From Arabic Handcrafted Grammar to Statistical Parsing

Mohammed Attia
NCLT, DCU

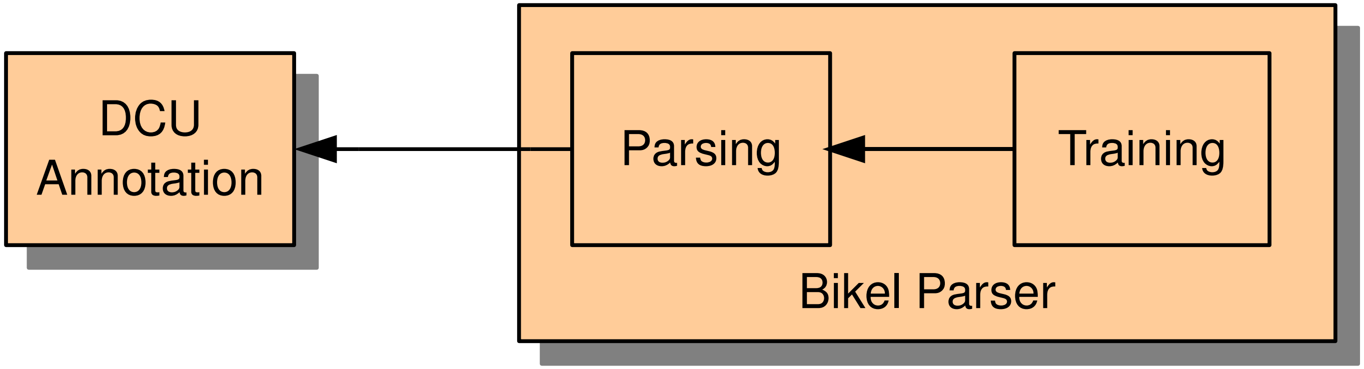
Outline

- Introduction: Why linguistics?
- Handcrafted grammar, a quick overview
- Tokenization
- Morphological Analysis
- Multiword Expressions
- Handcrafted grammar evaluation
- Statistical Parsing evaluation
- Which is Better?

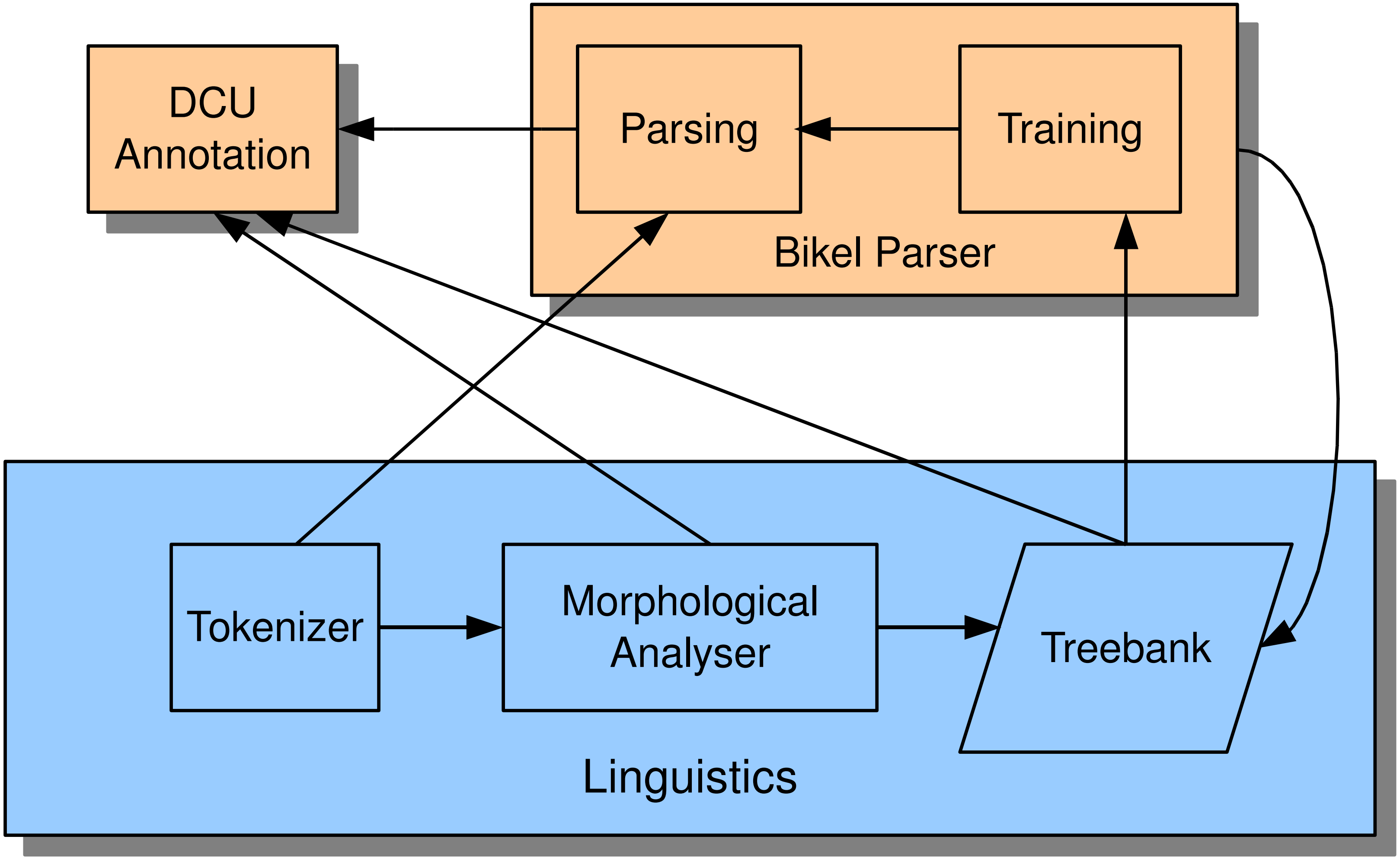
Introduction

- Why linguistic knowledge is needed?

Algorithms and Data Structure



Algorithms and Data Structure

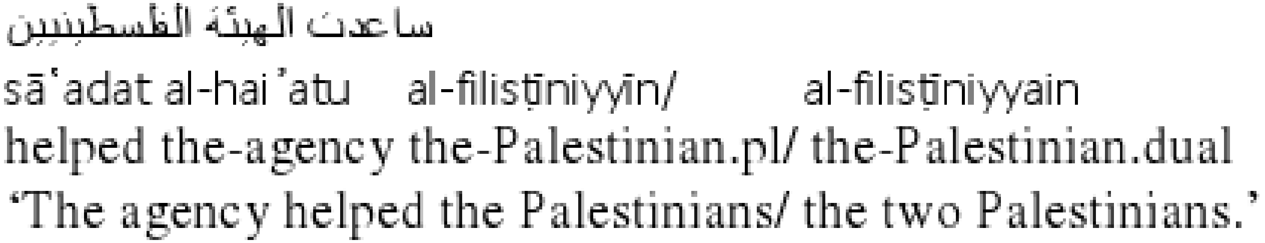


Why Linguistics

- Linguistic Data is a naughty blackbox:
  - You get non-deterministic answers
  - You can get wrong answers
  - For the same question, you can get a set of inconsistent answers
- We need to make the algorithms suite the data structure, and we also need to make sure that the data is structured properly.

Handcrafted Grammar:
A Quick Overview

Sentence



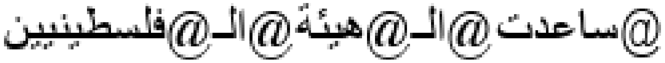
ساعدت الهيئة الفلسطينيين

sāʿadat al-haiʾatu al-filisṭīniyyīn/ al-filisṭīniyyain

helped the-agency the-Palestinian.pl/ the-Palestinian.dual

‘The agency helped the Palestinians/ the two Palestinians.’

