Arabic Language: Nature and Challenges

Mohammed Attia

The British University in Dubai

May 29, 2012

Outline

- Introduction
- Lexicons and Corpus Linguistics
- Morphology
- Syntactic Parsing
- Tokenization
- Multiword Expressions
- Statistical Parsing
- Which is better?
- Spelling Checking and Correction
- Integration with Applications

Introduction

Introduction

- Living language are just … living

Introduction

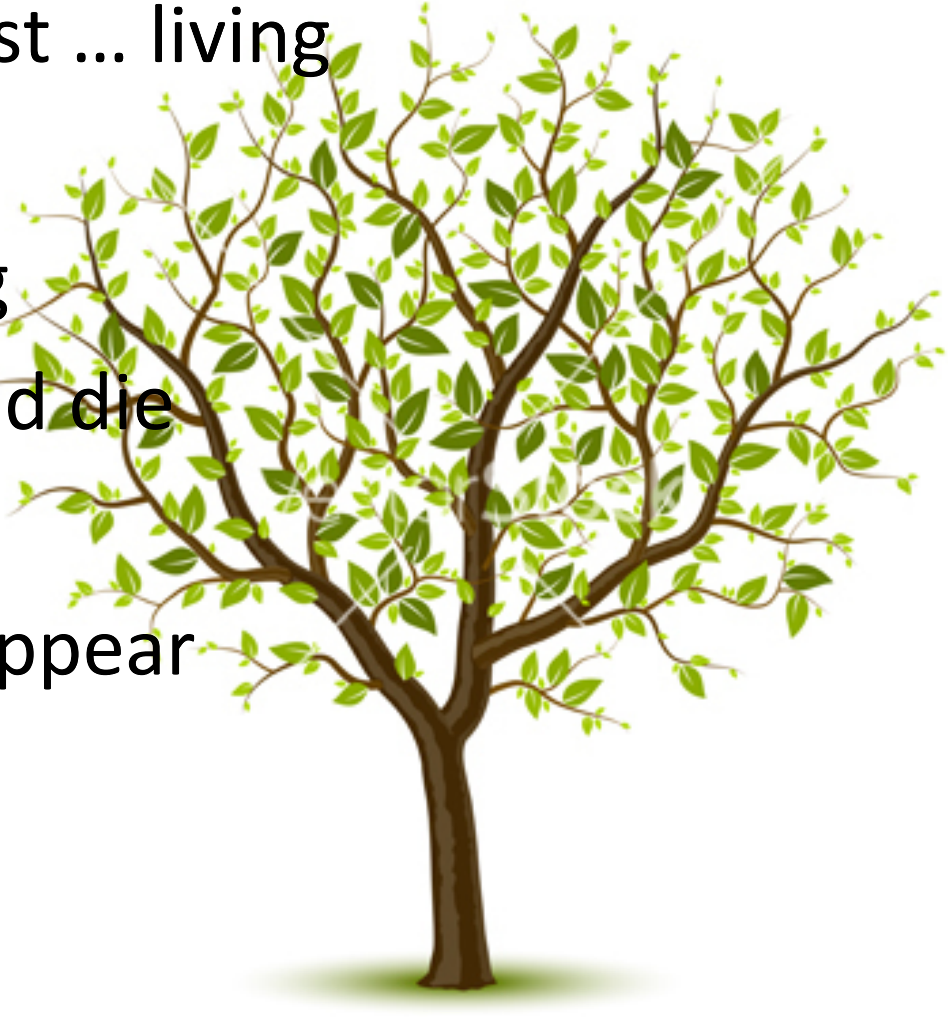
- Living language are just … living
- They grow

Introduction

- Living language are just … living
- Leaves fall off

Introduction

- Living language are just … living
- New leaves appear
- … constantly changing
- They may grow old and die
- They maybe reborn
- New languages may appear



Introduction

- Language reveals everything about us …

Introduction

- Language tell everything about us ….
- How rich or poor

Introduction

- Language tell everything about us ….
- How well-educated

Introduction

- Language tell everything about us….
- where we come from

Introduction

- Language tell everything about us ….
- what kind of work we do

Introduction

- Language tell everything about us ….
- our feelings and our sentiments

Introduction

- Language is the key to business expansion: Translation and localization

World translation business in 2011 = $30 billion

To: Rose Hall

Cc:

Subject: Re: Shopping List

On Dec 19, 2006, at 9:18 AM, Rose Hall <rosemarie.hall@mac.com> wrote:

Introduction

- And a repository of knowledge and information

Lexicons and Corpus Linguistics

Principles of Lexicography

مبادئ علم صناعة المعاجم

Definition of a dictionary

- A description of the vocabulary (حصيلة لغوية) used by members of a speech community (مجتمع يتحدث نفس اللغة). A dictionary deals with:
  - Conventions عرفي not idiosyncrasies شخصي
  - norms سائد not rarities نادر
  - Probable واقع not possible ممكن نظريا

- Lexical evidence
  - Subjective evidence
    - Introspection الاستبطان
    - informant-testing استشارة أصحاب المعرفة
  - Objective evidence
    - A corpus (ذخيرة النصوص) provides typifications (تصنيف للأنماط) of the language
      - A typical lexical entry means it is both "frequent" متكرر or "recurrent" دوري and "well-dispersed" موزع ومتفرق in a corpus.
      - A typical lexical entry belongs to the stable "core" of the language.

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford *University Press.*

Principles of Lexicography

- **Corpora and Dictionaries**
  - Brown Corpus, 1 million words, 1960s,
    → Citations for *American Heritage Dictionary*
  - Birmingham corpus, 20 million words, 1980s
    → Cobuild English Dictionary.
  - British National Corpus (BNC), 100 million words, 1990s set the standard (balance, encoding)
  - The Oxford English Corpus, one billion words, 2000s
    → Oxford English Dictionary
  - Longman Corpus Network, 330 million word
    → Longman Dictionaries

Principles of Lexicography

- **Dictionaries before Corpora**
  - Citation banks مراجع اقتباسية استشهادية
    - A citation is a short extract providing evidence for a word usage or meaning in authentic use.
  - Disadvantages
    - labour-intensive
    - instances of usage are authentic, but there is a big subjective element in their selection.
      - People tend to notice what is remarkable and ignore what is typical
      - bias towards the novel or idiosyncratic usages

Principles of Lexicography

- **Characteristics of a reliable corpus** (مواصفات ذخيرة النصوص)
  - The corpus does not favour high class language
  - The Corpus should be large and diverse
  - The corpus should be either synchronic or diachronic
  - The corpus should be well-balanced using “stratified sampling” (أخذ عينات بشكل نسبي)
  - The corpus should avoid skewing (الانحراف أو التحيز)

Principles of Lexicography

- **Lexical Profiling**
  - Word POS

    v, n, adj, adv, conj, det, interj, prep, pron
  - Valency Information

    subcat frames, other obligatory

    or optional syntactic constructions
  - Collocations

    commit a crime, sky blue, lame duck

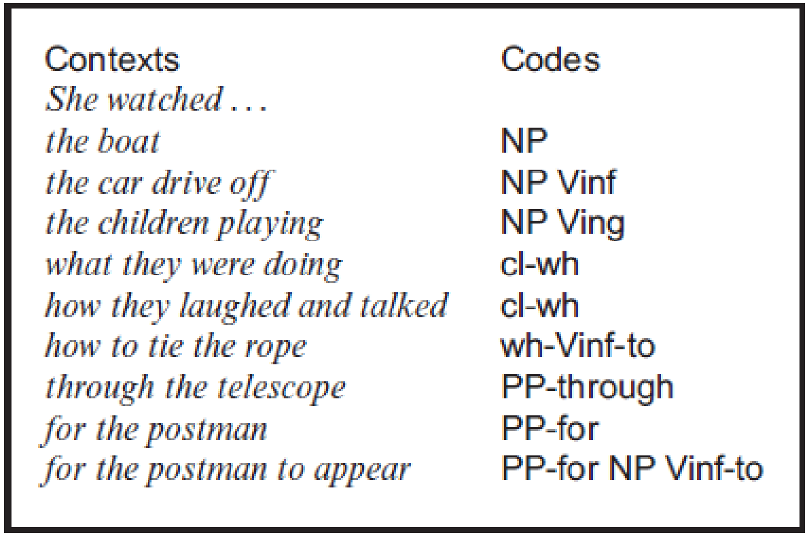
ظلام دامس، ارتكب جريمة، فعل فعلة

  - Colligational preferences

    was acquitted,

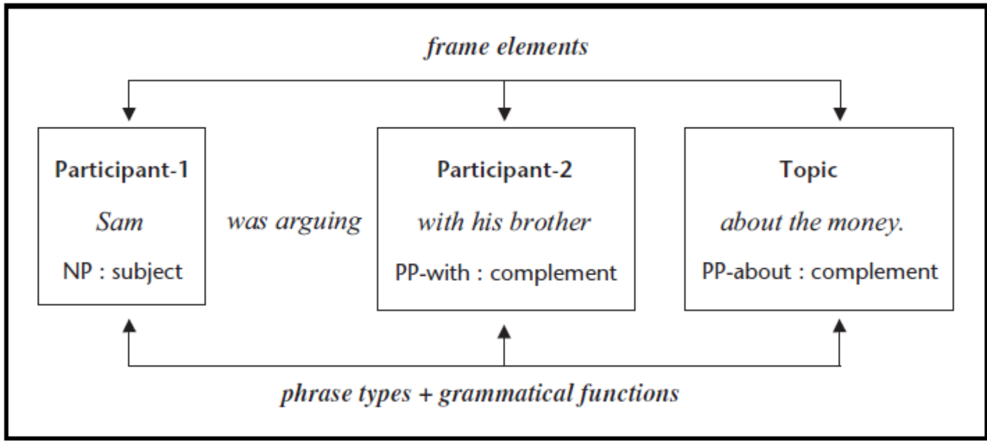
    trials (difficult experiences)

لم يأبه



Contexts	Codes
She watched . . .	
the boat	NP
the car drive off	NP Vinf
the children playing	NP Ving
what they were doing	cl-wh
how they laughed and talked	cl-wh
how to tie the rope	wh-Vinf-to
through the telescope	PP-through
for the postman	PP-for
for the postman to appear	PP-for NP Vinf-to

Some constructions for the verb *watch*



frame elements

Participant-1

Sam

NP : subject

was arguing

Participant-2

with his brother

PP-with : complement

Topic

about the money.

