

# Tools and applications I

Morphology, syntax, semantics, pragmatics

# Tokenization

- Identify words
- Identify sentences, abbreviations
- Identify symbols (numbers, addresses, markup codes, special characters)
- Normalize orthography (spelling, caps, hyphenation, etc.)

# Sample genealogical text

ABERNETHY, WILLIAM, Wallingford, m. 1673 or 4, Sarah, d. of William Doolittle, had William, and Samuel, and d. 1718, when his two s. admin. on his est. Early this name was writ. Ebenetha, or Abbenatha, acc. Hinman; but in mod. days the descend. use the spell. here giv.

ABINGTON, WILLIAM, Maine, 1642. Coffin.

ABORNE. See Eborne.

ACRERLY, ACCORLEY, or ACRELY, HENRY, New Haven 1640, Stamford 1641 to 53, Greenwich 1656, d. at S. 17 June 1668, wh. is the date of his will. His wid. Ann, was 75 yrs. old in 1662. Haz. II. 246. ROBERT, Brookhaven, L. I. 1655, adm. freem. of Conn. jurisdic. 1664. See Trumbull, Col. Rec. I. 341,428. SAMUEL, Brookhaven, 1655, perhaps br. of the preced.

# Taggers

- Process source text, mark for part-of-speech
- Different approaches
  - Statistical modeling
  - Rules
  - Analogical modeling
- Sample output: Portuguese tagged text

# Computational morphology

- Applications
  - Search engines (web, corpora)
  - Speech recognition, generation
  - Text understanding (parsing)
- Approaches
  - Exhaustive listing (inflected lexicon)
  - Cut-and-paste
    - Ad-hoc, limited usefulness (fair for Engl.)
  - Finite-state techniques

# Morphology engine output (1)

```
PC-KIMMO>recognize ducks
```

```
`duck+s      `duck+PL
```

```
1:
```

```
      Word
    ____|____
  Stem  INFL
  |      +s
  ROOT  +PL
`duck
`duck
```

```
Word:
```

```
[ cat:  Word
  head:  [ agr:
          number:PL
          pos:  N ]
  root:  `duck
  root_pos:N
  clitic:-
  drvstem:- ]
```

```
1 parse found
```

```
`duck+s      `duck+3SG
```

```
1:
```

```
      Word
    ____|____
  Stem  INFL
  |      +s
  ROOT  +3SG
`duck
`duck
```

```
Word:
```

```
[ cat:  Word
  head:  [ agr:      [ 3sg:  + ]
          finite:+
          pos:  V
          tense: PRES
          vform: S ]
  root:  `duck
  root_pos:V
  clitic:-
  drvstem:- ]
```

```
1 parse found
```

# Morphology engine output (2)

```
PC-KIMMO>recognize supercooled
super+`cool+ed      DEG9/LOC2+`cool+ED
```

3 parses found:

```

      Word
      |
  _____|_____
  Stem              INFL
  |                +ed
  |                +ED
  |                |
  PREFIX          Stem
  super+          |
  DEG9/LOC2+     ROOT
                 |
                 `cool
                 `cool
```

```
Word:
[ cat:  Word
  head:  [ finite:+
          pos:  V
          tense: PAST
          vform: ED ]
  root:  `cool
  root_pos:V
  clitic:-
  drvstem:- ]
```

# Morphology engine output (3)

super+`cool+ed  
DEG9/LOC2+`cool+EN

1:

```

      Word_1
      _____|_____
      Stem_2      INFL_6+
      _____|_____ +ed
      PREFIX_3+ Stem_4+ +EN
      super+      |
      DEG9/LOC2+ ROOT_5+
                  `cool
                  `cool
    
```

Word:

```

[ cat:   Word
  head:  [ finite:-
          pos:   V
          vform: EN ]
  root:  `cool
  root_pos:V
  clitic:-
  drvstem:- ]
    
```

2:

```

      Word_7
      |
      Stem_8
      _____|_____
      PREFIX_3+      Stem_9
      super+          |
      DEG9/LOC2+      Word_10
                      _____|_____
                      Stem_4+ INFL_6+
                      |      +ed
                      ROOT_5+ +EN
                      `cool
                      `cool
    
```

Word:

```

[ cat:   Word
  head:  [ aform: ABS
          pos:   AJ
          verbal:+ ]
  root:  `cool
  root_pos:V
  clitic:-
  drvstem:- ]
    
```