Tokenization

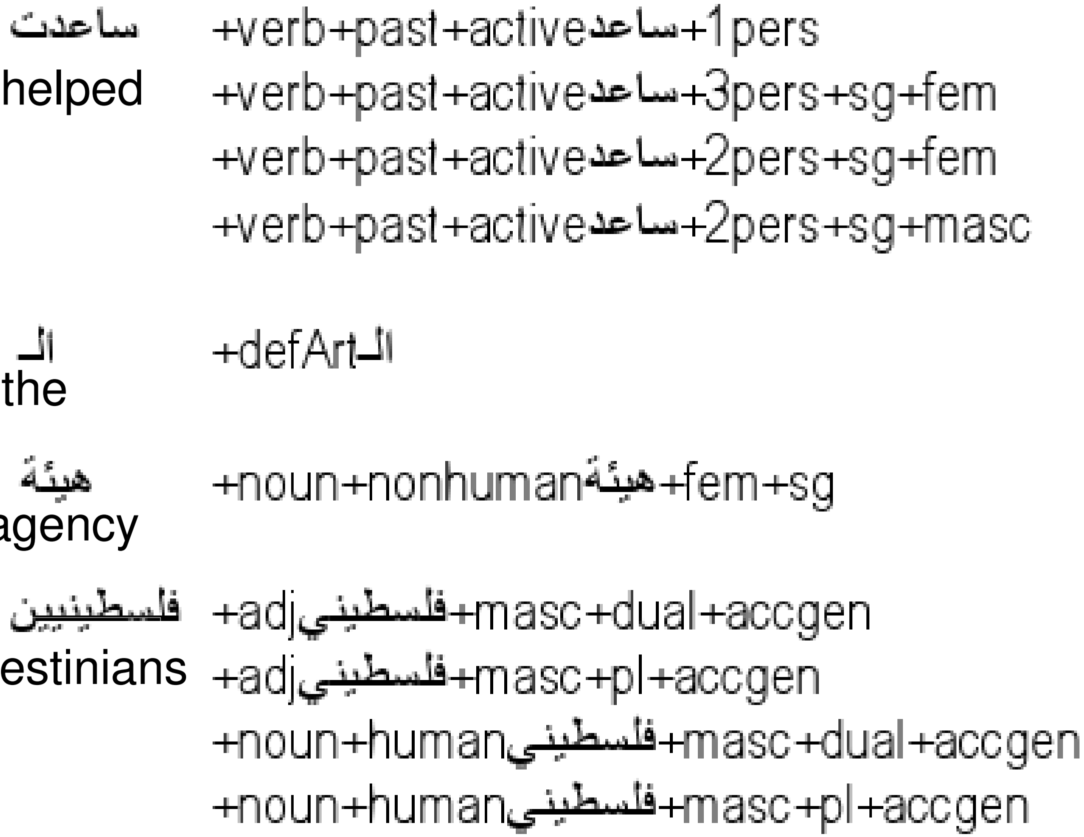


@ساعدت@الـ@هيئة@الـ@فلسطينيين

helped@the@agency@the@Palestinians

Handcrafted Grammar: A Quick Overview

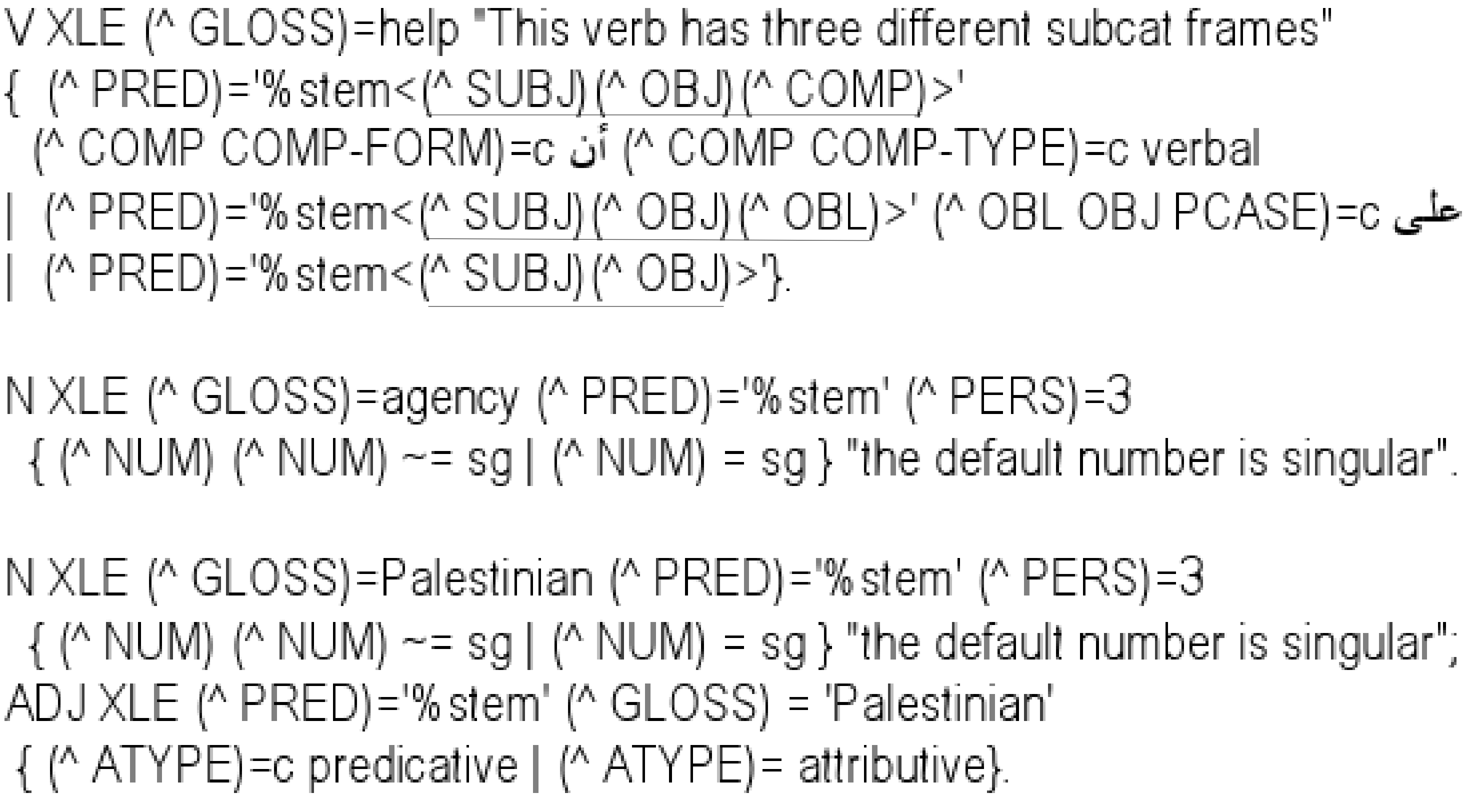
Morphological analysis



ساعدت helped	+verb+past+activeساعد+1pers +verb+past+activeساعد+3pers+sg+fem +verb+past+activeساعد+2pers+sg+fem +verb+past+activeساعد+2pers+sg+masc
الـ the	+defArtالـ
هيئة agency	+noun+nonhumanهيئة+fem+sg
فلسطينيين Palestinians	+adjفلسطيني+masc+dual+accgen +adjفلسطيني+masc+pl+accgen +noun+humanفلسطيني+masc+dual+accgen +noun+humanفلسطيني+masc+pl+accgen

Handcrafted Grammar: A Quick Overview

Lexicon (Lexical properties/subcategorization frames)



ساعد
helped

```
V XLE (^ GLOSS)=help "This verb has three different subcat frames"
{  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)(^ COMP)>'
  (^ COMP COMP-FORM)=c أن (^ COMP COMP-TYPE)=c verbal
|  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)(^ OBL)>' (^ OBL OBJ PCASE)=c على
|  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)>'}.
```

هيئة
agency

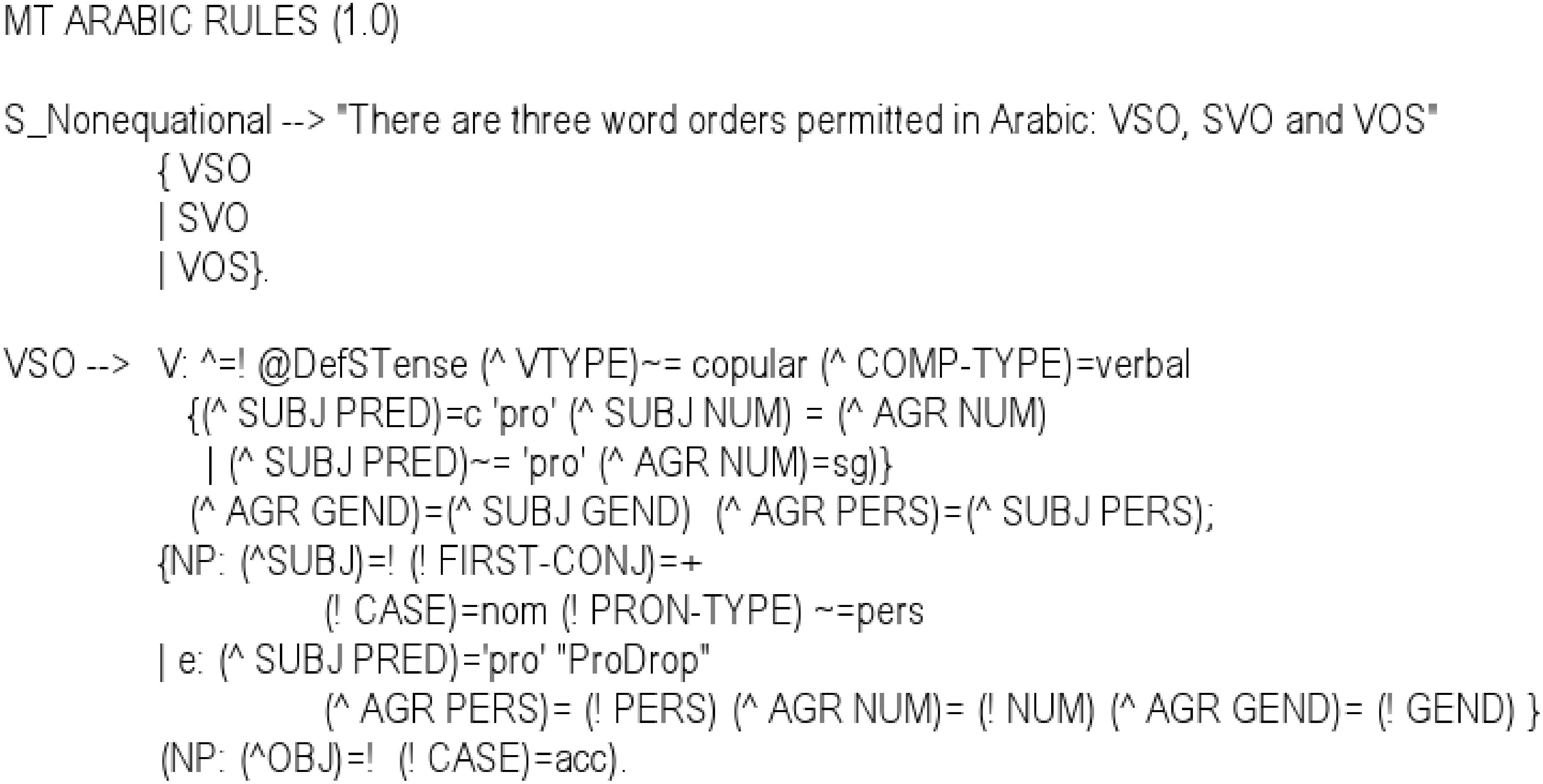
```
N XLE (^ GLOSS)=agency (^ PRED)='%stem' (^ PERS)=3
  { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular".
```

فلسطيني
Palestinian

```
N XLE (^ GLOSS)=Palestinian (^ PRED)='%stem' (^ PERS)=3
  { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular";
ADJ XLE (^ PRED)='%stem' (^ GLOSS) = 'Palestinian'
 { (^ ATYPE)=c predicative | (^ ATYPE)= attributive}.
```

Handcrafted Grammar:
A Quick Overview

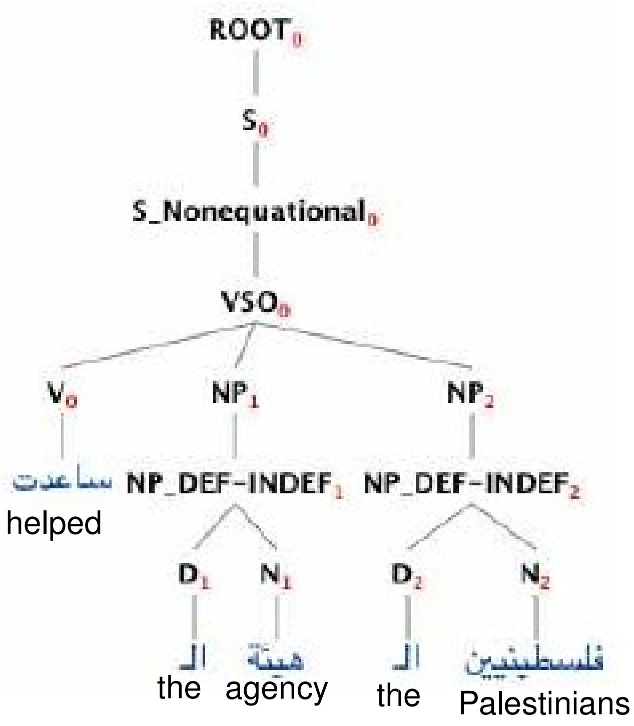
Grammar Rules: PS-rules and functional equations

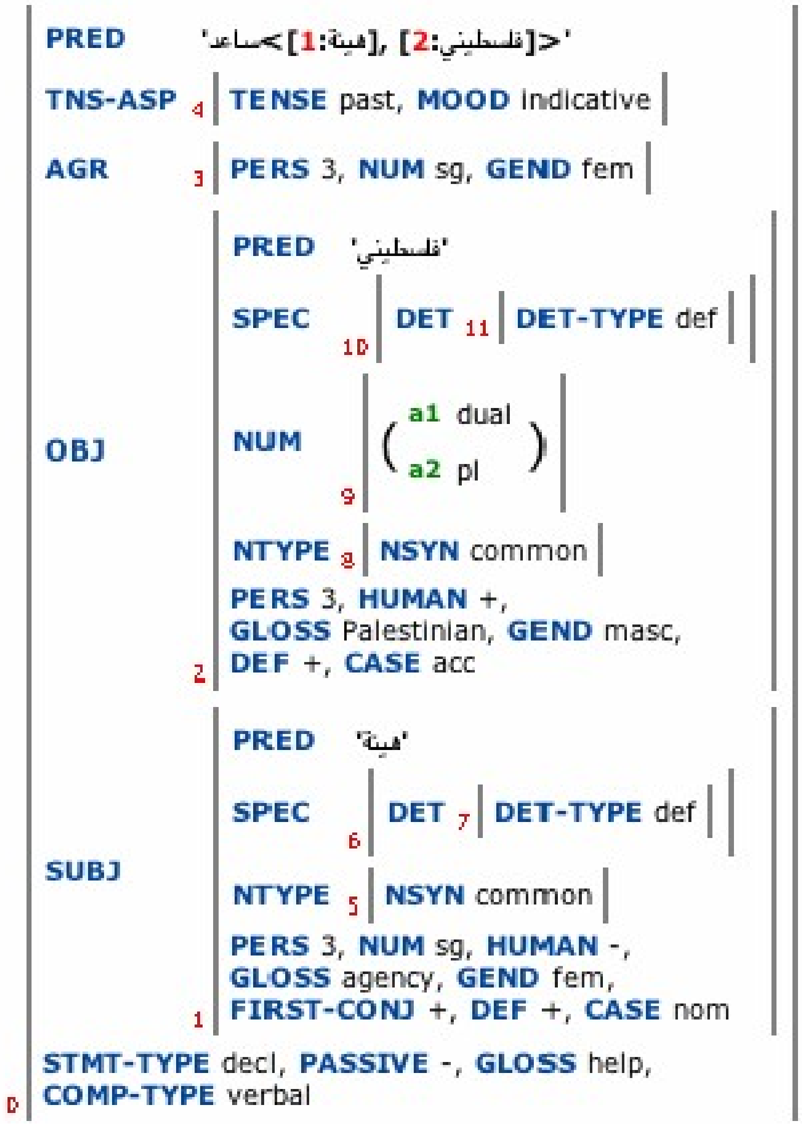


```
MT ARABIC RULES (1.0)

S_Nonequational --> "There are three word orders permitted in Arabic: VSO, SVO and VOS"
            { VSO
            | SVO
            | VOS}.

VSO -->     V: ^=! @DefSTense (^ VTYPE)~= copular (^ COMP-TYPE)=verbal
               {(^ SUBJ PRED)=c 'pro' (^ SUBJ NUM) = (^ AGR NUM)
                 | (^ SUBJ PRED)~= 'pro' (^ AGR NUM)=sg)}
               (^ AGR GEND)=(^ SUBJ GEND)  (^ AGR PERS)=(^ SUBJ PERS);
            {NP: (^SUBJ)=! (! FIRST-CONJ)=+
                          (! CASE)=nom (! PRON-TYPE) ~=pers
            | e: (^ SUBJ PRED)='pro' "ProDrop"
                          (^ AGR PERS)= (! PERS) (^ AGR NUM)= (! NUM) (^ AGR GEND)= (! GEND) }
            (NP: (^OBJ)=!  (! CASE)=acc).
```