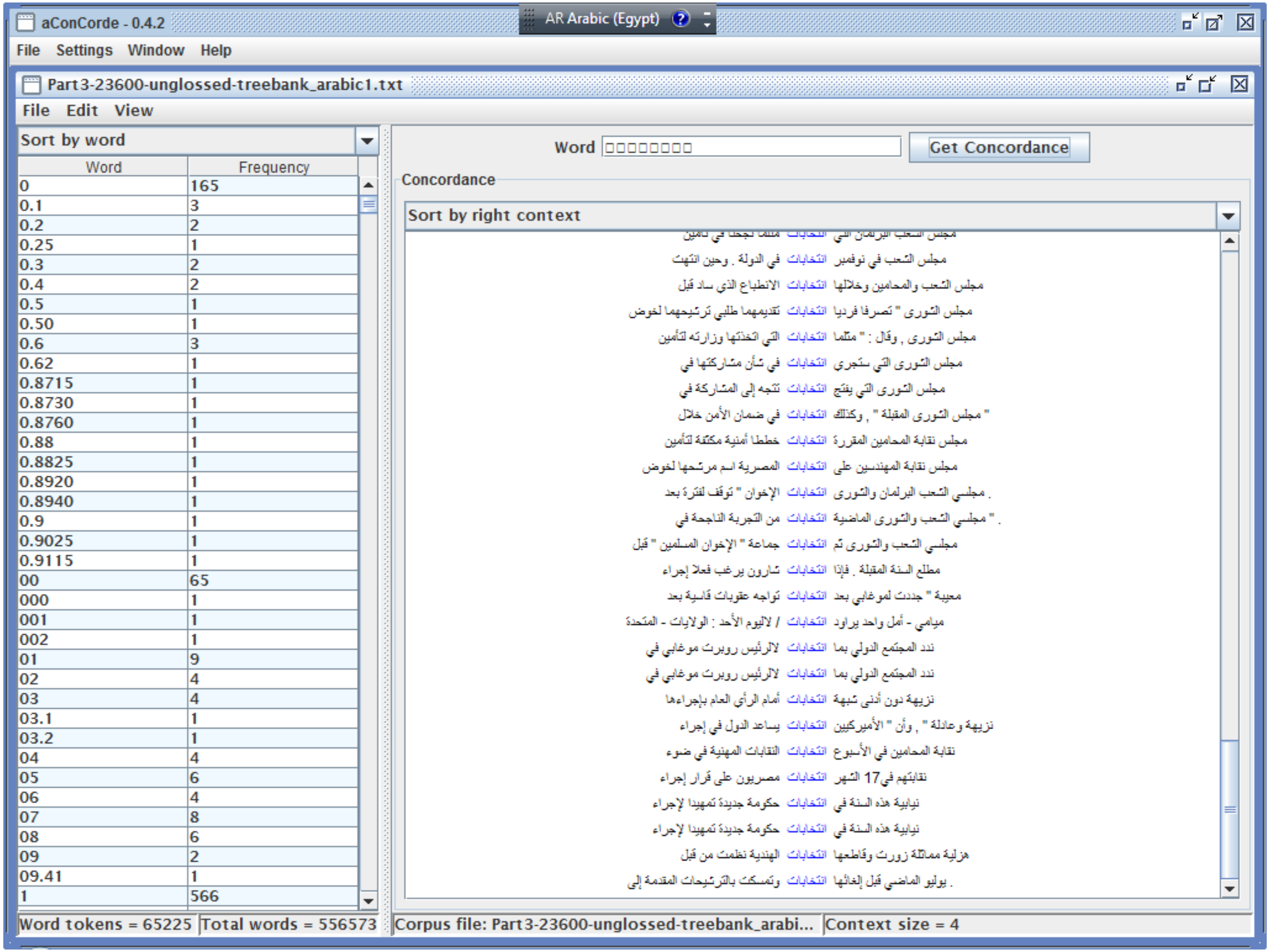
PP-about : complement

phrase types + grammatical functions

Principles of Lexicography

- **Lexical Profiling Software**
- Concordancers
- Word Sketch (Sketch Engine) - Adam Kilgarriff

Concordancer



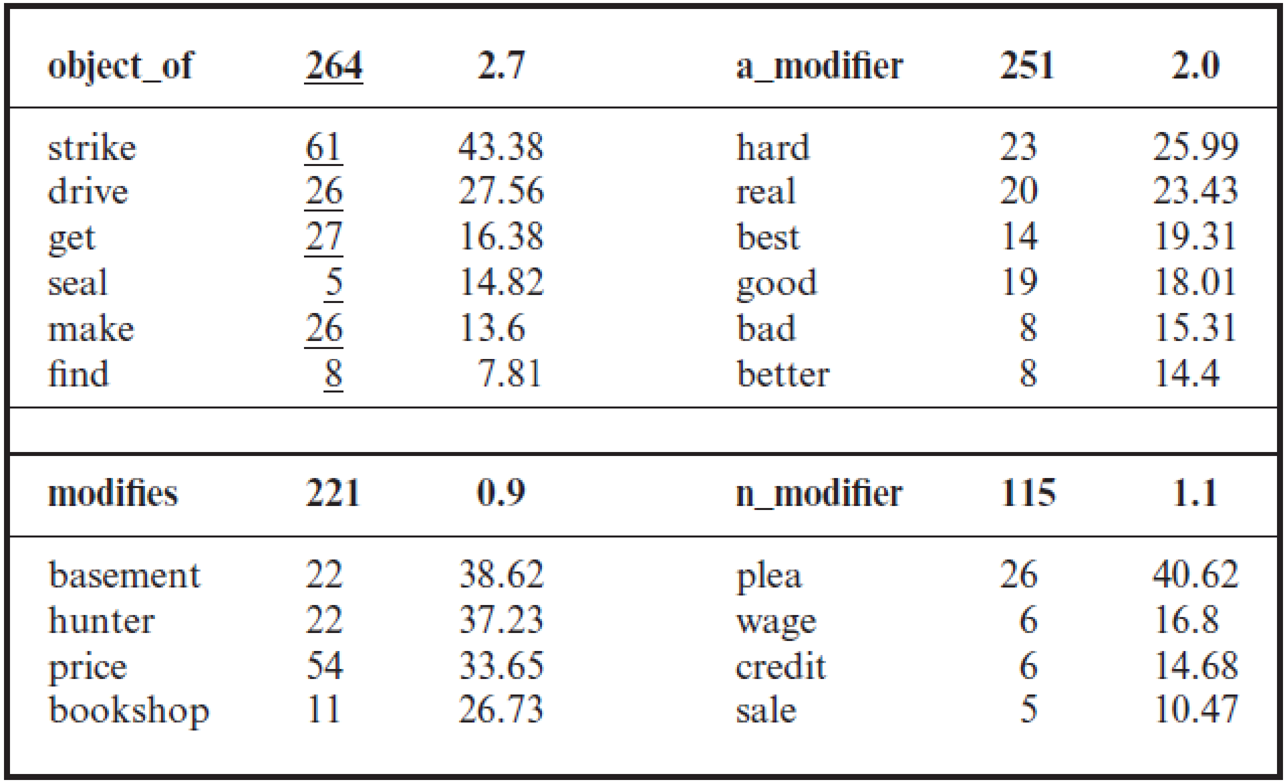
Word	Frequency
0	165
0.1	3
0.2	2
0.25	1
0.3	2
0.4	2
0.5	1
0.50	1
0.6	3
0.62	1
0.8715	1
0.8730	1
0.8760	1
0.88	1
0.8825	1
0.8920	1
0.8940	1
0.9	1
0.9025	1
0.9115	1
00	65
000	1
001	1
002	1
01	9
02	4
03	4
03.1	1
03.2	1
04	4
05	6
06	4
07	8
08	6
09	2
09.41	1
1	566

Word tokens = 65225 Total words = 556573

مجلس الشعب البرلمان التي انتخابات مثلما نجحا في تأمين
مجلس الشعب في نوفمبر انتخابات في الدولة . وحين انتهت
مجلس الشعب والمحامين وخلالها انتخابات الانطباع الذي ساد قبل
مجلس الشورى " تصرفا فرديا انتخابات تقديمهما طلبي ترشيحهما لخوض
مجلس الشورى , وقال : " مثلما انتخابات التي اتخذتها وزارته لتأمين
مجلس الشورى التي ستجري انتخابات في شأن مشاركتها في
مجلس الشورى التي يفتح انتخابات تتجه إلى المشاركة في
" مجلس الشورى المقبلة " , وكذلك انتخابات في ضمان الأمن خلال
مجلس نقابة المحامين المقررة انتخابات خططا أمنية مكثفة لتأمين
مجلس نقابة المهندسين على انتخابات المصرية اسم مرشحها لخوض
. مجلسي الشعب البرلمان والشورى انتخابات الإخوان " توقف لفترة بعد
. " مجلسي الشعب والشورى الماضية انتخابات من التجربة الناجحة في
مجلسي الشعب والشورى ثم انتخابات جماعة " الإخوان المسلمين " قبل
مطلع السنة المقبلة . فإذا انتخابات شارون يرغب فعلا إجراء
معيبة " جددت لموغابي بعد انتخابات تواجه عقوبات قاسية بعد
ميامي - أمل واحد يراود انتخابات / لااليوم الأحد : الولايات - المتحدة
ندد المجتمع الدولي بما انتخابات لالرئيس روبرت موغابي في
ندد المجتمع الدولي بما انتخابات لالرئيس روبرت موغابي في
نزيهة دون أدنى شبهة انتخابات أمام الرأي العام بإجراءها
نزيهة وعادلة " , وأن " الأميركيين انتخابات يساعد الدول في إجراء
نقابة المحامين في الأسبوع انتخابات النقابات المهنية في ضوء
نقابتهم في 17 الشهر انتخابات مصريون على قرار إجراء
نيابية هذه السنة في انتخابات حكومة جديدة تمهيدا لإجراء
نيابية هذه السنة في انتخابات حكومة جديدة تمهيدا لإجراء
هزلية مماثلة زورت وقاطعها انتخابات الهندية نظمت من قبل
. يوليو الماضي قبل إلغائها انتخابات وتمسكت بالترشيحات المقدمة إلى

Corpus file: Part3-23600-unglossed-treebank_arabi... Context size = 4

Sketch Engine



object_of	**264**	**2.7**	**a_modifier**	**251**	**2.0**
strike	61	43.38	hard	23	25.99
drive	26	27.56	real	20	23.43
get	27	16.38	best	14	19.31
seal	5	14.82	good	19	18.01
make	26	13.6	bad	8	15.31
find	8	7.81	better	8	14.4
modifies	**221**	**0.9**	**n_modifier**	**115**	**1.1**
basement	22	38.62	plea	26	40.62
hunter	22	37.23	wage	6	16.8
price	54	33.65	credit	6	14.68
bookshop	11	26.73	sale	5	10.47

Part of the Word Sketch for the noun *bargain*

1973 مستويات العربية المعاصرة
للسعيد محمد بدوي

- فصحى التراث
- فصحى العصر
- عامية المثقفين
- عامية المتنورين
- عامية الأميين

Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

- Modern Standard Arabic
  - The language of modern writing, prepared speeches and the language of the news
- Classical Arabic
  - The language of Arabia before Islam and after Islam until the Medieval Times
  - Present religious teaching, poetry and scholarly literature.
- Colloquial Arabic
  - Variety of Arabic spoken regionally and which differs from one country or area to another. They are to a certain extent mutually intelligible.