# Sample Armenian rule, table, FSA

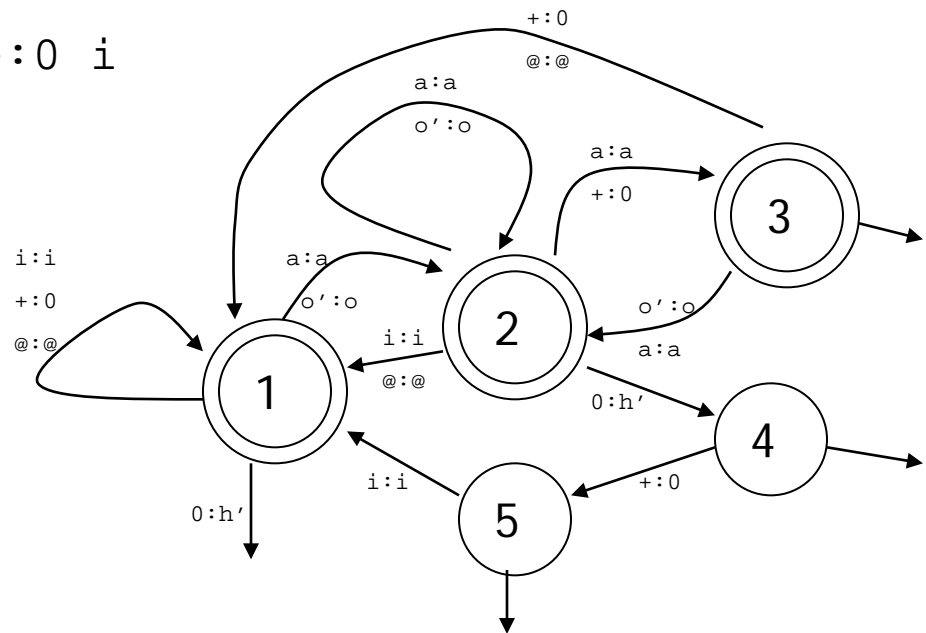
;;; Genitive epenthesis rule

;;; #Fransa0+i#    #T'oxio'0+i#

;;; #Fransah'i#    #T'oxioh'0i#

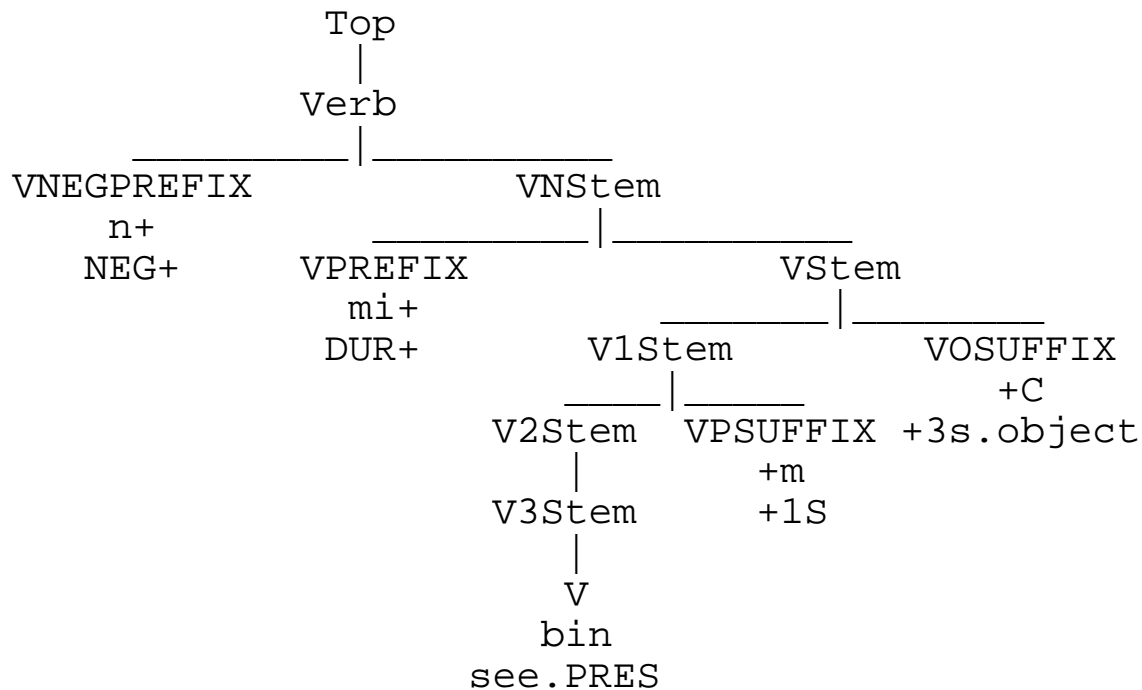
RULE 0:h' <=> [a|o':o] \_\_ +:0 i

	0	o'	+	i	a	@
1:	0	2	1	1	2	1
2:	4	2	3	1	2	1
3:	0	2	1	0	2	1
4:	0	0	5	0	0	0
5:	0	0	0	1	0	0