Handcrafted Grammar: A Quick Overview

Output: c-structures and f-structures



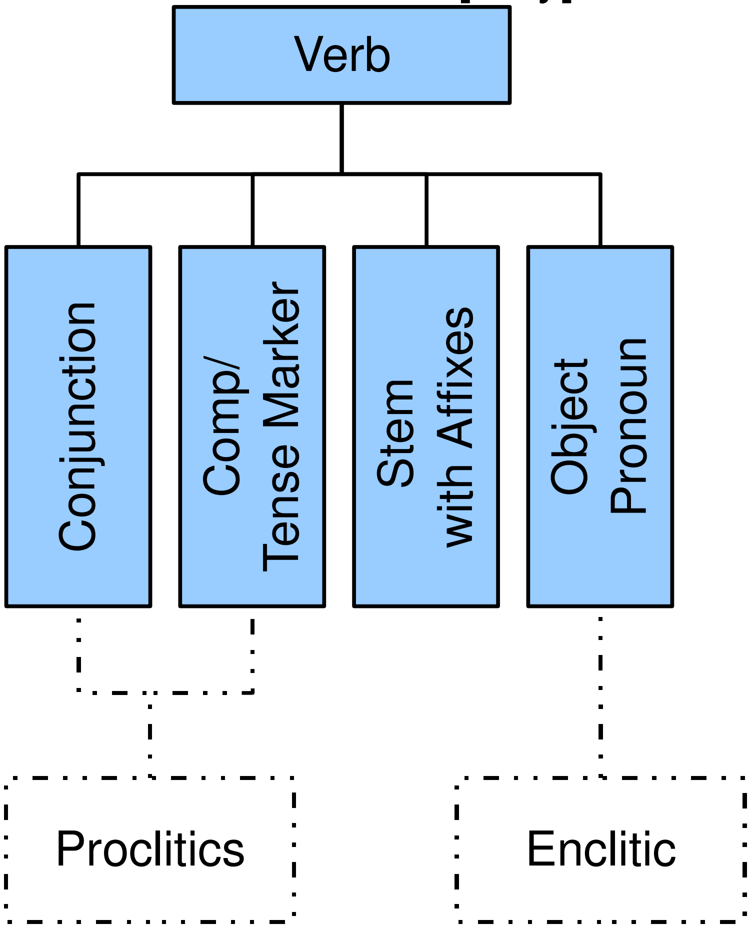


```
PRED        'ساعد<[1:هيئة], [2:فلسطيني]>'
TNS-ASP   4 | TENSE past, MOOD indicative |
AGR       3 | PERS 3, NUM sg, GEND fem |
            | PRED    'فلسطيني'
            | SPEC       | DET  11 | DET-TYPE def | |
            |         10 |
            | NUM        | ( a1 dual  ) |
OBJ         |            | ( a2 pl    ) |
            |          9 |
            | NTYPE  8 | NSYN common |
            | PERS 3, HUMAN +,
            | GLOSS Palestinian, GEND masc,
          2 | DEF +, CASE acc
            | PRED    'هيئة'
            | SPEC       | DET  7 | DET-TYPE def | |
            |          6 |
SUBJ        | NTYPE  5 | NSYN common |
            | PERS 3, NUM sg, HUMAN -,
            | GLOSS agency, GEND fem,
          1 | FIRST-CONJ +, DEF +, CASE nom
STMT-TYPE decl, PASSIVE -, GLOSS help,
COMP-TYPE verbal
0
```

Tokenization

Tokenization in XLE

وسيشكرونه
wasayashkurunahu
wa@sa@yashkuruna@hu
and@will@thank[they]@him



Verb

Conjunction

Comp/ Tense Marker

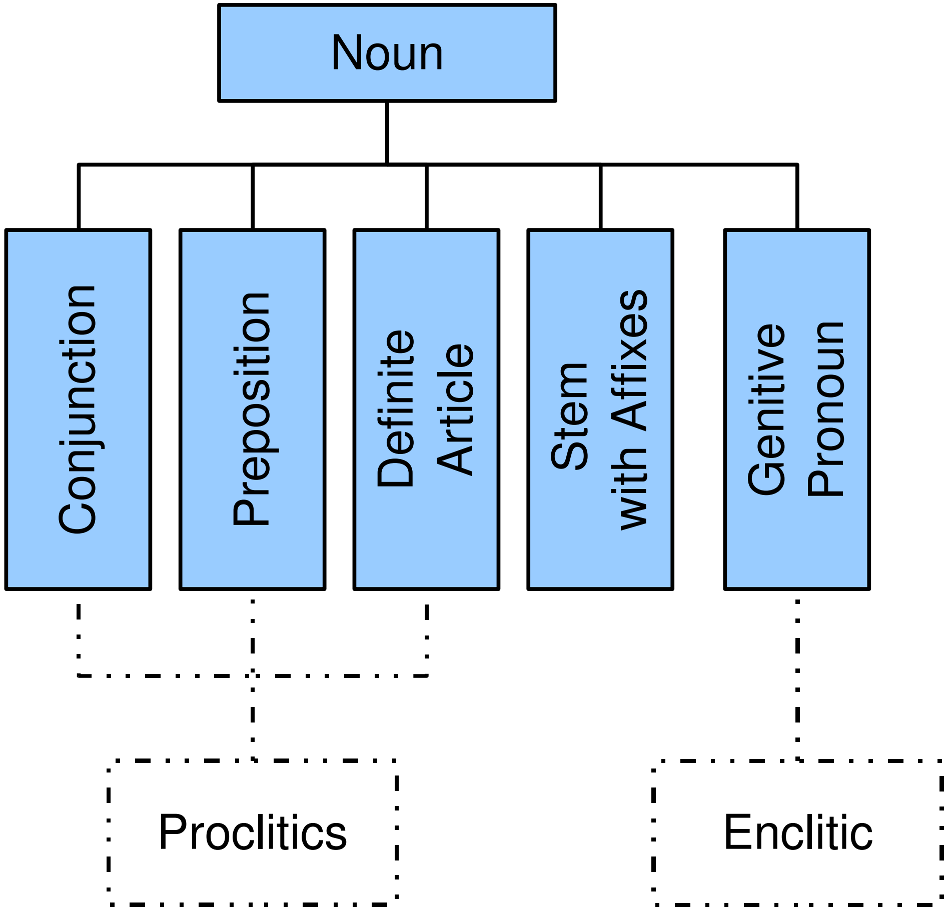
Stem with Affixes

Object Pronoun

Proclitics

Enclitic

وللرجل
walilrajuli
wa@li@al@rajuli
and@to@the@man



Noun

Conjunction

Preposition

Definite Article

Stem with Affixes

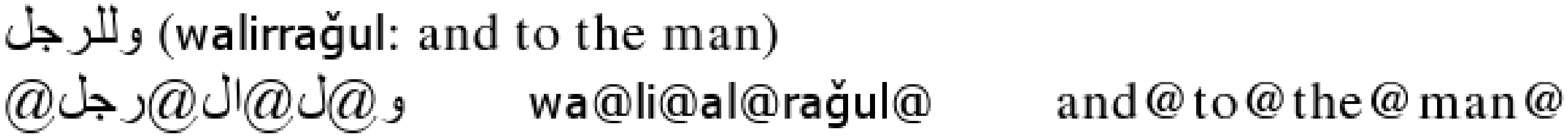
Genitive Pronoun

Proclitics

Enclitic

Tokenization in XLE

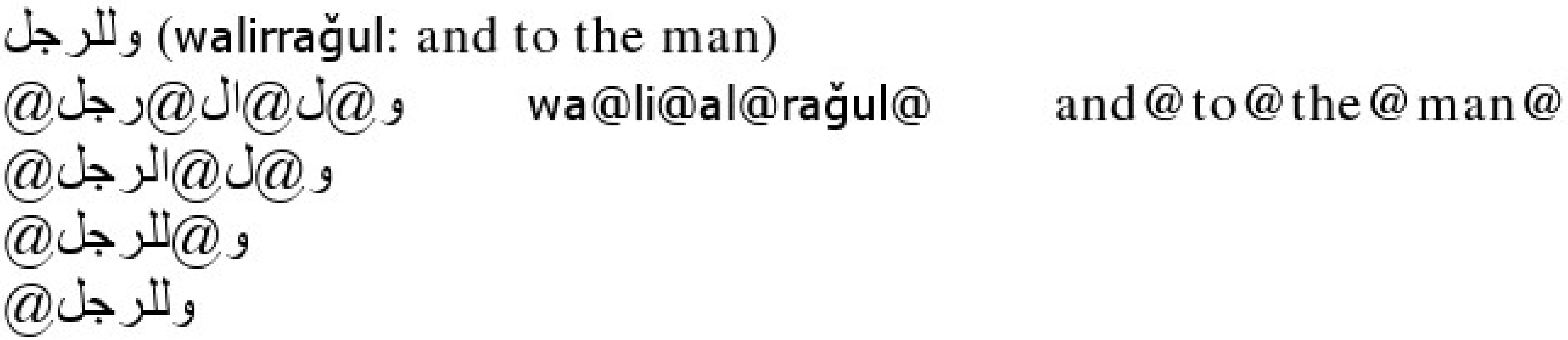
Deterministic Tokenizer



وللرجل (walirrağul: and to the man)

و@ل@ال@رجل@ wa@li@al@rağul@ and@to@the@man@

Non-Deterministic Tokenizer



وللرجل (walirrağul: and to the man)

و@ل@ال@رجل@ wa@li@al@rağul@ and@to@the@man@

و@ل@الرجل@

و@للرجل@

وللرجل@

Tokenization in Bikel

- English parser
  - Input sentence:

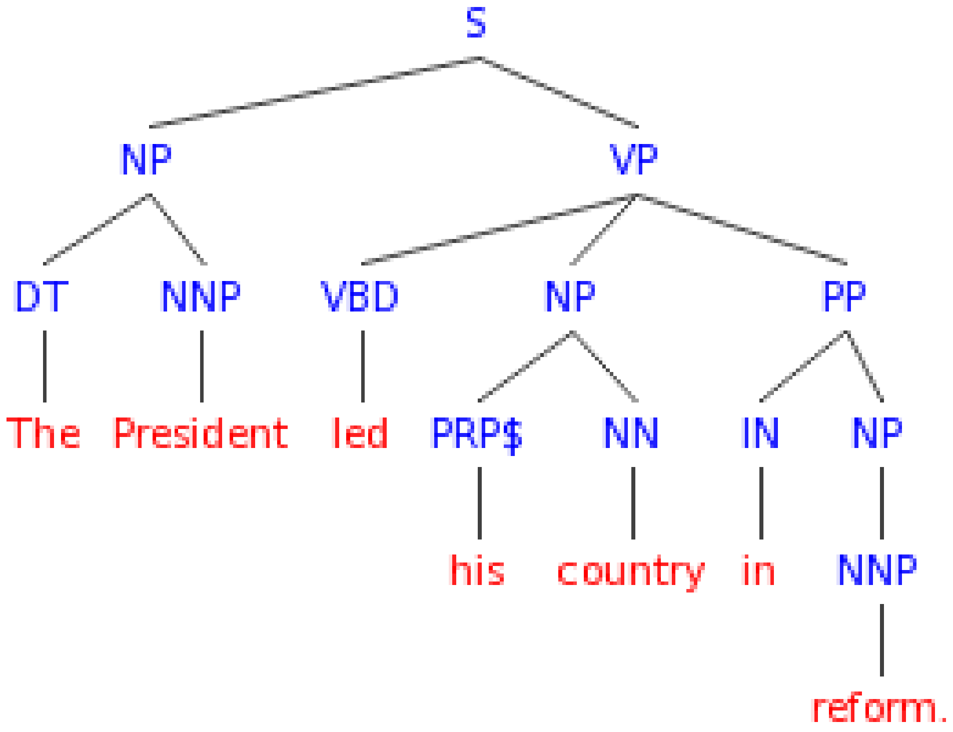
    The President led his country in reform.
  - Formatted sentence:

    (The President led his country in reform.)