Code Shifting – Code Switching – Diglossia – multi-layered diglossia

Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

- Modern Standard Arabic
  - Tendency for simplification
    - Some CA structures to die out
    - Structures marginal in CA started to have more salience
    - no strict abidance by case ending rules
  - A subset of the full range of structures, inflections and derivations available in CA
  - MSA conforms to the general rules of CA
  - How "big" or how "small" the difference (on morphological, lexical or syntactic levels) need more research and investigation

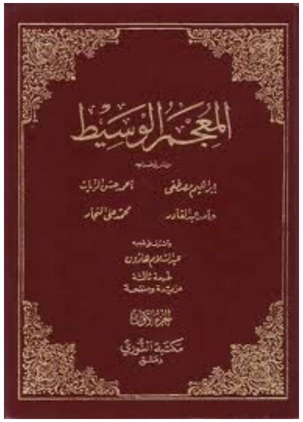
Review of Arabic lexicographic work

- *Kitab al-'Ain* by al-Khalil bin Ahmed al-Farahidi (died 789)

(refinement/expansion/organizational Improvement)

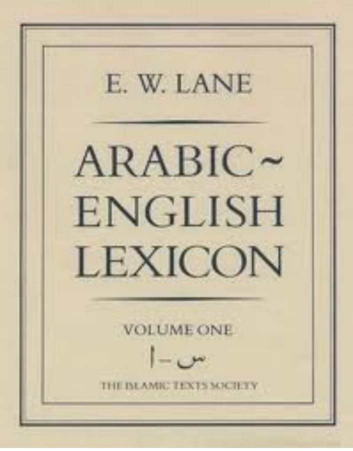
▼

- *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980)
- *al-Muheet* by al-Sahib bin 'Abbad (died 995)
- *Lisan al-'Arab* by ibn Manzour (died 1311)
- *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414)
- *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791)
- *Muheet al-Muheet* (1869) by Butrus al-Bustani
- *al-Mu'jam al-Waseet* (1960)



Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Edward William Lane's *Arabic-English Lexicon* (1876) indebted to *Taj al-Arous* by al-Zabidi
  - Hans Wehr's Dictionary of Modern Written Arabic (1961)
    - Size: 45,000 entries
    - Aim: Using scientific descriptive principles to describe present-day vocabulary through wide reading in literature of every kind
    - Application
      - Selection of works by high flying poets and literary critics such as Taha Husain, Taufiq al-Hakim, Mahmoud Taimur, al-Manfalauti, Jubran Khalil Jubran
      - Use of secondary sources (dictionaries) for expansion
      - Inclusion of rarities and classicisms that no longer formed a part of the living lexicon



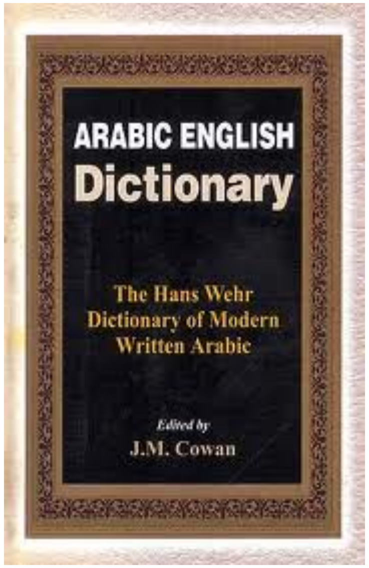
E. W. LANE

ARABIC~
ENGLISH
LEXICON

VOLUME ONE

س - ا

THE ISLAMIC TEXTS SOCIETY



ARABIC ENGLISH
Dictionary

The Hans Wehr
Dictionary of Modern
Written Arabic

Edited by
J.M. Cowan

Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Landau and Brill (1959) *A Word Count of Modern Arabic Prose*
    - A word count based on 270,000 words based on: 136,000 from the news (Moshe Brill, 1940) and 136,000 from 60 contemporary books on:
      fiction, literary criticism, history, biography, political science, religion, social studies, economics, travels and historical novels
    - 6,000 words in the news
    - 11,000 words in literature
    - 12,400 words in the combined list (does not include proper nouns)

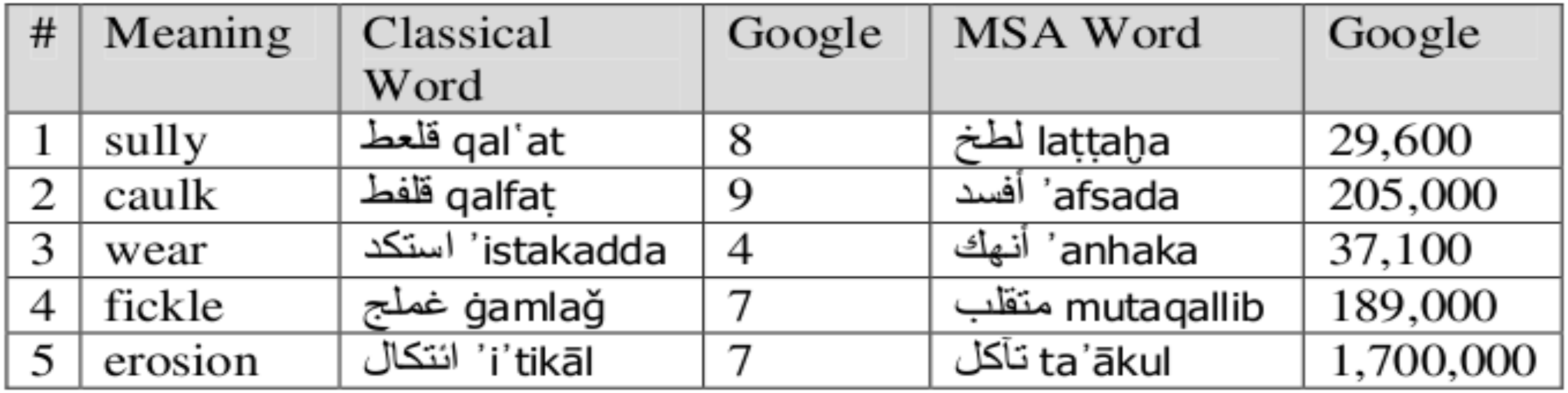
Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Van Mol's (2000) Arabic-Dutch learner's dictionary
    - COBUILD-style, Corpus-based (3 million words)
    - Manually constructed
    - Covers the whole range of the actual vocabulary in the corpus with 17,000 entries compared to 45,000 entries in Hans Wehr
    - 5% of frequent new words not found in Hans Wehr

Review of Arabic lexicographic work

- Buckwalter Arabic Morphological Analyzer (2002)
  - Size: 40,222 lemmas (including 2,034 proper nouns)
  - Includes many obsolete lexical items

(But how many?)



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

Google score for Classical vs. MSA entries

Corpus-based Lexicon

Largest corpus of modern Arabic to date
Arabic gigaword 1,200,000,000
= 16,000 large books
= 800 meters of bookshelves
Burj Khalifah is 830m

Avr reader reads 200 wpm
With 60% comprehension.

You will need 11 years 24/7
to read the Gigaword corpus

BurjDubaiSkyscraper.com
Imre Solt - 2009 ©

Review of Arabic lexicographic work

Buckwalter obsolete words: 8,400 obsolete words

صحراء: فَيْفاء فَدْفَد قَواء مَوْماة مَتْلَف سَبْسَب   **رمل**: هَيَلان وَعْس مِيعاس عِثْيَر

سرج: حِداجَة مَخْلُوفَة

حِمْل: ظَعِينَة حِدْج ظَعُون وِقْر

لجام: فدام كَعَم كِعام أَرْنَبَة شَكَم غِمامَة

راكب: حَدّاء

جمل: هَجِينَة

رداء: دِفِّيَّة بِشْتَة

حذاء: مبْذَل بَشْمَق زَرْبُول زَرْبُون صَرْمَة قَبْقاب

Review of Arabic lexicographic work

Not in Dictionaries: about 10,000 need to be added

سياسة: أمننة شرعنة أفروعربية إثني إقصائي تسييس محاصصة جبهوي جمهوعسكرية العصبوية شخصنة أمركة عصرنة

تكنولوجيا:

رقمنة، أتمتة، مَكْنَنة

فيس بوك، تويتر، تغريدة

هاتف \ جوال \ تليفون محمول

لاب توب

الهواتف الذكية

حوسبة

بريد إلكتروني

آي فون، دي في دي، سي دي

سبام، فيروس

ملتي ميديا

كمبيوتر لوحي، شاشة لمسية

شيفرة

اقتصاد: خصخصة ريعي يورو بورصة تعويم داو_جونز تضخم أسهم قيمة_دفترية مليار ترليون تجارة_إلكترونية

Morphological Lexicon AraComLex

- How our lexical database will be different from Buckwalter's. We include
  - only entries attested in a corpus
  - subcategorization frames
  - +/-human semantic information for nouns
  - Information on allowing passive and imperative inflection for verbs
  - Information on diptotes
  - detailed information about derived nouns/adjectives (active or passive participle or a verbal noun, *masdar*)
  - multi-word expressions
  - classification of proper nouns: person, place, organization, etc.
  - Frequency information
  - Citation in real examples

Corpus-based Lexicon

Lexicons as a truthful representation of the language as evidenced in a corpus

The Arabic Gigaword corpus

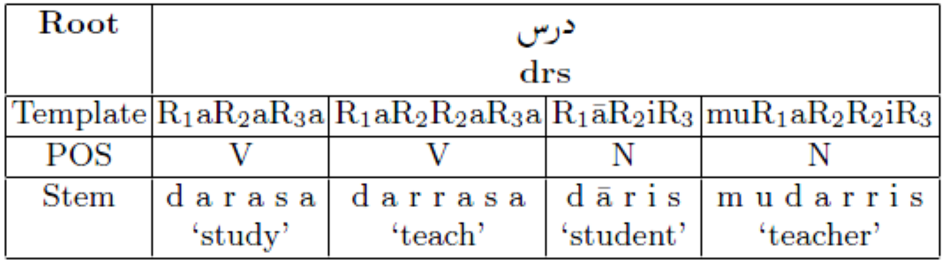
AraComLex

Our morphological analyser – based on a lexical database automatically derived from the Arabic Gigaword Corpus

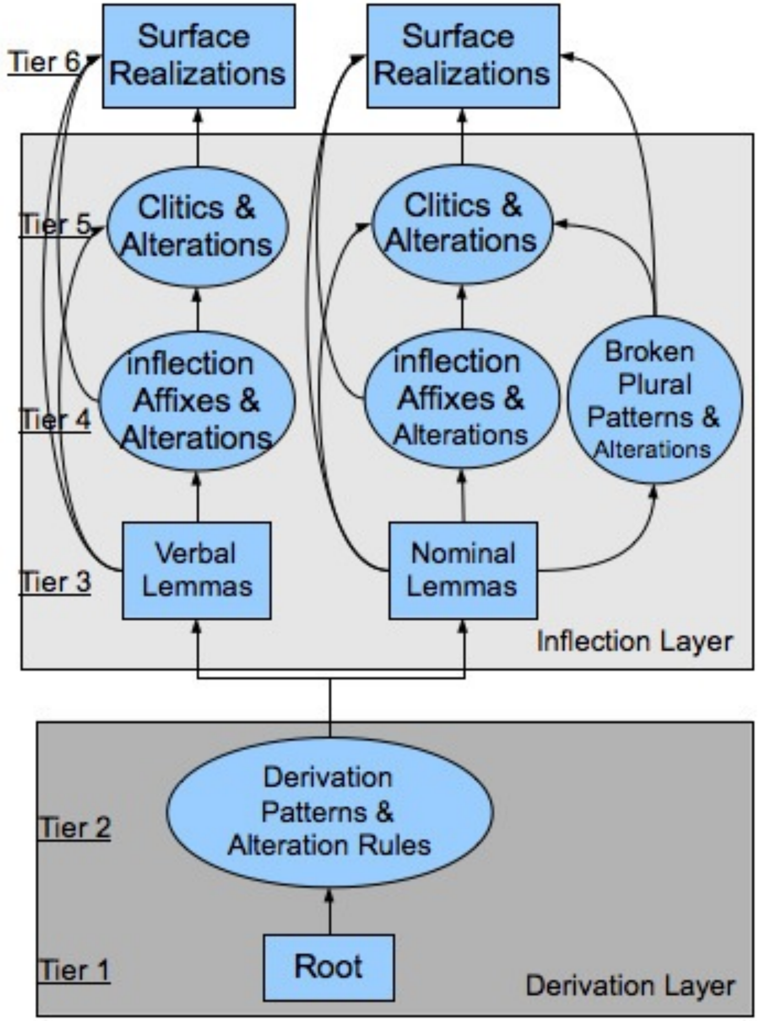
Morphology

Arabic Morphology

- Arabic Morphotactics



Root	درس drs			
Template	$R_1aR_2aR_3a$	$R_1aR_2R_2aR_3a$	$R_1\bar{a}R_2iR_3$	$muR_1aR_2R_2iR_3$
POS	V	V	N	N
Stem	darasa ‘study’	darrasa ‘teach’	dāris ‘student’	mudarris ‘teacher’