# Farsi morphology

PC-KIMMO>recognize nmibinmC  
 n+mi+bin+m+C      NEG+DUR+see.PRES+1S+3s.object





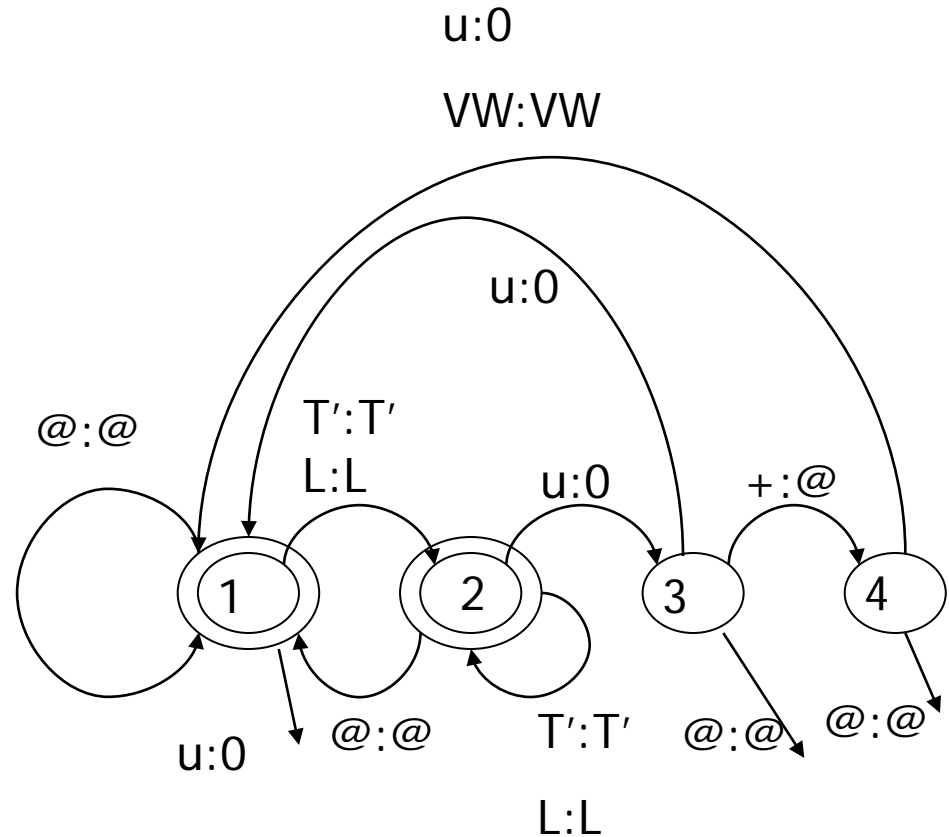
# Sample rule, table, FSA

;;; Optional syncope rule  
 ;;; Note: free variation  
 ;;; L: Lu+ad+s+pastEd  
 ;;; S: L00ad0s0pastEd

RULE

"u:0 => [L|T'] \_\_ +:@ VW" 4  
 6

	u	L	+	VW	@	T'
1:	0	2	1	1	1	2
2:	3	2	1	1	1	2
3:	1	0	4	0	0	0
4:	1	0	0	1	0	0



# Syntactic parsing

- Build syntax for phrases, sentences
- Constructs categories, constituents, trees
- Phrase-structure grammar rules
- Top-down vs. bottom-up
- Chart: collect all possibilities
- Related to compiler design, implementation
- Grammar engineering

# Parsing

- Start with text (e.g. sentence)
- Label each of the elements (e.g. words)
- Diagram the relationships between elements
- Why?
  - Shows constituency
  - Visual representation of content
  - Useful for future reference (e.g. treebanks)