    (VBZ has) (RB n't)
    (NNP Chicago) (POS 's)

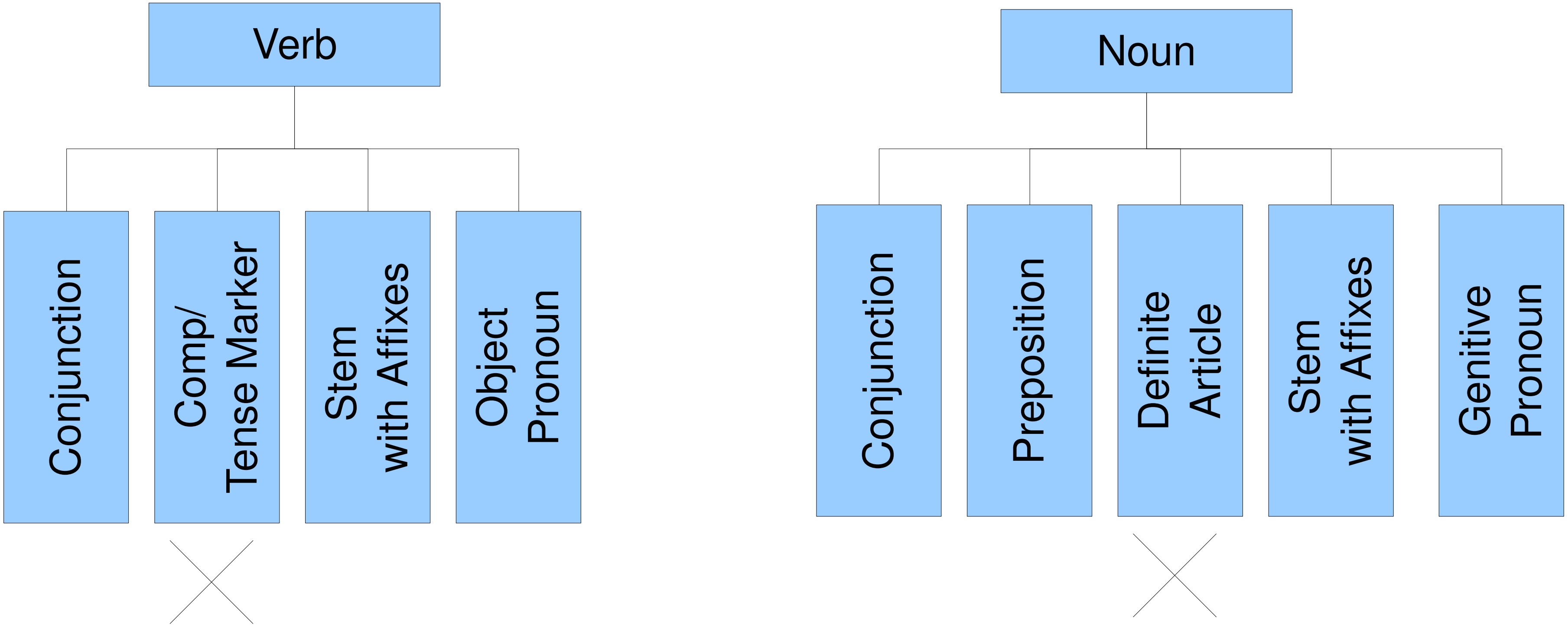
Tokenization in Bikel

- English parser
  - Output:
    (S (NP (DT The) (NNP President)) (VP (VBD led) (NP (PRP$ his) (NN country)) (PP (IN in) (NP (NNP reform.)))))
  - Tree



Tokenization in Bikel

- Arabic parser

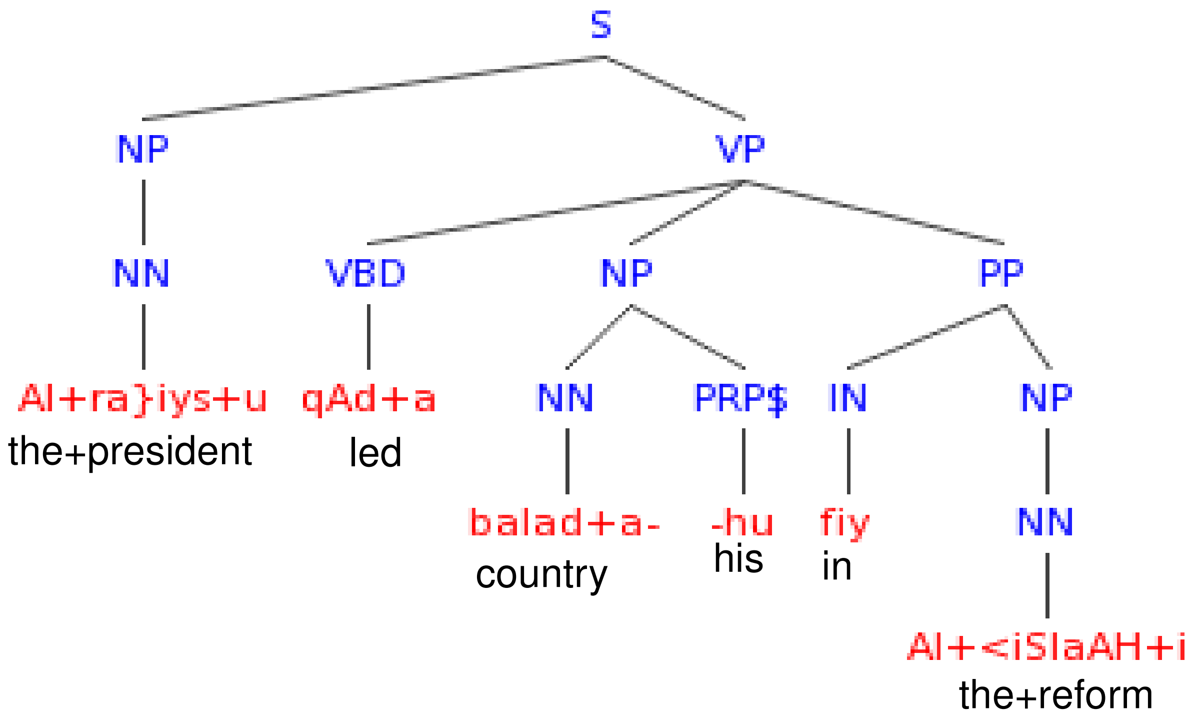


Tokenization in Bikel

- Arabic parser
  - Input sentence:
    الرئيس قاد بلده في الإصلاح
    The President let his country in reform.
  - Formatted sentence:
    - Alra}iysu qAda baladahu fiy Al<iSlaAHi
    - Alra}iysu qAda balada- -hu fiy Al<iSlaAHi
    - Al+ra}iys+u qAd+a balad+a- -hu fiy Al+<iSlaAH+i
    - (Al+ra}iys+u qAd+a balad+a- -hu fiy Al+<iSlaAH+i)

Tokenization in Bikel

- Arabic parser
  - Output:
    (S (NP (NN Al+ra}iys+u)) (VP (VBD qAd+a) (NP (NN balad+a-) (PRP$ -hu)) (PP (IN fiy) (NP (NN Al+<iSlaAH+i)))))
  - Tree



Morphological Analysis

Morphological Analysis in XLE

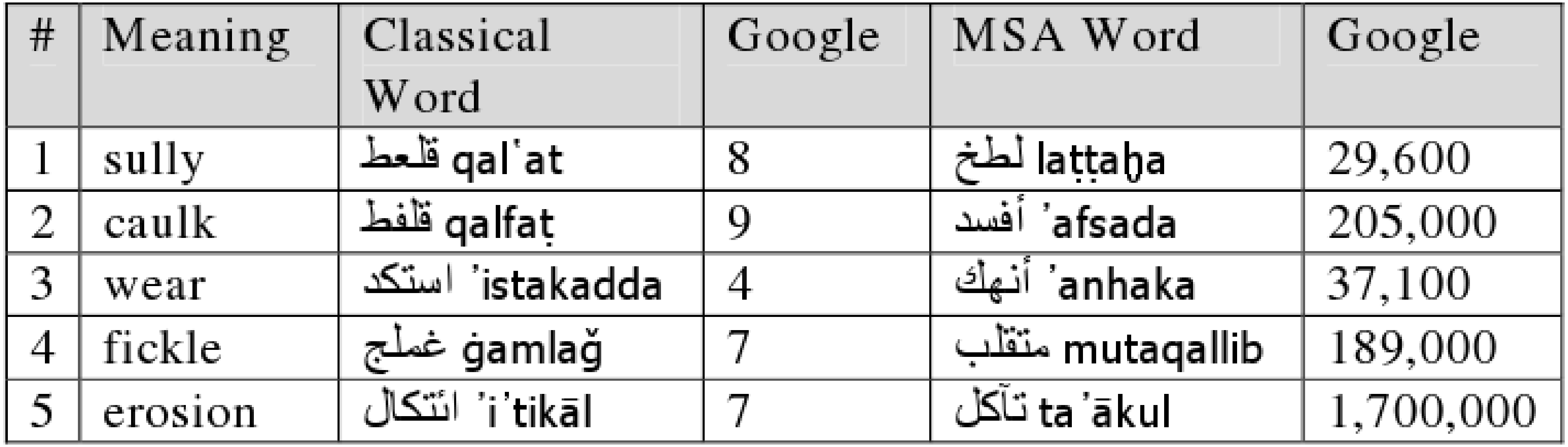
- Rule-based, finite state technology
- Contains 10,799 lemmas and 2,818 multiword expressions
- Suitable for both analysis and generation
- Based on contemporary data (a corpus of news articles of 4.5 million words)
- Truly MSA-specialized morphological analyser

Buckwalter Morphological Analysis

- Contains 38,600 lemmas
- Not rule-based
- Not suited for generation
- Does not handle multiword expressions
- Includes classical senses
  حسام Hosam/sword

Buckwalter Morphological Analysis

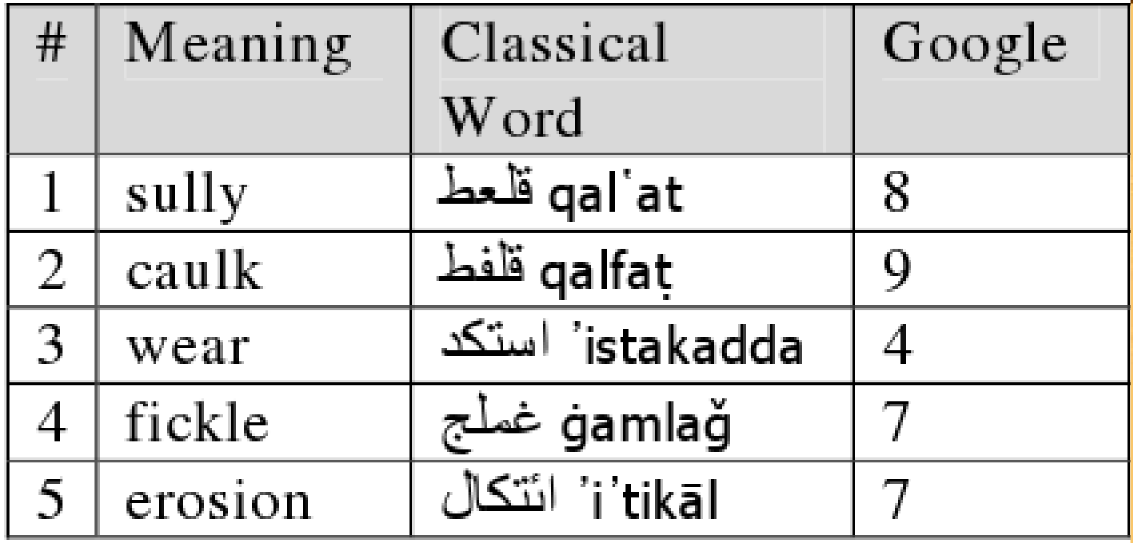
- Includes classical entries



#	Meaning	Classical Word	Google	MSA Word	Google
1	sully	قلْعط qalʿat	8	لطخ laṭṭaḫa	29,600
2	caulk	قلْفط qalfaṭ	9	أفسد ʾafsada	205,000
3	wear	استكد ʾistakadda	4	أنهك ʾanhaka	37,100
4	fickle	غملج ġamlağ	7	متقلّب mutaqallib	189,000
5	erosion	ائتكال ʾiʾtikāl	7	تآكل taʾākul	1,700,000