Arabic Morphology

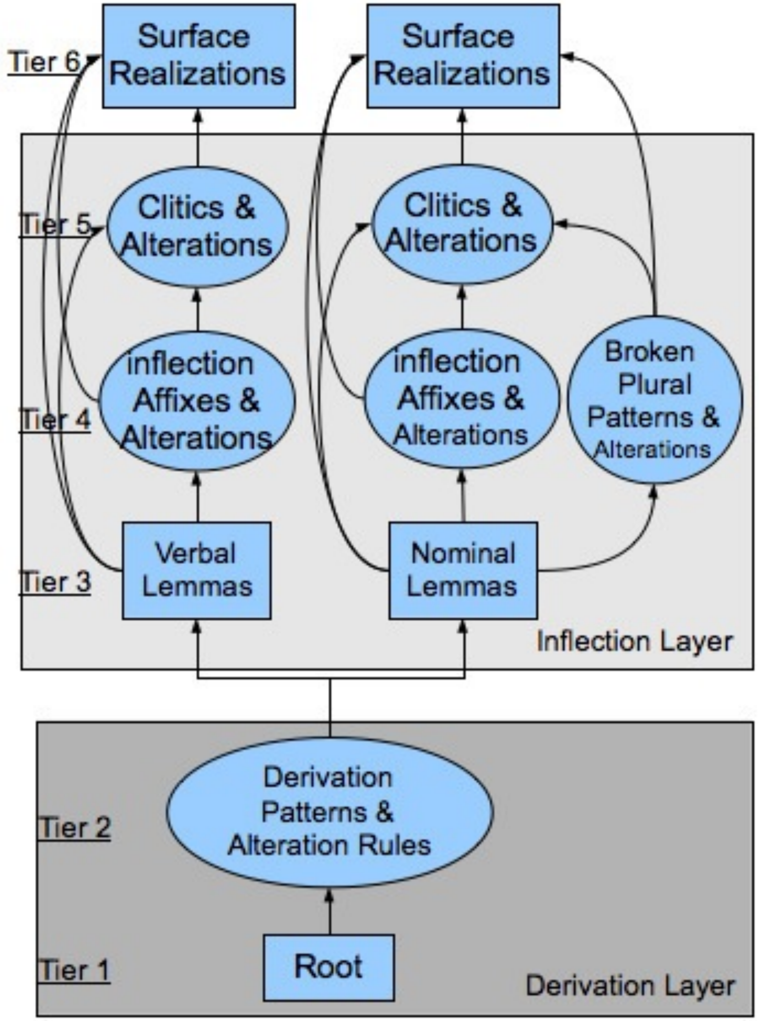
- Design Approach: Three approaches

1. Root-based Morphology
   Xerox Arabic FTM

2. Stem-based morphology
   Buckwalter

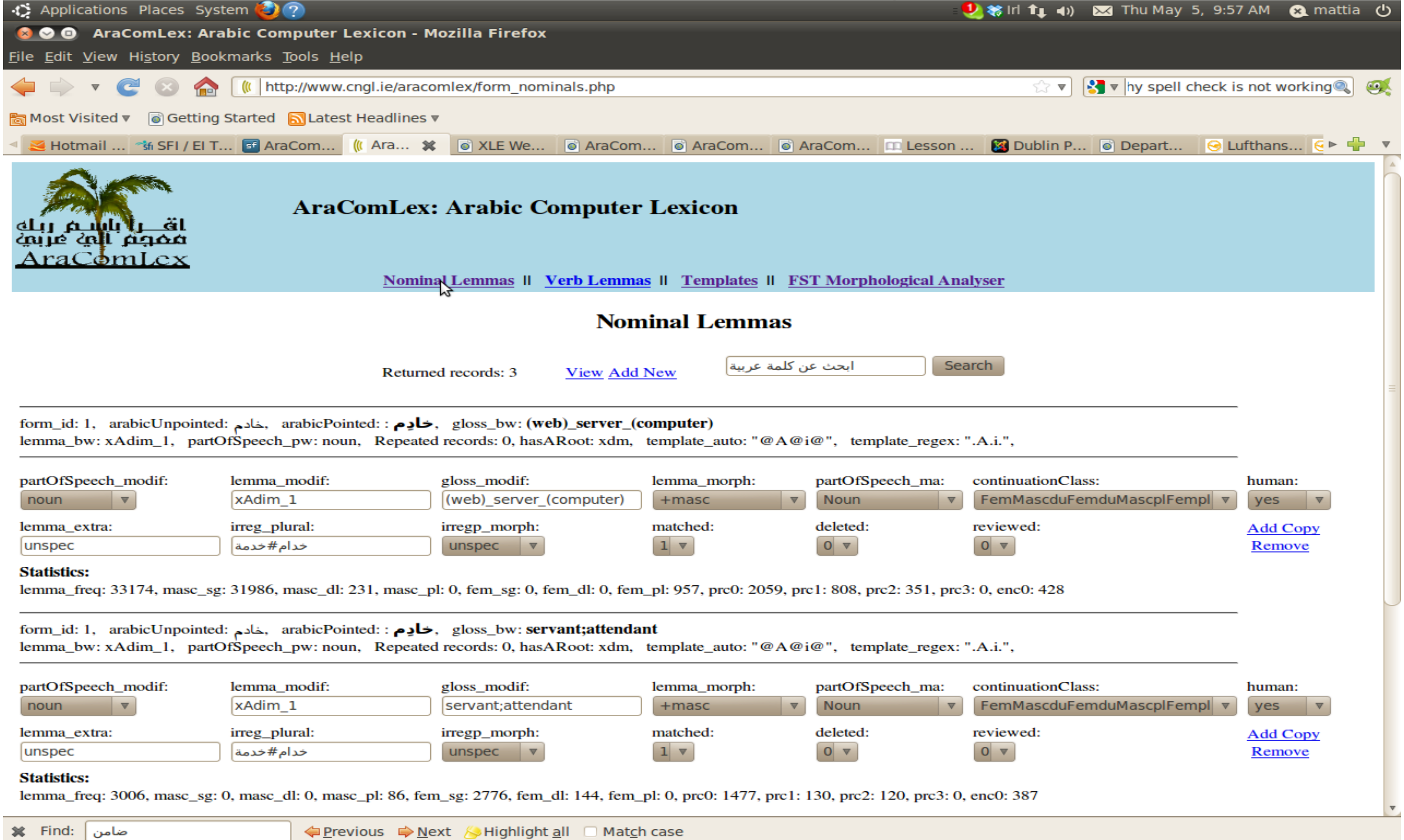
$kr	$akar	PV	thank;give thanks
$kr	$okurIV	thank;give thanks	

3. Lemma-based morphology



Morphological Lexicon - AraComLex

- AraComLex Lexicon Writing Application


Applications Places System
Irl Thu May 5, 9:57 AM mattia
AraComLex: Arabic Computer Lexicon - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.cngl.ie/aracomlex/form_nominals.php
hy spell check is not working
Most Visited ▾ Getting Started Latest Headlines ▾
Hotmail ... SFI / EI T... AraCom... Ara... XLE We... AraCom... AraCom... AraCom... Lesson ... Dublin P... Depart... Lufthans...
اقـرا باسـم ربك
معجم آلي عربي
AraComLex
Find: ضامن
Previous Next Highlight all ☐ Match case

AraComLex: Arabic Computer Lexicon

Nominal Lemmas || Verb Lemmas || Templates || FST Morphological Analyser

Nominal Lemmas

Returned records: 3 View Add New ابحث عن كلمة عربية Search

form_id: 1, arabicUnpointed: خادم, arabicPointed: : **خادِم**, gloss_bw: **(web)_server_(computer)**
lemma_bw: xAdim_1, partOfSpeech_pw: noun, Repeated records: 0, hasARoot: xdm, template_auto: "@A@i@", template_regex: ".A.i.",

partOfSpeech_modif:	lemma_modif:	gloss_modif:	lemma_morph:	partOfSpeech_ma:	continuationClass:	human:
noun ▾	xAdim_1	(web)_server_(computer)	+masc ▾	Noun ▾	FemMascduFemduMascplFempl ▾	yes ▾
lemma_extra:	irreg_plural:	irregp_morph:	matched:	deleted:	reviewed:	Add Copy
unspec	خدام#خدمة	unspec ▾	1 ▾	0 ▾	0 ▾	Remove

Statistics:
lemma_freq: 33174, masc_sg: 31986, masc_dl: 231, masc_pl: 0, fem_sg: 0, fem_dl: 0, fem_pl: 957, prc0: 2059, prc1: 808, prc2: 351, prc3: 0, enc0: 428

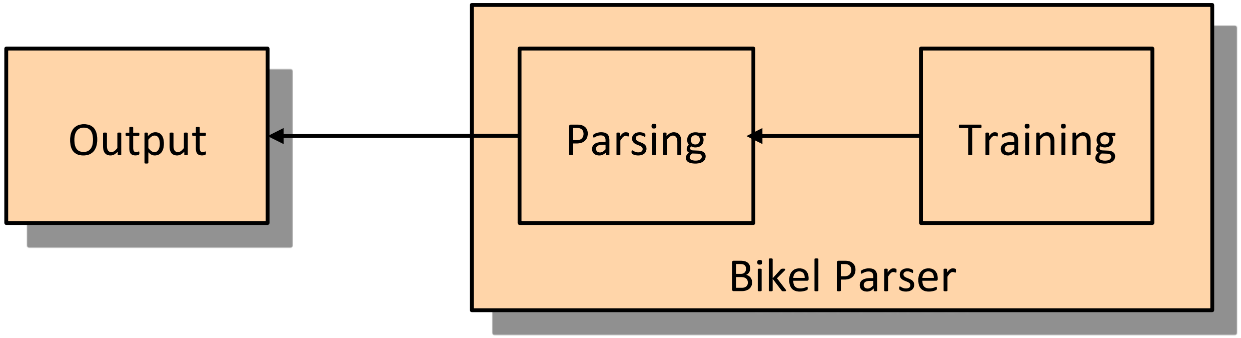
form_id: 1, arabicUnpointed: خادم, arabicPointed: : **خادِم**, gloss_bw: **servant;attendant**
lemma_bw: xAdim_1, partOfSpeech_pw: noun, Repeated records: 0, hasARoot: xdm, template_auto: "@A@i@", template_regex: ".A.i.",

partOfSpeech_modif:	lemma_modif:	gloss_modif:	lemma_morph:	partOfSpeech_ma:	continuationClass:	human:
noun ▾	xAdim_1	servant;attendant	+masc ▾	Noun ▾	FemMascduFemduMascplFempl ▾	yes ▾
lemma_extra:	irreg_plural:	irregp_morph:	matched:	deleted:	reviewed:	Add Copy
unspec	خدام#خدمة	unspec ▾	1 ▾	0 ▾	0 ▾	Remove