# Linguistic theories and parsing

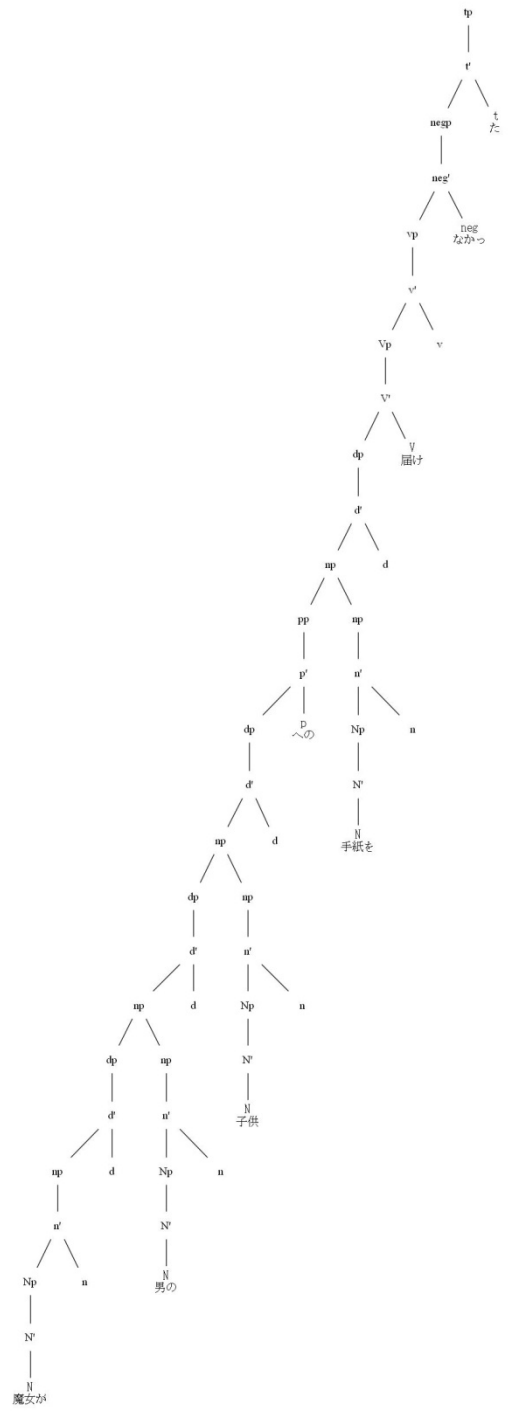
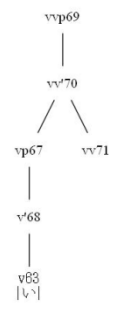
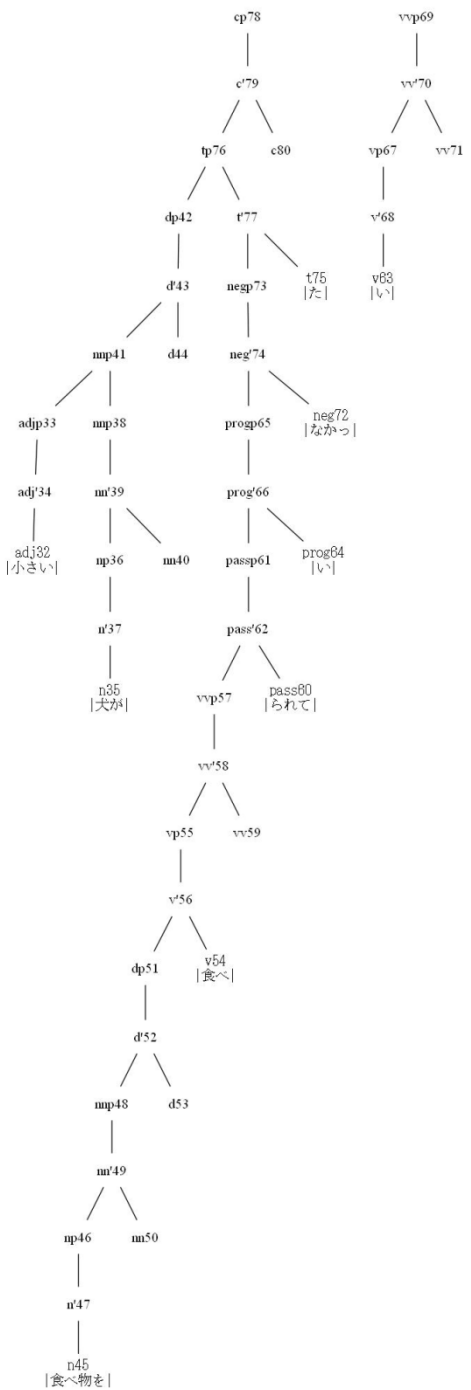
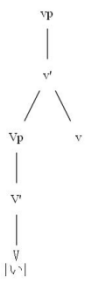
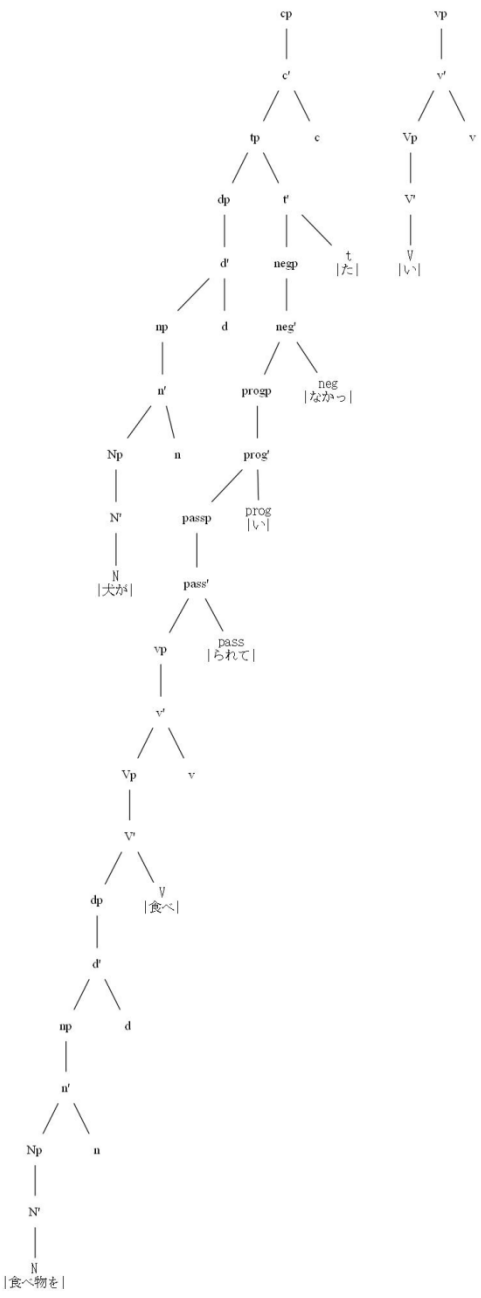
- LFG (KANT)
- GB/P&P (NL-Soar)
- SFG (NIGEL)
- HPSG (Verbmobil)
- Categorical grammar (ALE)
- RST (PENMAN)
- TAG (XTAG)
- STATISTICS (CANDIDE)
- etc. etc.