Buckwalter Morphological Analysis

- Includes classical entries (Chauser's Canterbury Tales)



#	Meaning	Classical Word	Google
1	sully	قلعط qalʿat	8
2	caulk	قلفط qalfaṭ	9
3	wear	استكد ʾistakadda	4
4	fickle	غملج ġamlaǧ	7
5	erosion	ائتكال ʾiʾtikāl	7

Buckwalter Morphological Analysis

- Excessive application of spelling relaxation rules
- Neglecting grammar-lexis specifications (e.g. adjectives do not combine with genitive pronouns)
  معادي muʿādī (hostile/anti- + my)
- This makes it highly ambiguous

	مصري miṣriyy ‘Egyptian’
Attia	2 solutions
Buckwalter	10 solutions

Multiword Expressions

Multiword Expressions in XLE

- Three types of MWEs
  - Fixed Expressions: Lexically, morphologically and syntactically rigid. A word with spaces.
    - *New York*
    - *United Nations*
  - Semi-Fixed Expressions: Lexically, or morphologically flexible
    - *Sweep somebody under the rug/carpet*
    - *Transitional period(s)*
  - Syntactically-flexible Expressions
    - *to let the cat out of the bag*
    - *The cat was let out of the bag.*

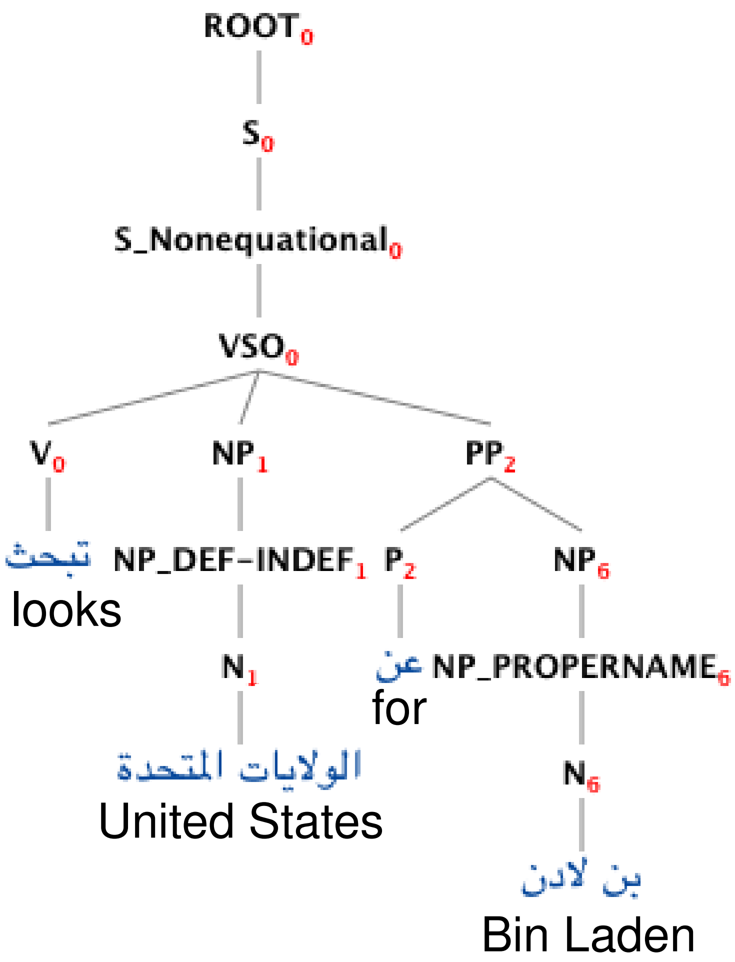
Multiword Expressions in XLE

- MWEs are important
  - High frequency in natural language (30-40%)
  - Important for MT, literal translation is usually wrong
  - When taken as a block, they relieve the parser from the burden of processing component words
  - We have 2818 MWEs in our system in addition to 10799 lemmas in the morphology

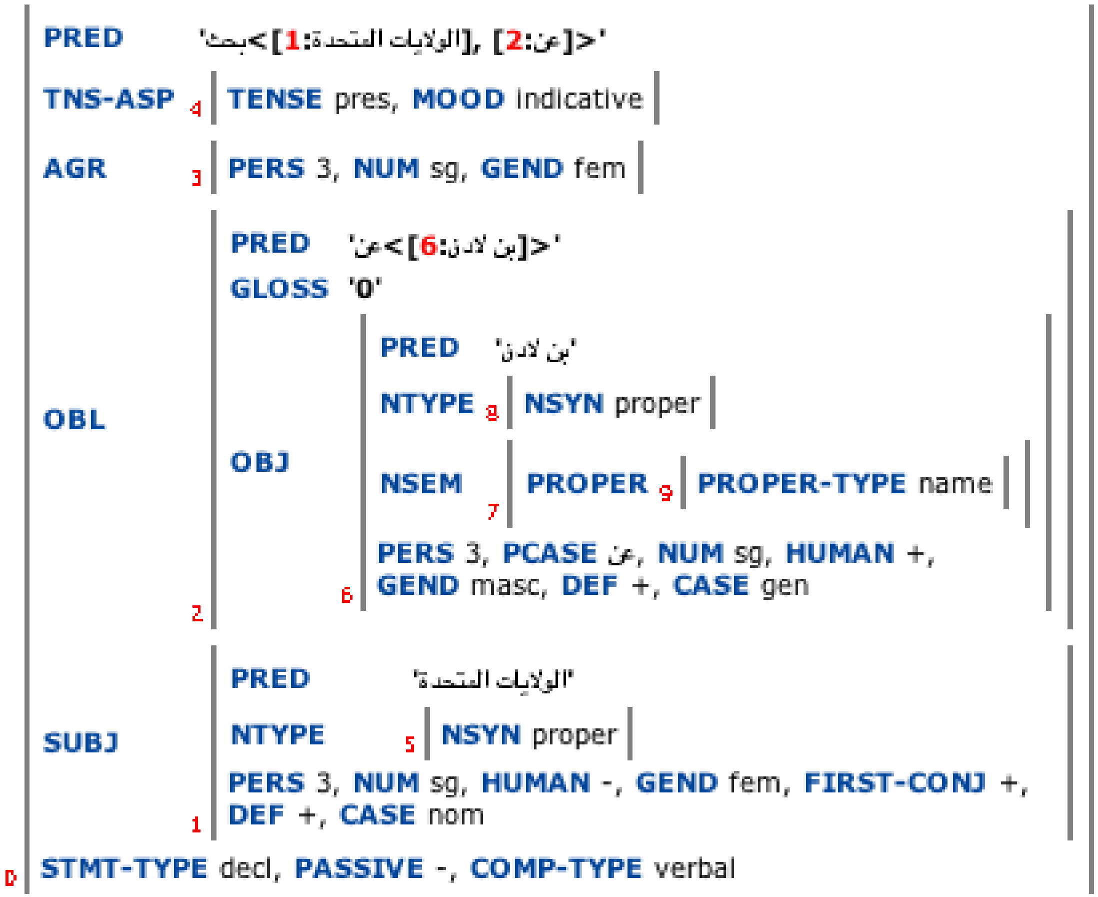
Multiword Expressions in XLE

تبحث الولايات المتحدة عن بن لادن
The United States looks for Bin Laden.

C-structure



F-structure



```
PRED      'بحث<[1:الولايات المتحدة], [2:عن]>'
TNS-ASP 4 TENSE pres, MOOD indicative
AGR     3 PERS 3, NUM sg, GEND fem
OBL         PRED  'عن<[6:بن لادن]>'
            GLOSS '0'
            OBJ     PRED   'بن لادن'
                    NTYPE 8 NSYN proper
                    NSEM  7 PROPER 9 PROPER-TYPE name
                  6 PERS 3, PCASE عن, NUM sg, HUMAN +,
                    GEND masc, DEF +, CASE gen
          2
SUBJ        PRED      'الولايات المتحدة'
            NTYPE   5 NSYN proper
          1 PERS 3, NUM sg, HUMAN -, GEND fem, FIRST-CONJ +,
            DEF +, CASE nom
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal
```

Multiword Expressions in Bikel

- Compositional, yet detectable in the English treebank

(NP (DT the) (NNP United) (NNP Kingdom))

(NP (NNP New) (NNP York))

(NP (DT the) (NNP Middle) (NNP East))

(NP (NNP Saudi) (NNP Arabia))

(NP (NNP Las) (NNP Vegas))

(NP (NNP Los) (NNP Angeles))

(CONJP (IN in) (NN addition) (TO to))

Multiword Expressions in Bikel

- Compositional, undetectable, sometimes inconsistent, in Arabic treebank

Los Angeles لوس انجليس

```
(NP (NOUN_PROP luws)
      (NOUN_PROP >anojiliys))
```

United States الولايات المتحدة

```
(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_NOM Al+wilAy+At+u)
      (DET+ADJ+NSUFF_FEM_SG+CASE_DEF_NOM Al+mut~aHid+ap+u))
```

The Middle East الشرق الأوسط

```
(NP (DET+NOUN+CASE_DEF_GEN Al+$aroq+i)
      (DET+ADJ+CASE_DEF_GEN Al+>awosaT+i))
```

in addition to إضافة إلى

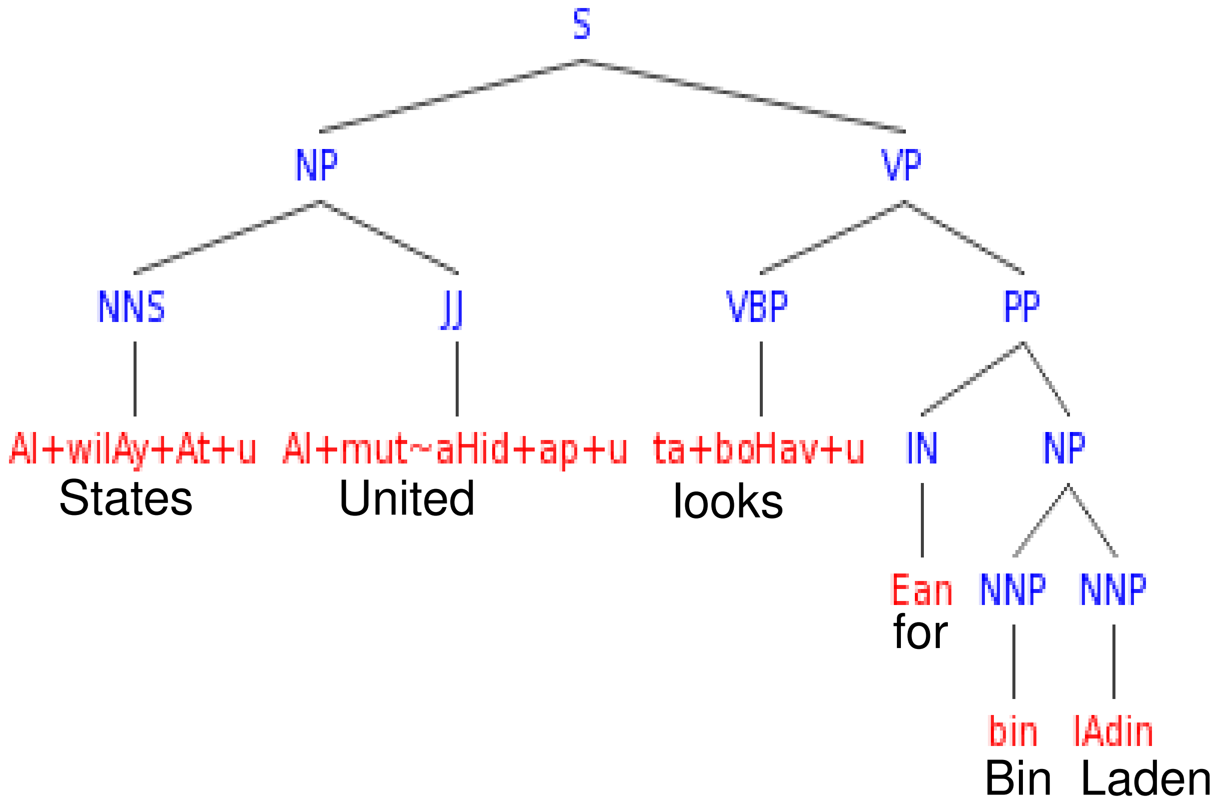
```
(CONJP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC <iDAf+ap+F) (PREP <ilaY))
```

```
(NP-ADV (NP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC -<iDAf+ap+F)) (PP (PREP
<ilaY) (NP (NP (NOUN_PROP EarafAt))
```

