} & {S-};



- Persian link grammar sample for verbs:

words.v UNKNOWN-WORD.v:

(({VMdur-} & {VMneg-}) or {VMbe-}) & {@AV-} & {0- or CCOB-}

& {@AV-} & {PP-} & {@AV-} & {VMT+} & (VMP+ or CCF+ or VMPP+ or [RW+]);

- Links inflexional morphemes together just like words
- Morphemes which are adjacent, separated by ZWNJ, or by space are linked equally

نمی بینمتان

persianparse 'nmi binmtAn'

```
+----VMneg----+
|      +-VMdur+-VMP-+---VMO--+
|      |      |      |      |
n.vmn mi.vmd bin.vp m.vmp tAn.vmo
```

کتاب‌هایتان

persianparse 'ktAb-hAitAn'

```
+-----M-----+
+--NMS--+      |
|      |      |
ktAb.n hA.nms tAn.pme
```

Parser can handle nouns & verbs not found in the lexicon

```
fdsaf fdsafasf fdsafftim:  
fdsaf fdsafasf fdsaff t im
```

```
      +-----On-----+-----VMP-----+  
      +-----M-----+           +---VMT---+   |  
      |               |           |         |   |  
fdsaf[?].n fdsafasf[?].n fdsaff[?].v t.vmt im.vmp
```

```
fdsaf fdsafasf krdim:  
fdsaf fdsafasf kr d im
```

```
      +-----On-----+-----VMP-----+  
      |               +-----K-----+---VMT---+   |  
      |               |           |         |   |  
fdsaf[?].n fdsafasf[?].nk kr.vk d.vmt im.vmp
```

Coordination, subordination, & relative clauses handled recursively

کتابی را که دیروز خریده بودم امروز صبح تمام کردم

(Amtrup et al., 2000b)

```
echo -e 'ketAbi rA keh diruz xarideh budam emruz SobH tamAm kardam' | stemmer.pl -R -u | persianlg
```

```
+-----On-----+
+-----REL-----+-----C-----+-----AV-----+
+----PA----+ | +----VMPP----+ +----VMP-----+ | +----AV-----+----VMP-----+
+-NMSi+ | | +--AV--+VMT--+ +--VPPP--+VMT--+ | | | +----K--+VMT--+ |
| | | | | | | | | | | | | | | | | |
ktAb.n i.nms rA.acc kh.sub diruz.av xri.v d.vmt h.per bu.vpperf d.vmt m.vmp |mruz.av SbH.av tmAm.ajk kr.vk d.vmt m.vmp
```

سخت است ولی فکر می‌کنم که آسانتر خواهد شد

adapted from (Megerdooonian, 2000)

```
persianparse 'saxt ast uali fekr mi-konam keh AsAntar xuAhad Cod'
```

```
+-----C-----+
+-----CC-----+ | +-----P-----+
| +-----K-----+ | | | +----VFUT----+
+----P--+CCF--+ | +--VMdur+VMP+--SUB--+ +--AJM--+ +--VMP--+ +--VMT+
| | | | | | | | | | | | | | | |
sxt.aj |st.vip uli.cc fkr.nk mi.vmd kn.vks m.vmp kh.sub ]sAn.aj tr.ajm xuAh.fut d.vmp C.vi d.vmt
```

Separated Complex Predicates

Accepts complex predicates separated by accusative enclitics & PPs

آنها کتکتان زدند

```
+-----Spn3p-----+
|           +-----K-----+-----VMP-----+
|           |           +---VMO---+---VMT+           |
|           |           |           |           |           |
]nhA.pn ktk[?].nk tAn.vmo z.vk d.vmt nd.vmp
```

آنها دست به کتابهای شما زدند

```
+-----Spn3p-----+
|           +-----K-----+-----+           |
|           |           +-----PP-----+           |
|           |           +-----EZ-----+           +-----VMP-----+
|           |           +---PO---+---NMSp---+           +---M---+           +-VMT+           |
|           |           |           |           |           |           |           |           |
]nhA.pn dst.nk bh.pp ktAb.n hA.nms e.ez CmA.pn z.vk d.vmt nd.vmp
```

Separated Complex Predicates 2

15/18

Accepts complex predicates separated by future auxiliary verbs

ما دست خواهیم زد

mA dst xuAhim zd

```
+-----Spn1-----+
|           +-----K-----+
|           |           +---VMPK-+---IK-+-VMP+
|           |           |           |           |           |
```


mA.pn dst.nk xuAh.fut im.vmp z.vk d.vmp

- 60 link categories including 9 morphological links
- Parses subordinate clauses, relative clauses, coordinated clauses, auxiliary verbs, unknown words, (separated) complex predicates
- ≈ 1000 sentences per minute
- Command-line & web interfaces

- Integration with Soar artificial intelligence system
- Information extraction

- Export output to constituent trees
- Expand lexicon
- Investigate theoretical implications of non-linear forms

mā dæst xāh-im zæd



- Amtrup, J. W., Megerdooonian, K., and Zajac, R. (2000a). Rapid development of translation tools: Application to Persian and Turkish. In *COLING*, pages 982–986.
- Amtrup, J. W., Rad, H. M., Megerdooonian, K., and Zajac, R. (2000b). Persian-English machine translation: An overview of the Shiraz project. Memoranda in Computer and Cognitive Science MCCS-00-319, Computing Research Lab, New Mexico State University.
- Antworth, E. (1990). *PC-KIMMO: A two-level processor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, TX.
- Koskenniemi, K. (1983). Two-level model for morphological analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 683–685.
- Megerdooonian, K. (2000). A computational analysis of the Persian noun phrase. Memoranda in Computer and Cognitive Science MCCS-00-321, Computing Research Lab, New Mexico State University.
- Sleator, D. and Temperley, D. (1993). Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*.

- Web: <http://home.byu.net/jmd56>
- Email
 - Jon Dehdari: jonsafari@byu.edu
 - Deryle Lonsdale: lonz@byu.edu