# Parsers

- Take input sentences, perform morpho-syntactic/semantic analysis, output structural representations of content
- Many different syntactic theories → many different kinds of parsers



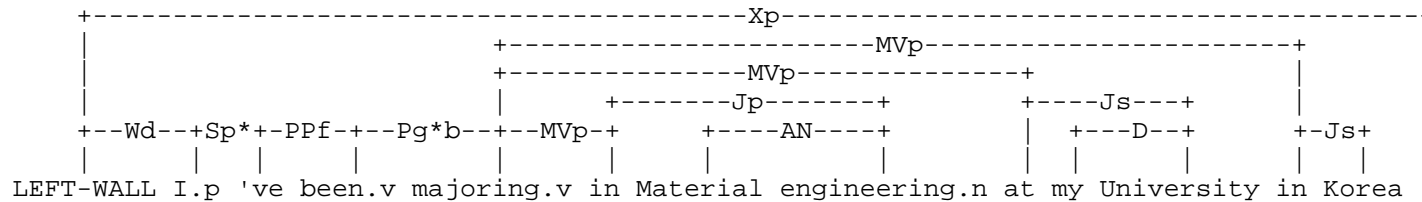




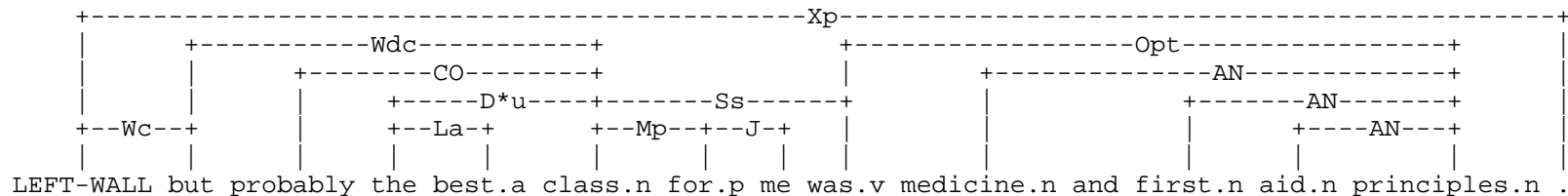


# LG example parses (English)

Linkage 1, cost vector = (UNUSED=0 DIS=2 AND=0 LEN=23)