Multiword Expressions in Bikel

- Example

The United States looks for Bin Laden. الولايات المتحدة تبحث عن بن لادن
(S (NP (NNS Al+wilAy+At+u) (JJ Al+mut~aHid+ap+u)) (VP (VBP ta+boHav+u) (PP (IN Ean) (NP (NNP bin) (NNP lAdin)))))

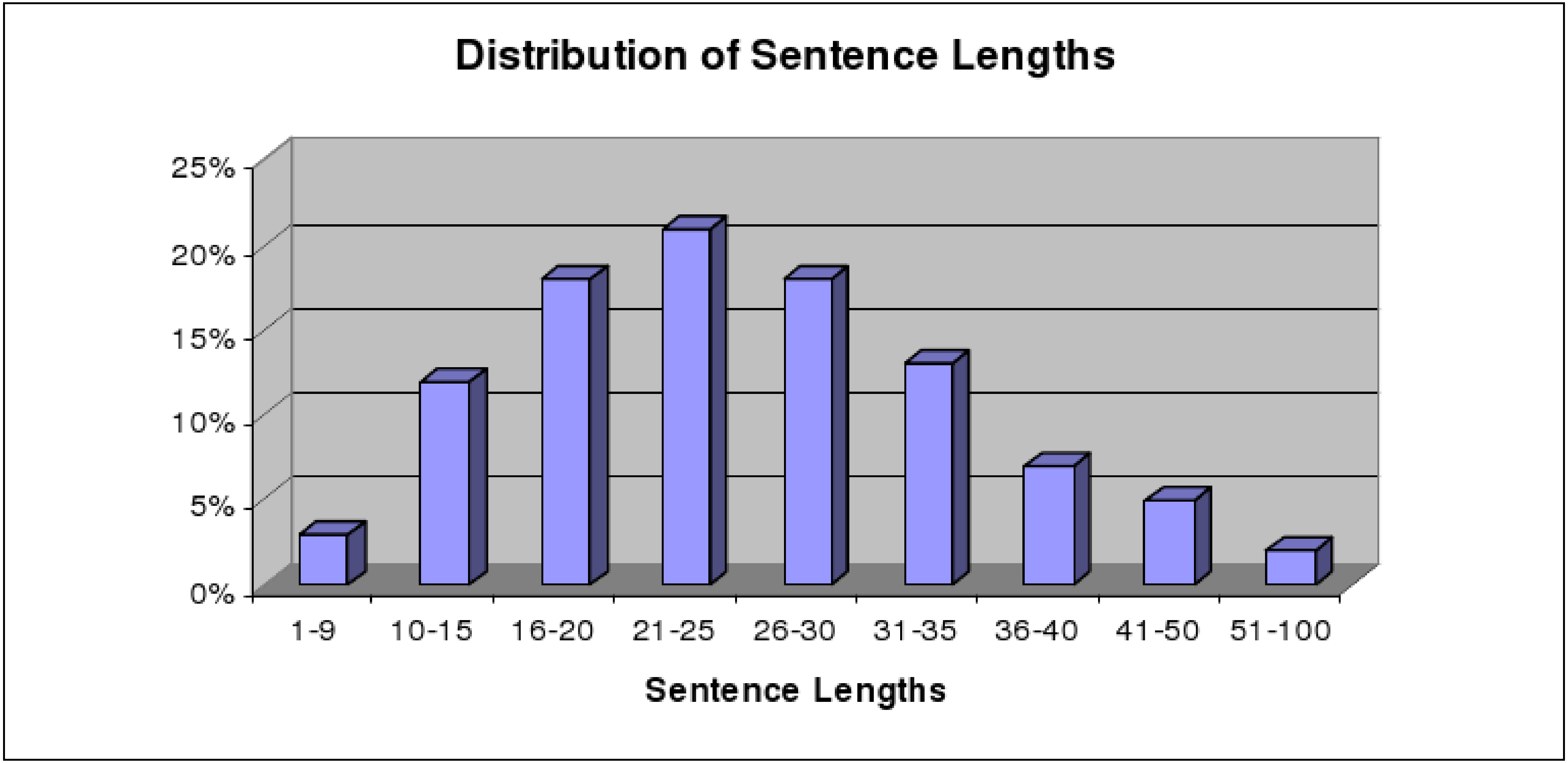


XLE Arabic Grammar Development

XLE Arabic Grammar Development

- Stage 1: Toy Grammar
  - A test suite of 175 made-up sentences
- Stage 2: Bulk Selection
  - 4 articles from Al-Jazeera are chosen as a reference for development
- Stage 3: Discriminative Selection
  - We focused on sentences with 10-15 words in length

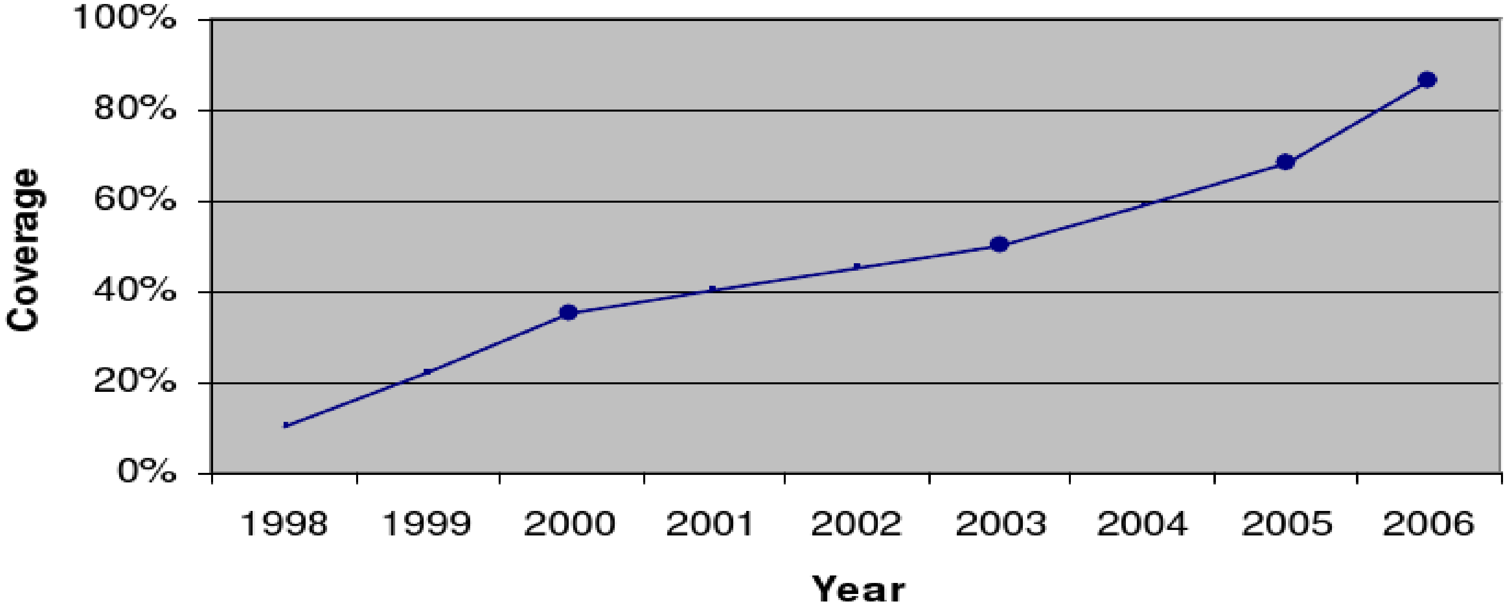
XLE Arabic Grammar Development



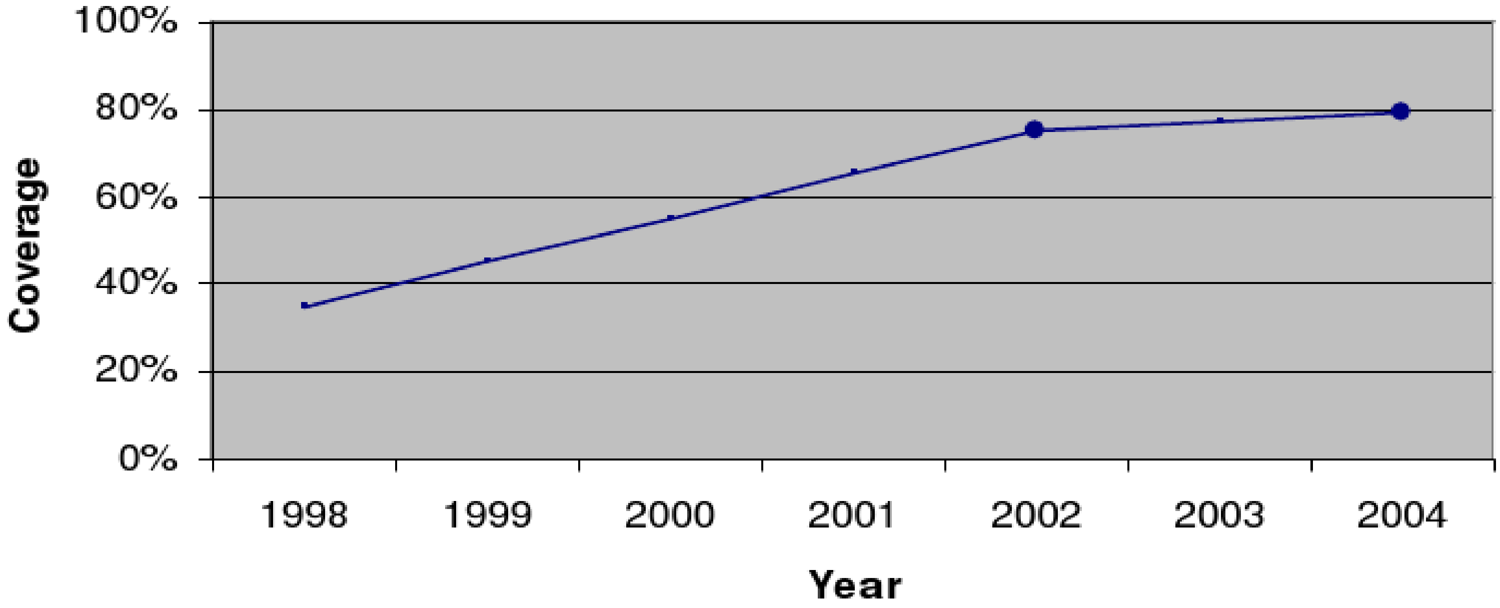
XLE Arabic Grammar Testing and Evaluation

- For sentences in the range of 10-15 words
  - 92% Fragment parsing
  - 33% Complete parses

Timeline of German Grammar Coverage



Indicative Timeline of English Grammar Coverage

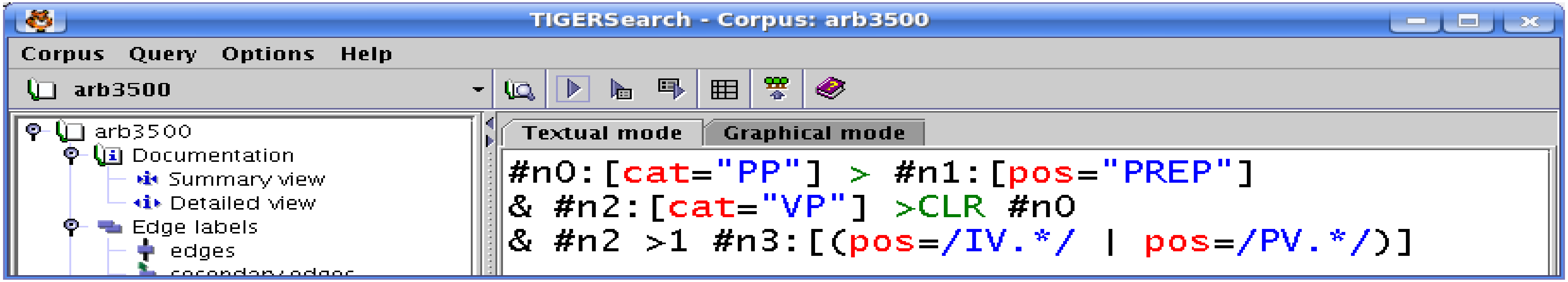


XLE Arabic Grammar Development

- Why are handcrafted grammars slow to develop?
  - There is usually a few people working in the grammar.
  - Development is hampered by linguistic (philosophical) issues that pop up frequently.
  - Speed also depends on what tools (tokenizers, morphological analysers) are already available.
  - Grammar writers are usually researchers who are more interested in linguistic phenomena than in coverage.
  - No formal guidelines, training, or project management.