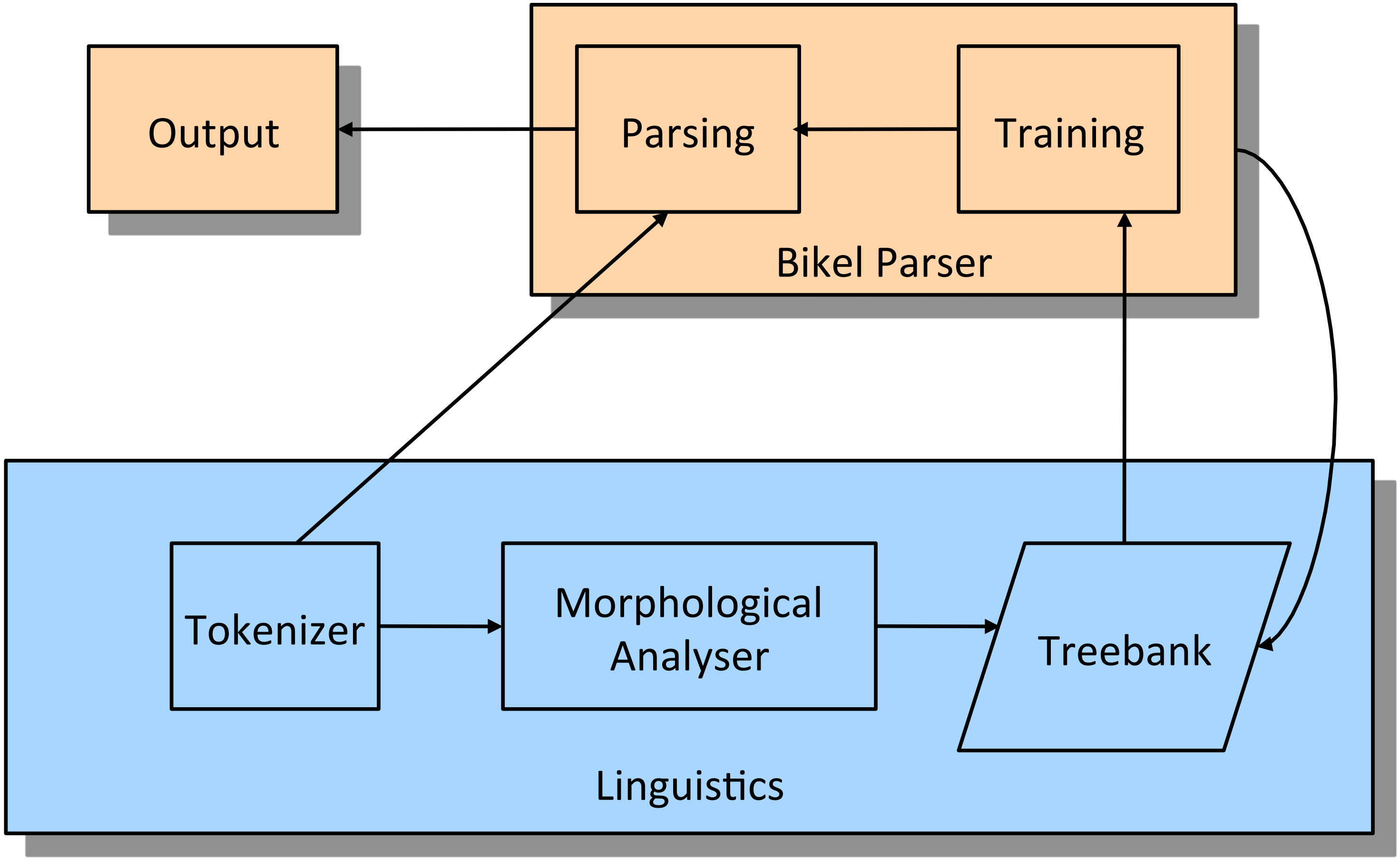
Statistics:
lemma_freq: 3006, masc_sg: 0, masc_dl: 0, masc_pl: 86, fem_sg: 2776, fem_dl: 144, fem_pl: 0, prc0: 1477, prc1: 130, prc2: 120, prc3: 0, enc0: 387

Syntactic Parsing

Algorithms and Data Structure



Algorithms and Data Structure

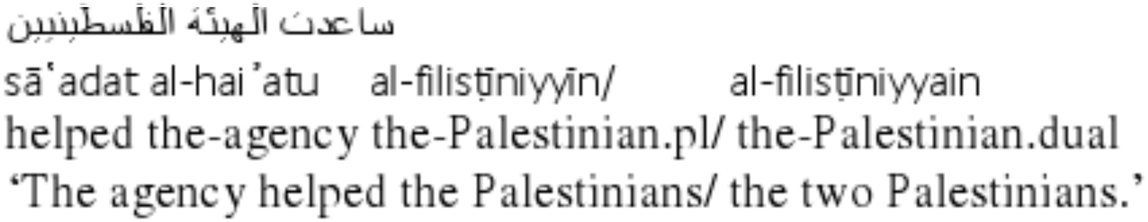


Why Linguistics

- 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



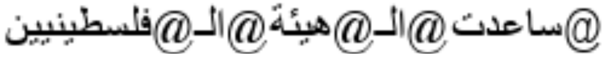
ساعدت الهيئة الفلسطينيين

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

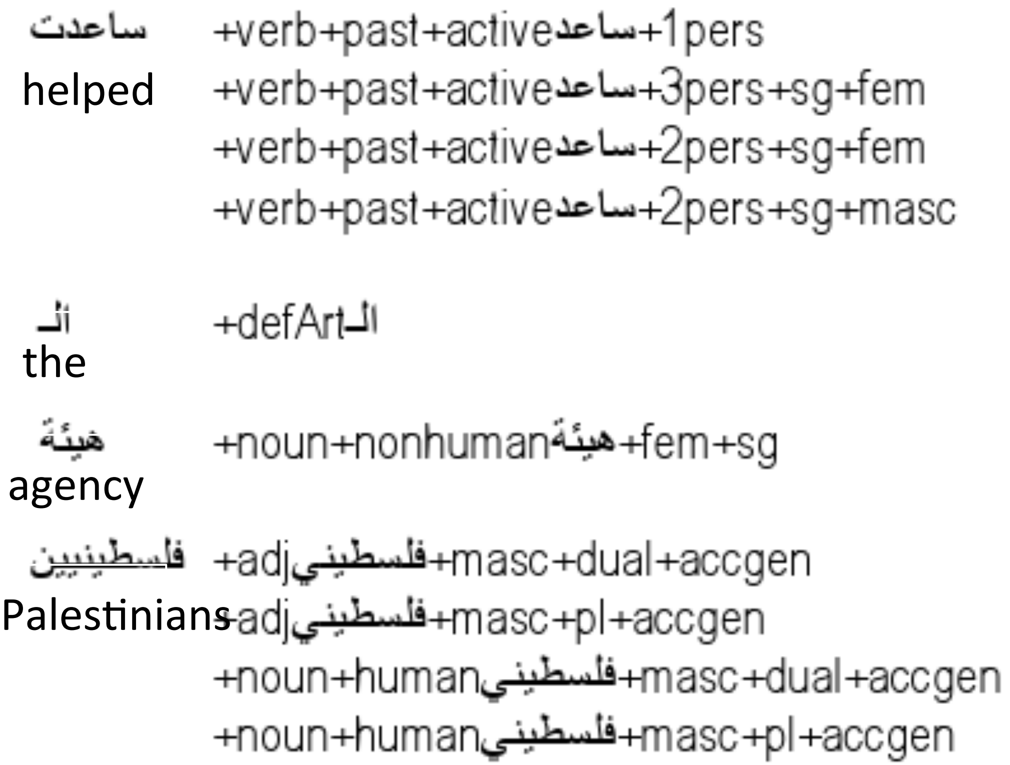


@ساعدت@الـ@هيئة@الـ@فلسطينيين

helped@the@agency@the@Palestinians

Handcrafted Grammar:
A Quick Overview

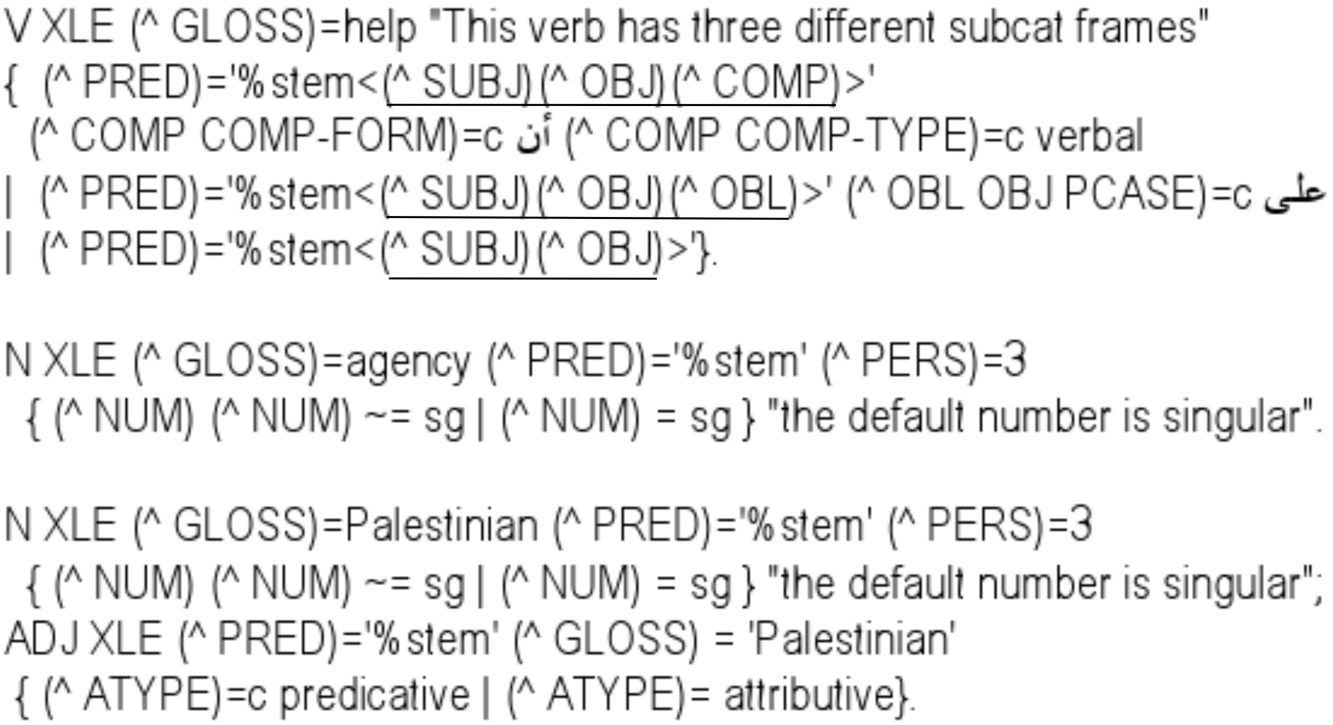
Morphological analysis



ساعدت	+verb+past+activeساعد+1pers
helped	+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	+adjفلسطيني+masc+dual+accgen
Palestinians	+adjفلسطيني+masc+pl+accgen
	+noun+humanفلسطيني+masc+dual+accgen
	+noun+humanفلسطيني+masc+pl+accgen