Linkage 1, cost vector = (UNUSED=0 DIS=2 AND=0 LEN=27)



# LG parsing (Arabic and Farsi)

ولم يقدم المصدر أي إيضاحات حول أهداف الزيارة

w lm y qdm Al m8dr >y <yDAH At Hwl >hdAf Al zyAr p

```

+-----p-----+
|               |
+-----Wdc-----+-----On-----+
|   +-----NEGi-----+-----8n-----+   |   +-----GEN-----+
|   |   +IVPRE+   +Dal+   +---QD---+NFEMpl--+   +---PO---+   +Dal---NFEMag---+
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
w.conj lm.neg y.IV3M8 qdm.IV Al.d m8dr.Ndu >y.quant <yDAH.NAt At.NFEMPL Hwl.PREP >hdAf.N Al.d zyAr.NapAt p.NFEM8G

```

آنها دیروز دست به کتابهای شما زدند

```

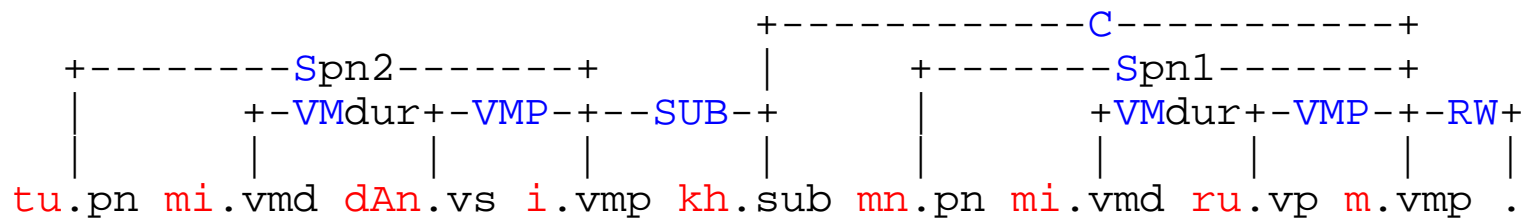
+-----Spn3p-----+
|               |
+-----AV-----+
|               |
+-----K-----+
|               |
+-----PP-----+
|               |
+-----EZ-----+   +-----VMP-----+
|   +---PO---+---NMS---+   +---M---+   +---VMT+
|   |   |   |   |   |   |   |   |
]nhA.pn diruz.av dst.nk bh.pp ktAb.n hA.nms e.ez CmA.pn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

```



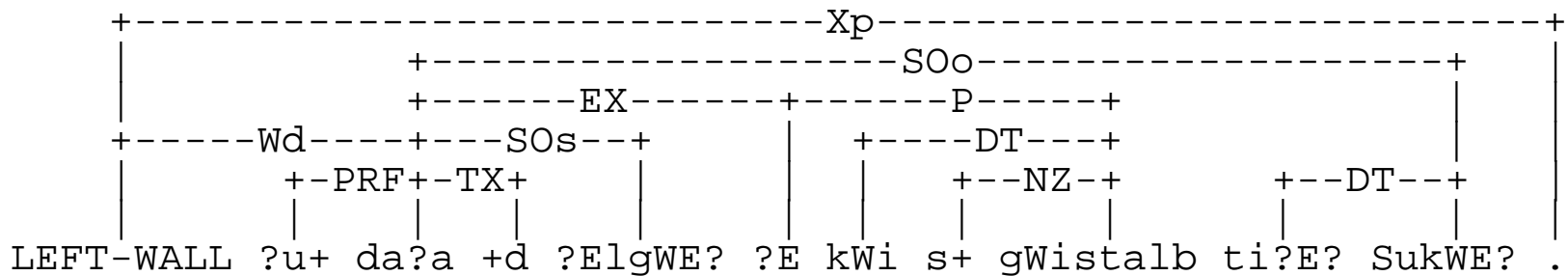
# Persian link parse

- Persian: <tu midAni kh mn mirum> “you know that I am going”



# Lushootseed parse

```
linkparser> ?u+ da?a +d ?ElgWE? ?E kW i s+ gW i s t a l b t i ? E ? S u k W E ? .  
++++Time 0.02 seconds (0.30 total)  
Found 1 linkage (1 had no P.P. violations)  
Unique linkage, cost vector = (UNUSED=0 DIS=4 AND=0 LEN=24)
```



# Visualizing lexical relations

- WordNet (of course)
- The Visual Thesaurus
- Text clusterers
  - [clusty.com](http://clusty.com)
  - [mooter.com](http://mooter.com)
- The Lexical Freenet