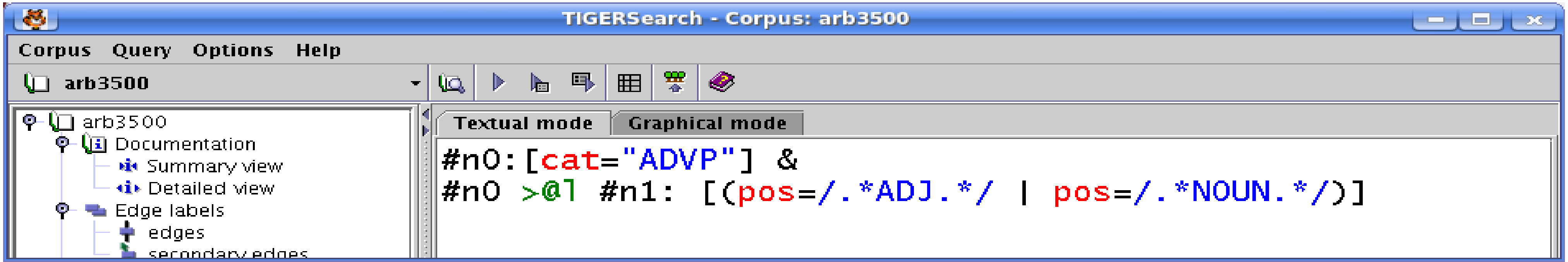
How can Arabic handcrafted grammar coverage be improved?

- Tripling the size of the morphology now 10,000 entries + 3,000 MWEs
  - This can now be done using statistical tools
    - 1195 verbs that subcategorize for prepositions (3500 sents treebank)



```
#n0:[cat="PP"] > #n1:[pos="PREP"]
& #n2:[cat="VP"] >CLR #n0
& #n2 >1 #n3:[(pos=/IV.*/ | pos=/PV.*/)]
```

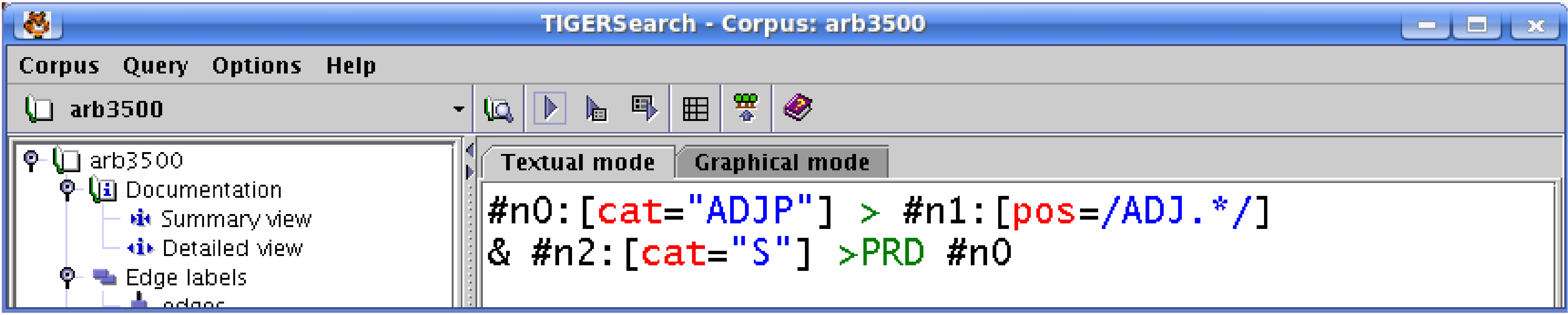
- 161 adjectives and nouns that function as adverbs



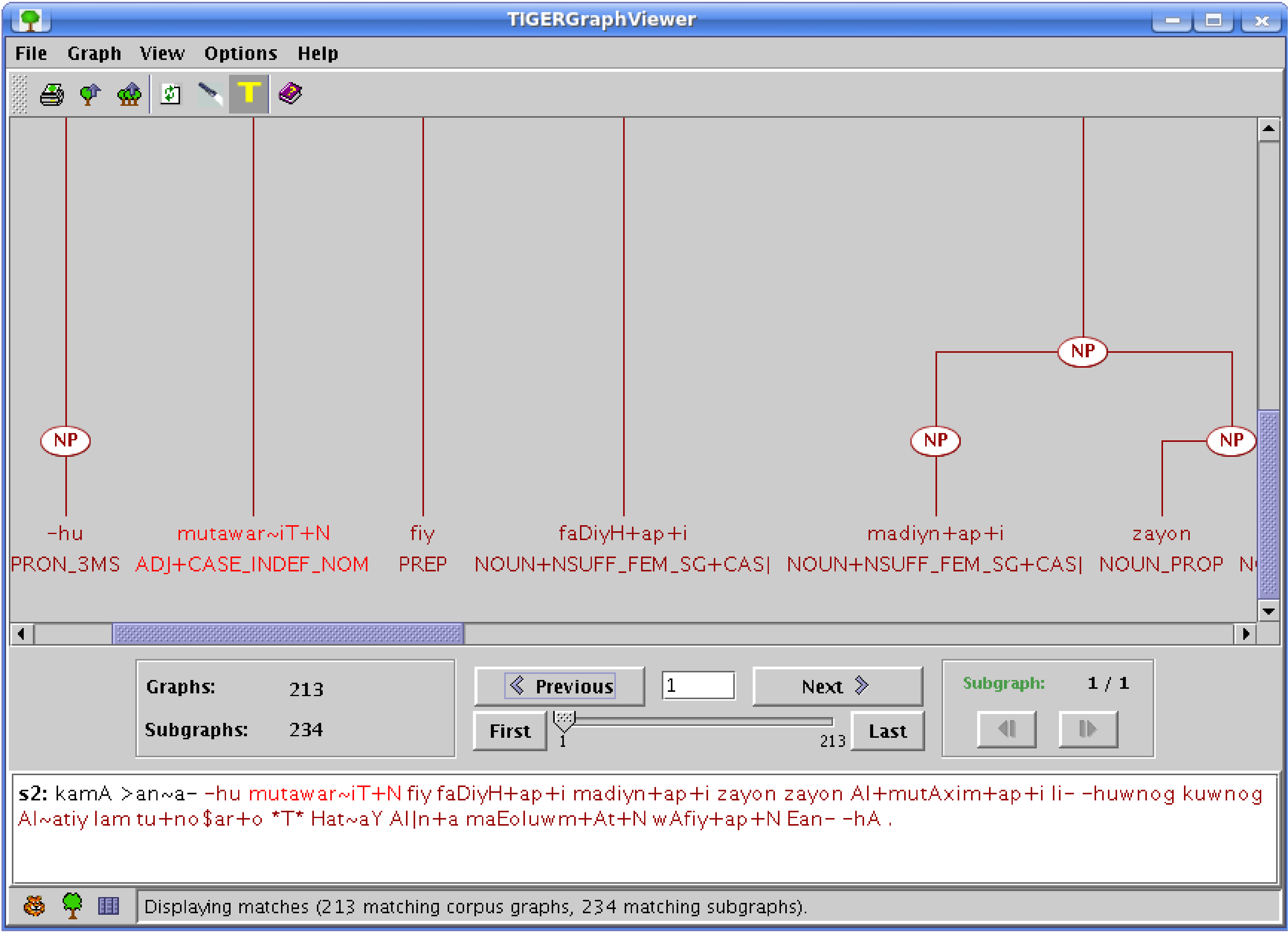
```
#n0:[cat="ADVP"] &
#n0 >@l #n1: [(pos=/.*ADJ.*/ | pos=/.*NOUN.*/)]
```

How can Arabic hand-crafted grammar be improved?

- Acquire statistics about the frequency of constructions
  - Adjectives that function as a predicate in a copula construction



```
#n0:[cat="ADJP"] > #n1:[pos=/ADJ.*/]
& #n2:[cat="S"] >PRD #n0
```



Graphs: 213

Subgraphs: 234

1

Subgraph: 1 / 1

s2: kamA >an~a- -hu mutawar~iT+N fiy faDiyH+ap+i madiyn+ap+i zayon zayon Al+mutAxim+ap+i li- -huwnog kuwnog Al~atiy lam tu+no$ar+o *T* Hat~aY Al|n+a maEoluwm+At+N wAfiy+ap+N Ean- -hA .

Displaying matches (213 matching corpus graphs, 234 matching subgraphs).

Bikel Arabic Parser Evaluation

- Coverage of the statistical parser on sentence <= 40 words:
  - Arabic: 75.4%
  - Chinese: 81%
  - English: 87.4%

    (Bikel, 2004)
  - Arabic is “far below” the required standard.

    (Kulick et al., 2006)

Bikel Arabic Parser Evaluation

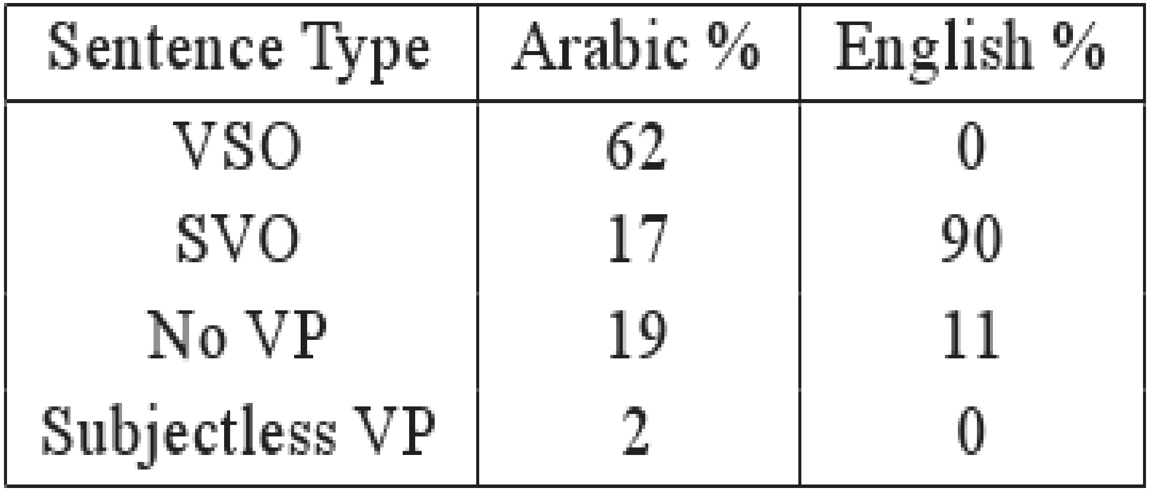
- Why Arabic performs poorly? (Kulick et al. 2006)
  - The ATB tag set is very large and dynamic, this is why they are mapped into 20 PTB tags. The tagset reduction is extreme and important information is lost.
    - Verb
      - IV3FS+IV+IVSUFF_MOOD:I
      - IV3MS+IV+IVSUFF_MOOD:J
      - PV+PVSUFF_SUBJ:3MS
      - IVSUFF_DO:3MP
    - Noun
      - NOUN+CASE_DEF_ACC
      - DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN
      - NOUN+NSUFF_FEM_SG+CASE_DEF_GEN

Bikel Arabic Parser Evaluation

- Why Arabic performs poorly? (Kulick et al. 2006)
  - Average sentence length in Arabic is 32 compared to 23 in English
  - Significant number of POS tag inconsistencies, for example *lys* is tagged as NEG_PART and PV
  - 5% of VP in Arabic have non-verbal heads
  - Base Noun Phrases (NPB) are 30% in English compared to 12% in Arabic.
  - Construct states in Arabic *roughly* correspond to possession constructions in English

Bikel Arabic Parser Evaluation

- Why Arabic performs poorly? (Kulick et al. 2006)
  - Arabic has a much greater variance in sentence structure than English.