Handcrafted Grammar: A Quick Overview

Lexicon (Lexical properties/subcategorization frames)أ



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

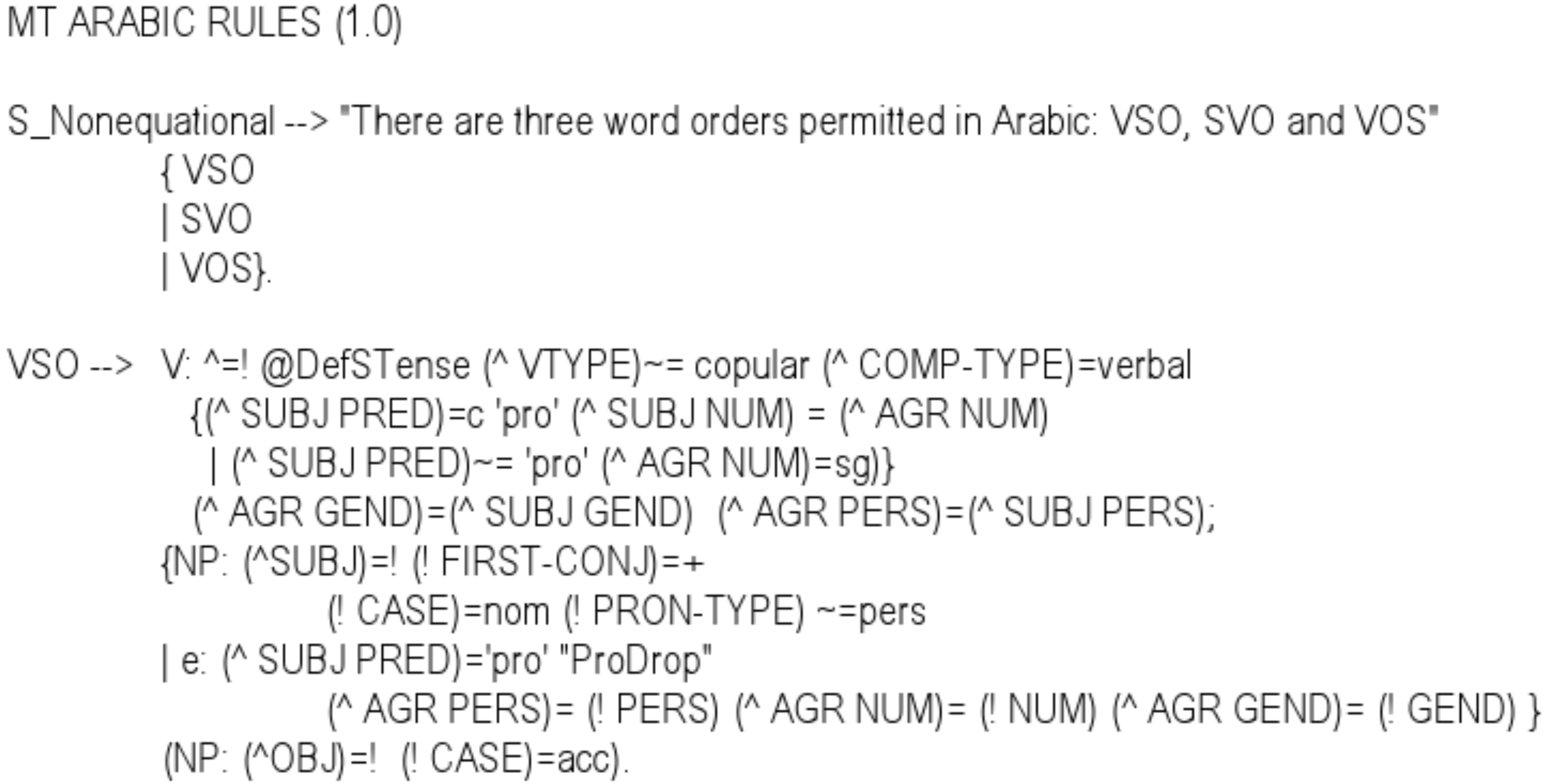
```
N XLE (^ GLOSS)=agency (^ PRED)='%stem' (^ PERS)=3
  { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular".
```

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

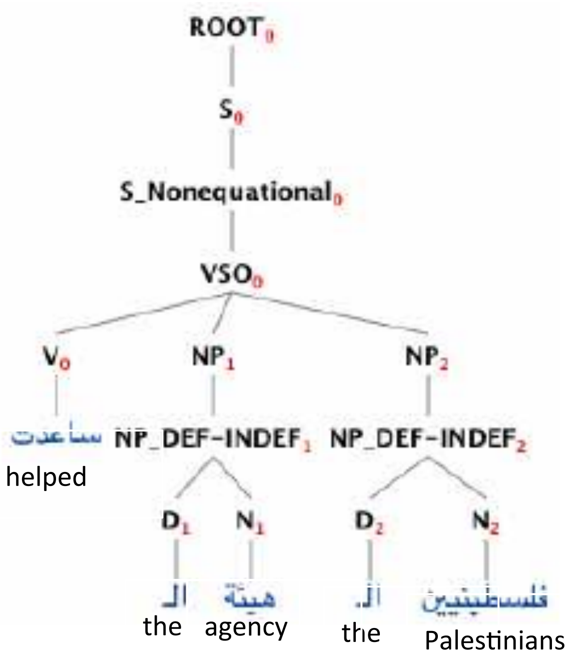
Grammar Rules: PS-rules and functional equations

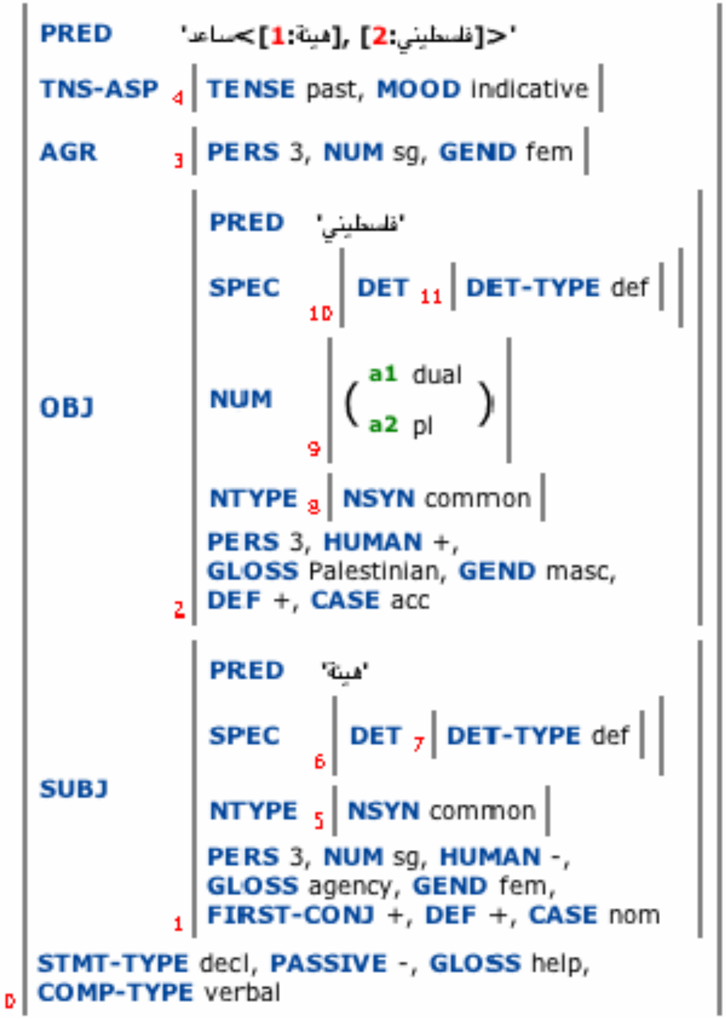


```
MT ARABIC RULES (1.0)

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



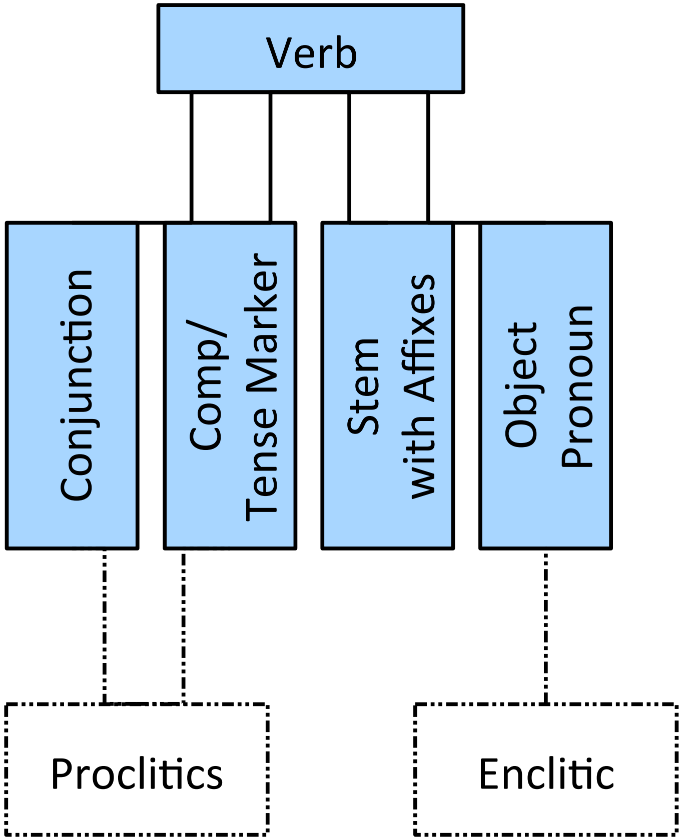