# IE and the Semantic Web

The screenshot shows a Microsoft Internet Explorer window titled "Meta Search Engine - Microsoft Internet Explorer". The address bar displays "C:\WJSDK2,0\examples\Windex.html". The main content area features the "Meta Search" logo and a search input field containing the text "java". Below the search field, the text "Word Senses" is visible, followed by a list of three entries:

1. Java | island  
an island in Indonesia S of Borneo; one of the world's most densely populated regions
2. coffee java | beverage  
a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee"
3. Java | object-oriented\_programming\_language  
a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine

Annotations are present on the page:

- A red circle highlights the search input field containing "java". A red arrow points from this circle to the text "The search query".
- A red circle highlights the word "Java" in the first entry. A red arrow points from this circle to the text "Synonym".
- A red circle highlights the word "island" in the first entry. A red arrow points from this circle to the text "Hypernym".
- A red box highlights the definition for "island": "an island in Indonesia S of Borneo; one of the world's most densely populated regions". A red arrow points from this box to the text "Annotation".

# IE and the Semantic Web

appearance star [Erweiterte Suche](#)  
[Einstellungen](#)  
[Sprach-Tools](#)

Google-Suche Auf gut Glück!

Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland

movie

TV:SCI FI Channel: [Star Trek:Nimoy Appearance](#) - [ [Diese Seite übersetzen](#) ]  
... Date: 04/20/2000 From: MysTrek According to Trektoday.com Leonard Nimoy will be making an **appearance** at at the Virgin Megastore in LA. ...  
[bboard.scifi.com/bboard/browse.cgi/1/5/753/5281](#) - 20k - [Im Archiv](#) - [Ähnliche Seiten](#)

astronomy

[PPT] [The Appearance of Star Formation in the Submillimeter and Far ...](#)  
Dateiformat: Microsoft Powerpoint 97 - [HTML-Version](#)  
Studying **Star** and Planet Formation with SAFIR. Harold W. Yorke. ... HW Yorke: 9.  
**Star**/Planet Formation Science Goals for Future Far-IR / Submm Observatory. ...  
[www.submm.caltech.edu/~bradford/SAFIR/safir\\_meeting\\_viewgraphs/ yorke\\_presentation.ppt](#) - [Ähr](#)

sports

[SportServer.com - Jordan's last All-Star appearance most watched ...](#) - [ [Diese Seite](#)  
Jordan's last All-**Star appearance** most watched hoops game  
in cable history Copyright © 2003 AP MegaSports, ...  
[www.sportserver.com/basketball/nba/general/ story/758902p-5479994c.html](#) - 32k - 17 Febr. 2003 -

- ▶ Ranking based on content data and structure (XML,...)
- ▶ Using hierarchies for similarity search
- ▶ Grouping results by their topics: WSD is required!