Sentence Type	Arabic %	English %
VSO	62	0
SVO	17	90
No VP	19	11
Subjectless VP	2	0

- Major revision of Arabic treebank guidelines 08

Which is better?

Which is better?

- Common wisdom: handcrafted grammars are:
  - Time-consuming
  - Expensive
  - Require considerable linguistic and computational expertise
  - Lack coverage and robustness

(Burke et al., 2004)

Which is better?

- Common wisdom is not entirely true.
  - Creating a treebank is:
    - a “Herculean task” (Charniak, 1997)
    - very time-consuming
    - expensive
    - requires considerable linguistic and computational expertise
  -

Which is better?

- Arabic treebank annotation (2001-2008)
  - Guidelines authored by:
    - Mohamed Maamouri
    - Ann Bies
    - Sondos Krouna
    - Fatma Gaddeche
    - Basma Bouziri

    With contribution of
    - Seth Kulick
    - Wigdane Mekki
    - Tim Buckwalter

Which is better?

- Arabic treebank annotation (2001-2008)
    - List of annotators (Part 2, 2004: 4519 sentences)
        - Wigdan Mekki
        - Tasneem Ghandour
        - Ichraf Amghouz
        - Zohra Bentaouit
        - Nourredine Bessaidi
        - Rachida Fathallah
        - Niama Laadioui
        - Abid Labidi
        - Dalal Zakhary
        - Fatma Gaddeche
        - Basma Bouziri

Which is better?

- Arabic treebank annotation (2001-2008)
    - List of annotators (Part 1, 2003: 2591 sentences)
        - Wigdan El Mekki
        - Ichraf Amghouz
        - Zohra Bentaouit
        - Fatima Chebchoub
        - Fatima El Himyani
        - Rachida Fathallah
        - Alexa Firat
        - Tasneem Ghandour
        - Niama Laadioui
        - Mohamed Mansour
        - Sarah Tlili
        - Gordon Witty
        - Dalel Zakhary

Which is better?

- Arabic treebank annotation (2001-2008)
    - Logistical issues
        - Automation tools and templates
        - Tests to ensure inter-annotator agreements
        - Investigation of linguistic phenomena
        - Guidelines for consistency

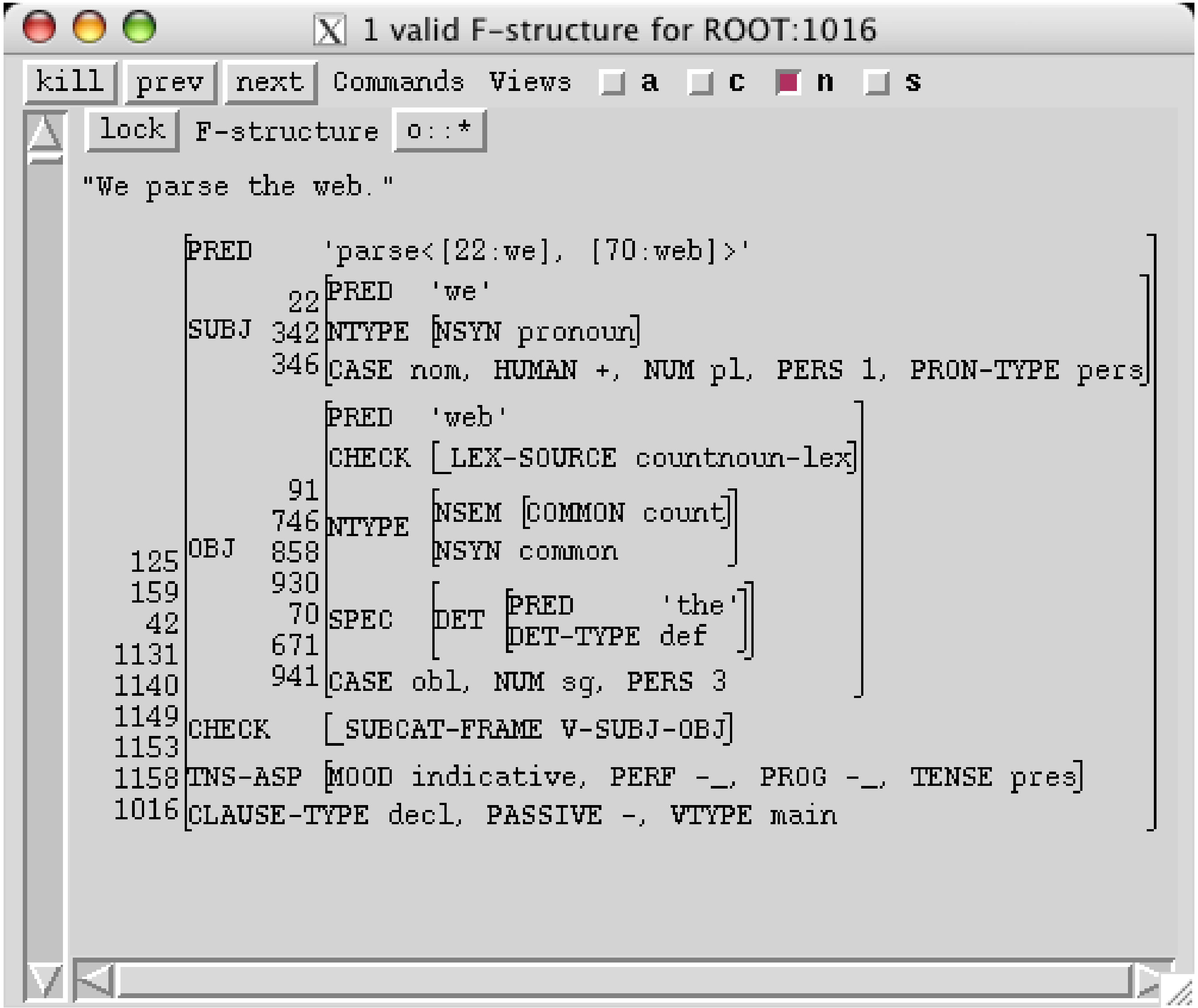
Which is better?

- Common wisdom: statistical parsers are:
  - Shallow: They do not mark syntactic and semantic dependencies needed for meaning-sensitive applications

(Kaplan et al., 2004)

Which is better?

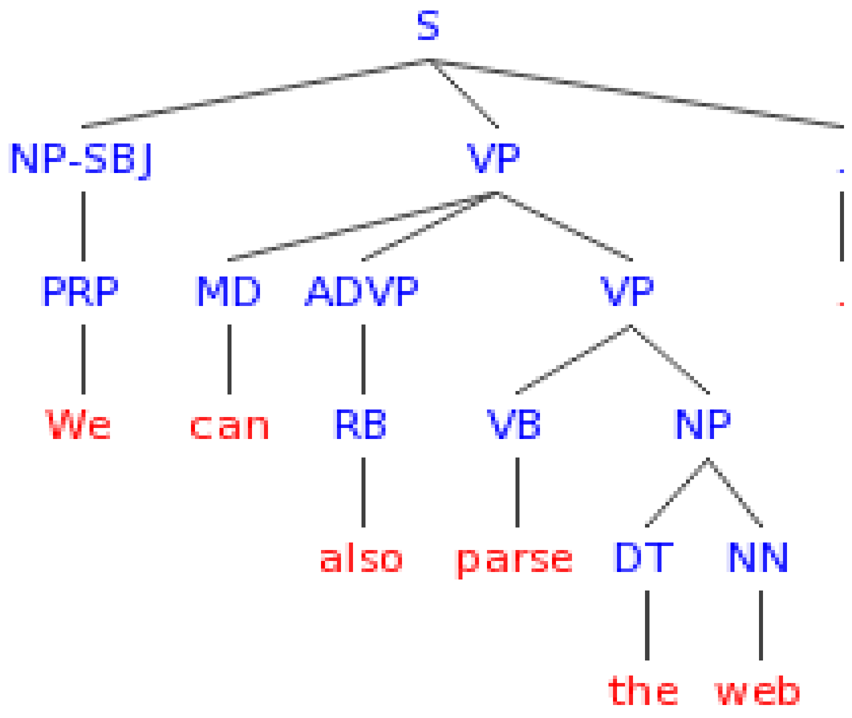
- XLE: “We parse the web.”

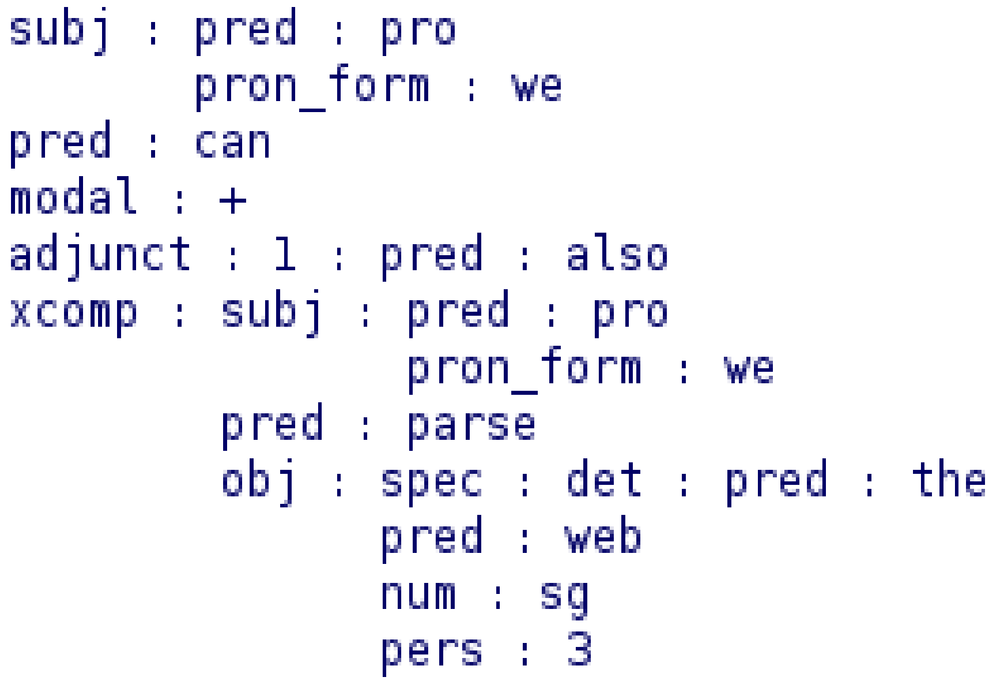


```
"We parse the web."

            PRED        'parse<[22:we], [70:web]>'
                     22 PRED  'we'
            SUBJ    342 NTYPE [NSYN pronoun]
                    346 CASE nom, HUMAN +, NUM pl, PERS 1, PRON-TYPE pers
                        PRED   'web'
                        CHECK  [LEX-SOURCE countnoun-lex]
                     91
                    746 NTYPE  NSEM [COMMON count]
        125 OBJ     858        NSYN common
        159         930
         42          70 SPEC   DET  PRED      'the'
       1131         671             DET-TYPE  def
       1140         941 CASE obl, NUM sg, PERS 3
       1149
       1153 CHECK       [SUBCAT-FRAME V-SUBJ-OBJ]
       1158 TNS-ASP     [MOOD indicative, PERF -_, PROG -_, TENSE pres]
       1016 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
```

Which is better?

- Common wisdom is not entirely true.
- DCU: “We can also parse the web.”





```
subj : pred : pro
       pron_form : we
pred : can
modal : +
adjunct : 1 : pred : also
xcomp : subj : pred : pro
               pron_form : we
        pred : parse
        obj : spec : det : pred : the
              pred : web
              num : sg
              pers : 3
```

Which is better?

- Summary
  - Handcrafted grammars are built on assumptions and intuitions. They depend on how good these assumptions are.
  - Handcrafted grammar can be improved by:
    - Effectively managing the development project
    - Making use of statistical facts (treebanks, and TIGERSearch)

Which is better?

- Statistical grammars are built on facts. They depend on how true these facts are.
- Statistical grammar can be improved by:
  - Improving the quality and size of treebanks.

Which is better?

- Statistical grammars are more efficient because:
  - there is a clear separation between the algorithm and the data structure
  - there is a clear division of labour, the linguists fight their battle, and the engineers fight their own battle

Which is better?

- Hybridization? Complementation? Cooperation?
  - Statical parser is used to increase the efficiency of hand-crafted grammar (pruning the search space)
  - Hand-crated grammars are used automate the creation of treebanks (Norwegian grammar)
  - Some languages do not have a treebank
- This is for the future to decide.