```
PRED        'ساعد<[1:هيئة], [2:فلسطيني]>'
TNS-ASP   4 | TENSE past, MOOD indicative |
AGR       3 | PERS 3, NUM sg, GEND fem |
            | PRED    'فلسطيني'
            | SPEC    | DET 11 | DET-TYPE def | |
            |      10
            | NUM     | ( a1 dual )
OBJ         |         |   a2 pl
            |       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,
0 COMP-TYPE verbal
```

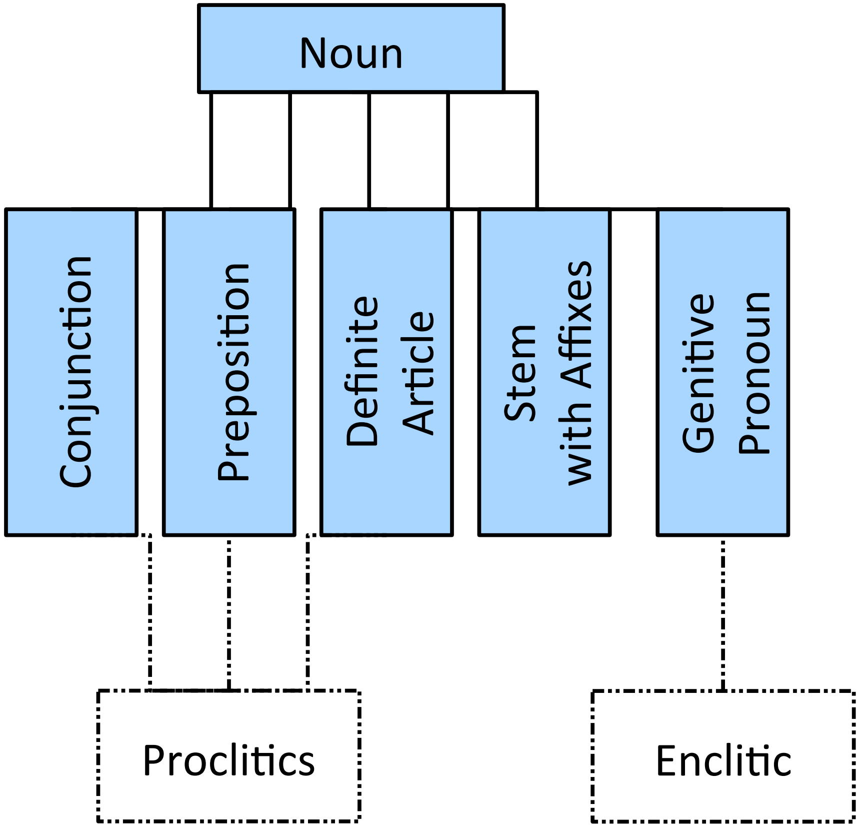
Tokenization

Tokenization in XLE

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

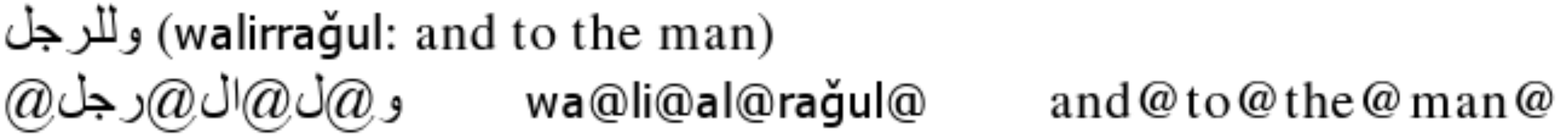


وللرجل
walilrajuli
wa@li@al@rajuli
and@to@the@man



Tokenization in XLE

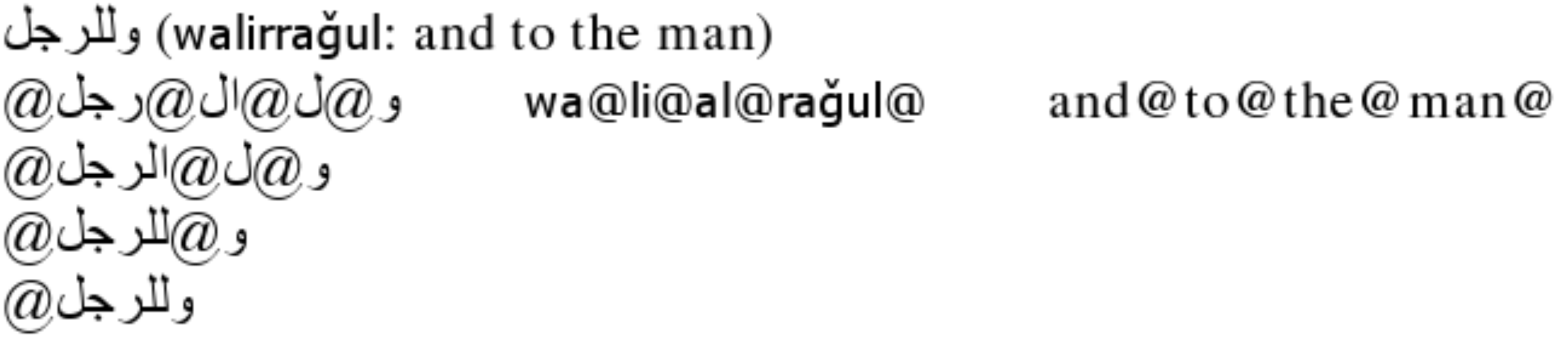
Deterministic Tokenizer



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.)

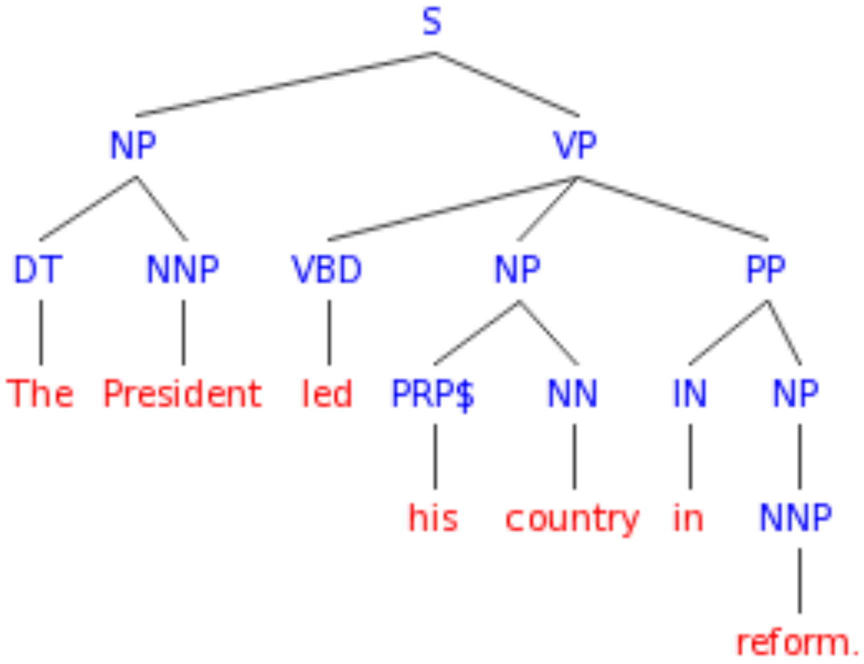
(VBZ has) (RB n't)
(NNP Chicago) (POS 's)

Tokenization in Bikel

- English parser
  - Output:

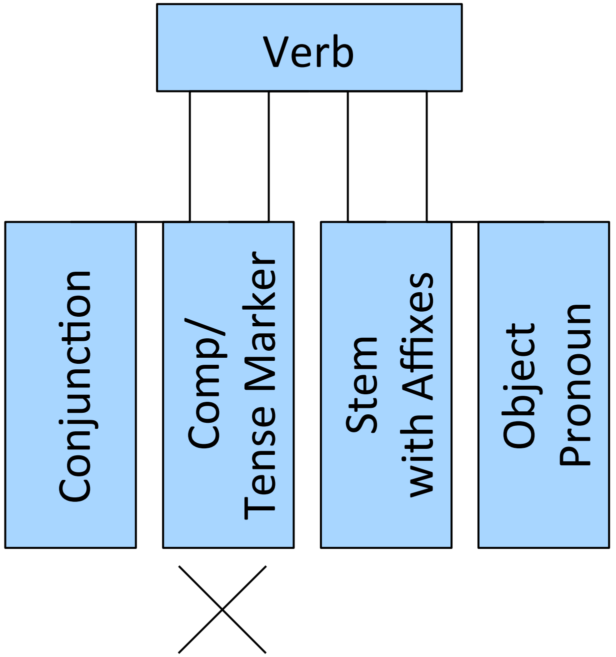
```
(S (NP (DT The) (NNP President)) (VP (VBD led) (NP (PRP$ his) (NN country)) (PP (IN in)
(NP (NNP reform.)))))
```

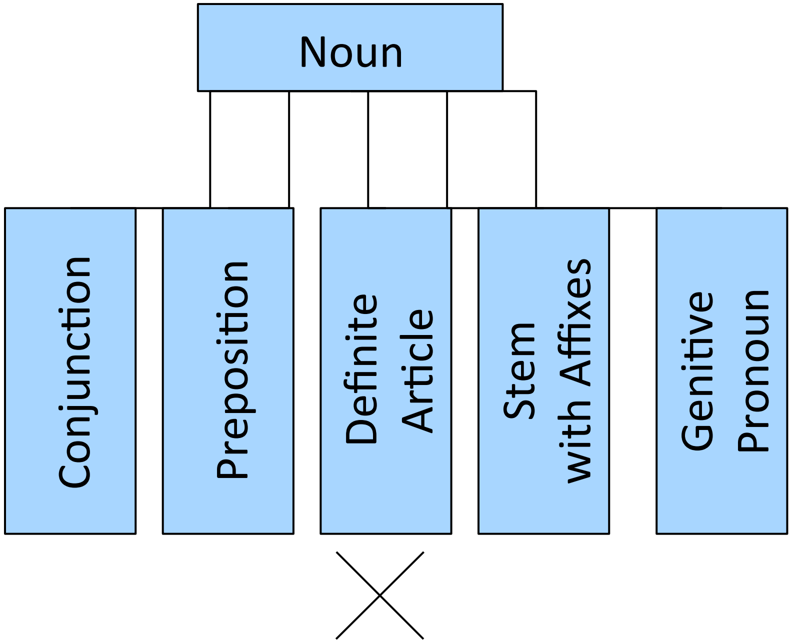
  - Tree



Tokenization in Bikel

- Arabic parser





Tokenization in Bikel

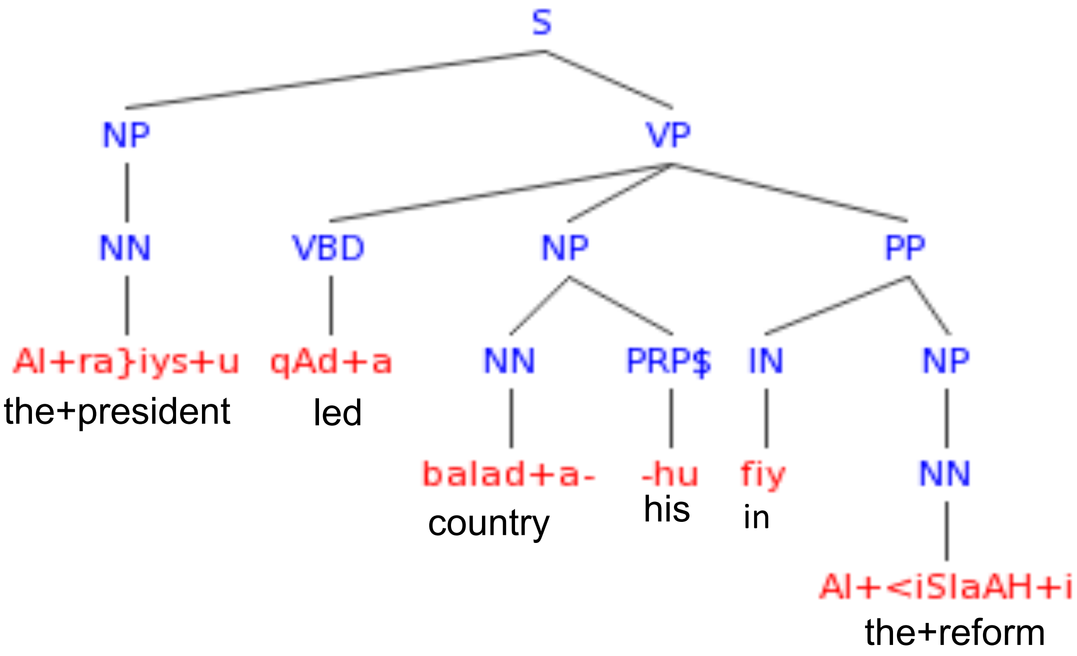
- Arabic parser
  - Input sentence:

    الرئيس قاد بلده في الإِصلاح

    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



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*
    - أبو ظبي – رأس الخيمة – جبل علي
    - أبو فروة – فرس النبي
  - Semi-Fixed Expressions: Lexically, or morphologically flexible
    - *Sweep somebody under the rug/carpet*
    - *Transitional period(s)*
    - مدينة ملاهي \ مدينة ترفيهية
    - قاعدة عسكرية
  - Syntactically-flexible Expressions
    - *to let the cat out of the bag*
    - *The cat was let out of the bag.*
    - دراجة نارية \ دراجة الرجل النارية
    - وضعت الحرب أوزارها \ وضعت أوزارها

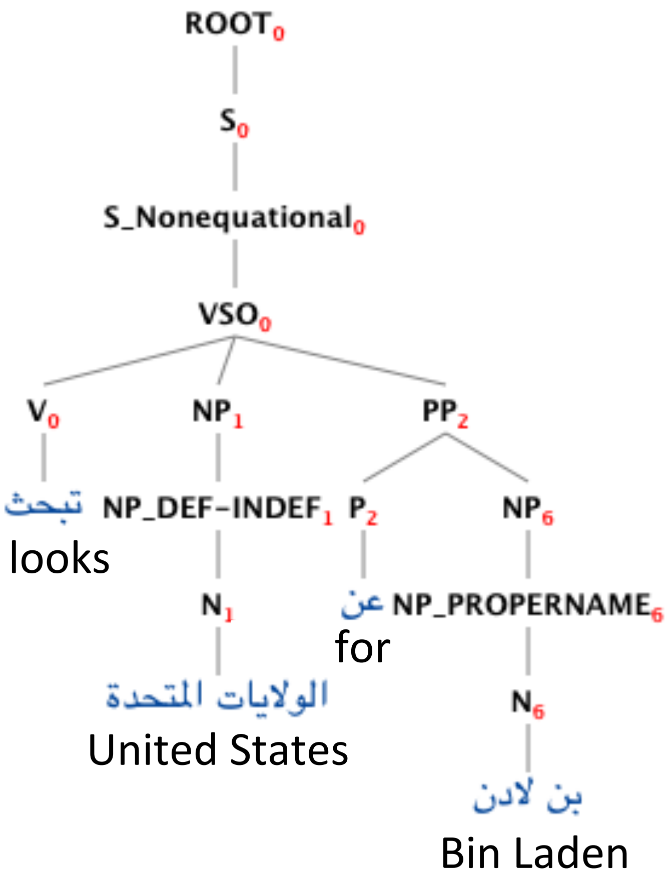
Multiword Expressions

- 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 collected 34,658 MWEs in addition to 45,202 Named Entities

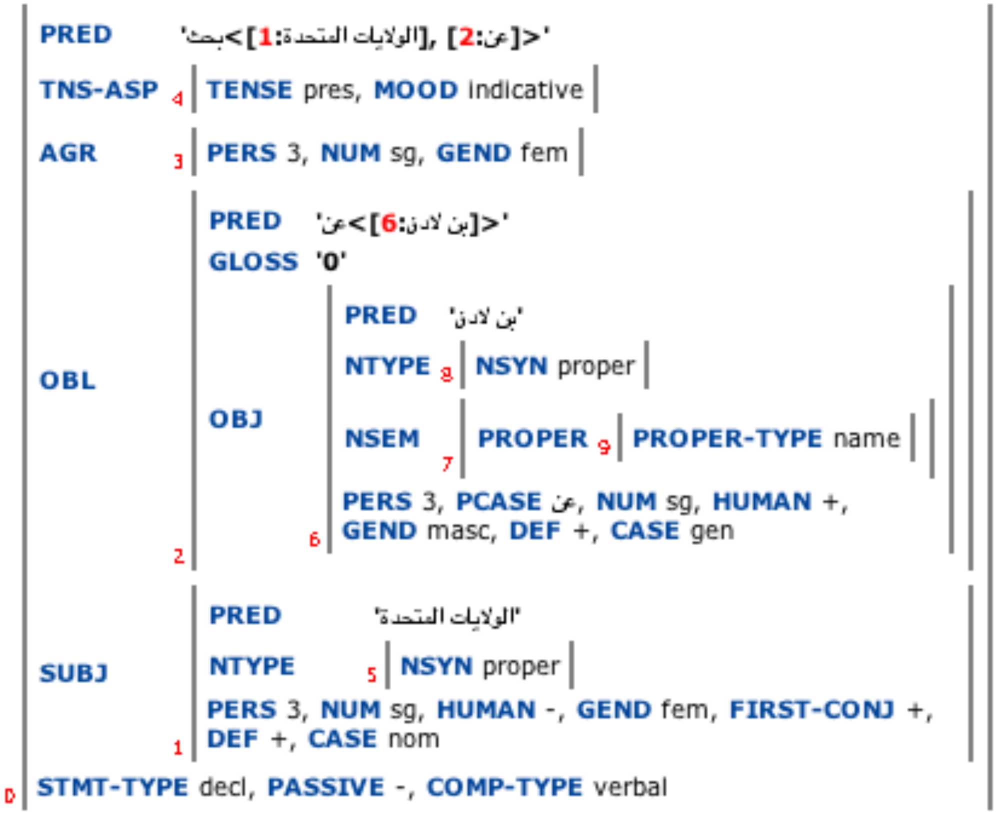
Multiword Expressions in XLE

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]
             [PRED  'عن<[6:بن لادن]>'
             GLOSS '0'
                    [PRED    'بن لادن'
                     NTYPE 8 [NSYN proper]
  OBL               
             OBJ     NSEM    [PROPER 9 [PROPER-TYPE name]]
                           7
                     PERS 3, PCASE عن, NUM sg, HUMAN +,
                   6 GEND masc, DEF +, CASE gen]
           2 ]
             [PRED            'الولايات المتحدة'
  SUBJ       NTYPE          5 [NSYN proper]
             PERS 3, NUM sg, HUMAN -, GEND fem, FIRST-CONJ +,
           1 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 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))
```