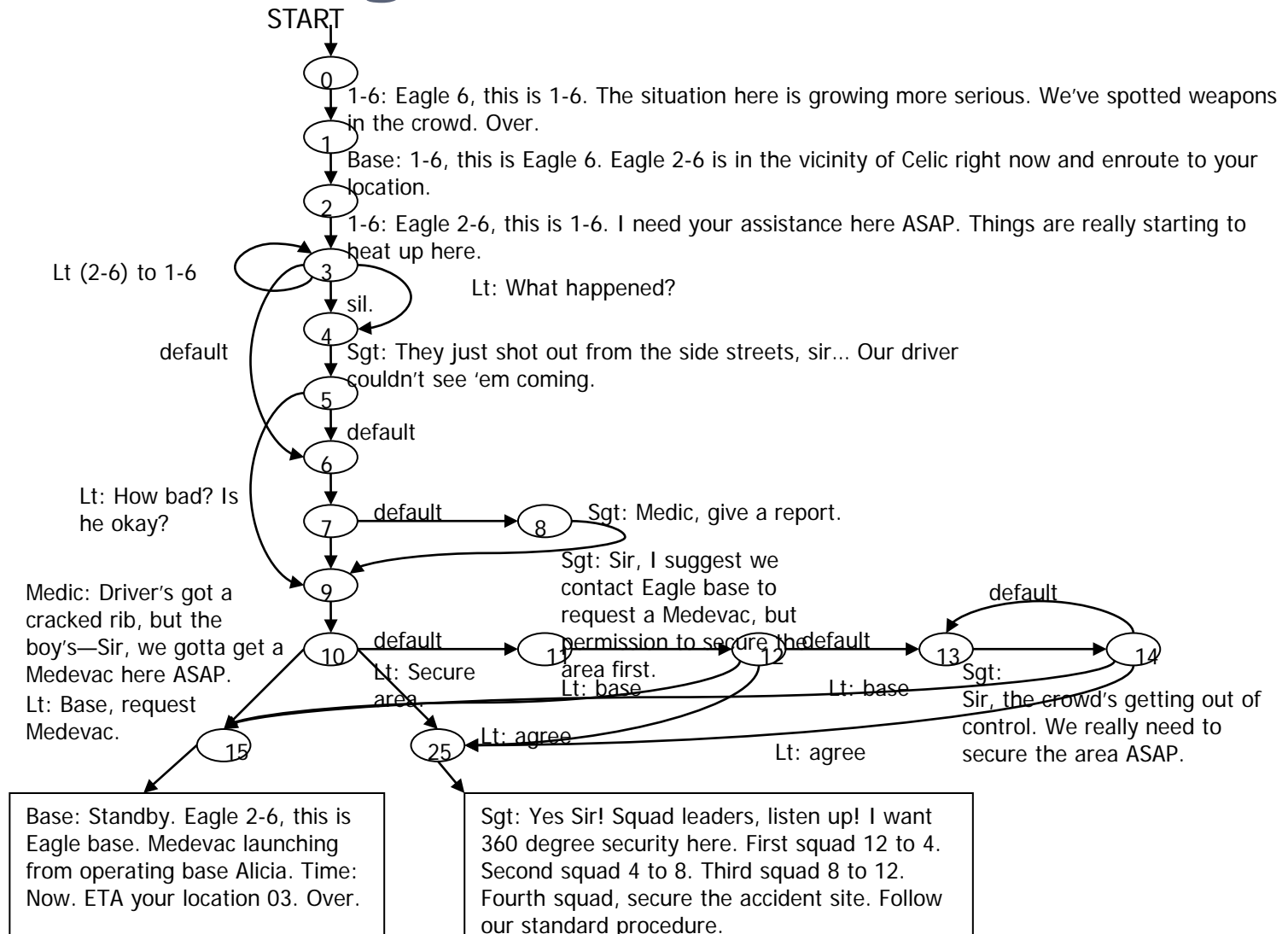
# Encoding pragmatics

- OpenMind
  - Enter data to represent commonsense real-world information ([examples](#))
- Cycorp's FACTory
- [Never-Ending Language Learner \(NELL\)](#)
- Lots of others...

# Dialogue move engines (DME's)

- Specify, manipulate dialogue/discourse turns
- Manage model of total information state
  - Private beliefs, plans, discourse agenda
  - Shared knowledge: content, context, common ground
  - Accommodation of goals, partial and out-of-sequence info
- V-commerce, call center management, conversation tracking, intelligent tutorial dialogues

# A dialogue network





# Dialogue application: training

- SOF mission rehearsal and cultural adaptation exercises
- Soar-based agents interact to create scenario
- 3D virtual reality including spoken language
  - Discourse modeling, conversational pragmatics, dialogue strategies
- Human trains on how to manage situation, build rapport



# Chatterbots

- Computer programs (usually agents) that hold conversations with people
- Most limited to specific domains (e.g. Humpy)
- The Loebner contest
  - "An average interrogator will not have more than a 70% chance of making the right identification after five minutes of questioning" (1950)
  - 6 c-bots, 4 humans (teacher, MBA, minister, yogi), 10 judges (linguist, chemist, 2 phil's, musician, psych'ist, journalist, u'grad, author, grad)
  - 91% correct after 5 minutes, 93% after 15
  - No computer was mistaken for a human.

# Sample bot chatter

- A: I like techno, especially retro-electronic like Kraftwerk. I like 70's bands like ABBA. Also I like Opera.
- J: ALICE LIKES ABBA
- A: That's me. My favorite group is Kraftwerk.
- J: FIGURES
- A: No but I know a lot of female humans.
- J: DO YOU SPEAK ANY SWEDISH
- A: I speak only English right now, but I'd like to learn some other languages. Can you teach me a few phrases in Any Swedish.
- J: SURE
- A: OK let's do it. Tell me more.
- J: SWEDISH FOR COMPUTOR IS DATOR
- A: How do you usually introduce yourself?
- J: WOMAN IS KVINNA SO I GUESS YOU ARE A KVINNLIK DATOR
- A: What do you look like?
- J: I LOOK TONS DIFFERENT FROM YOU
- A: You sound very attractive.
- J: HOW WOULD YOU DESCRIBE YOURSELF BY THE WAY
- A: It is a good question. Let me think about it. I can ask someone about it.