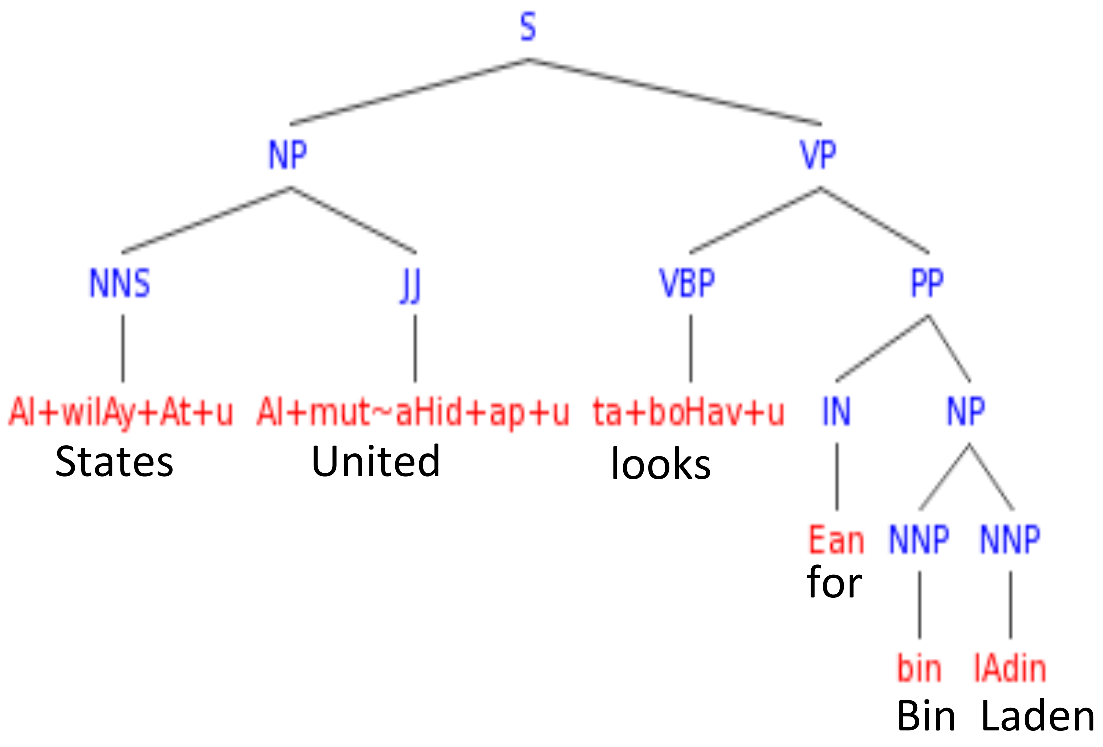
in addition to إضافة إلى

```
(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)))))



Statistical Parsing

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

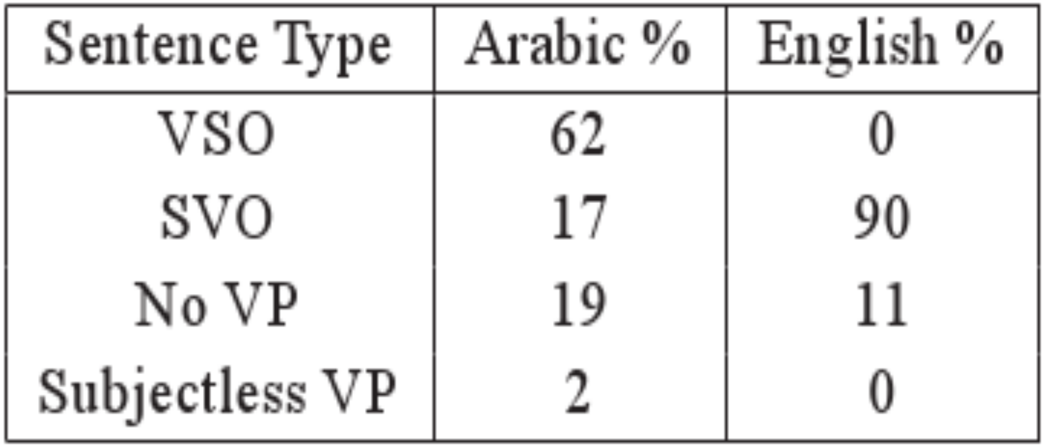
- Why Arabic performs poorly? (Kulick et al. 2006)
  - The ATB tag set is very large and dynamic, this is why they are mapped into 20 PTB tags. The tagset reduction is extreme and important information is lost.
    - Verb
      - » IV3FS+IV+IVSUFF_MOOD:I
      - » IV3MS+IV+IVSUFF_MOOD:J
      - » PV+PVSUFF_SUBJ:3MS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

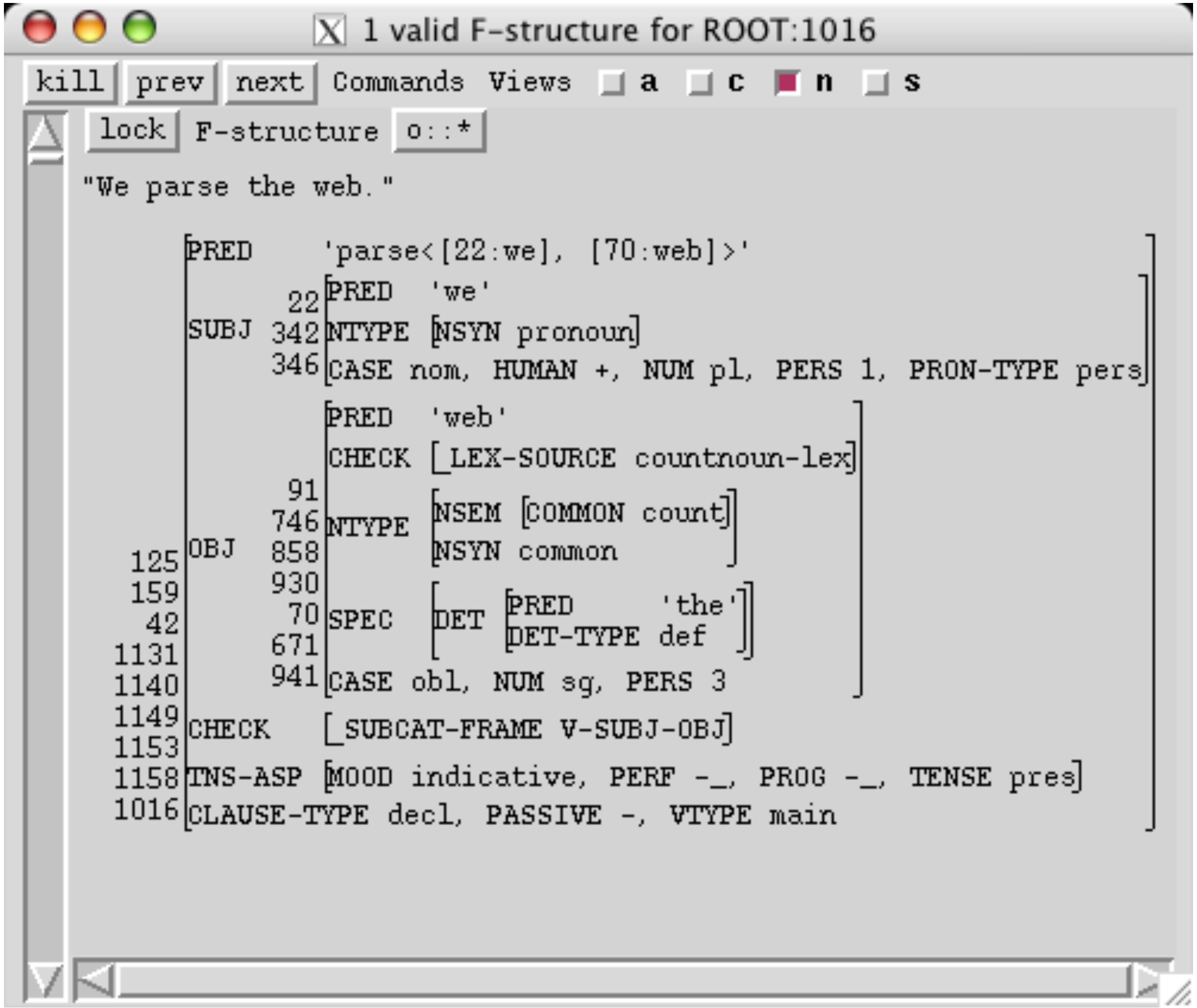
Which is better?

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

(Kaplan et al., 2004)

Which is better?

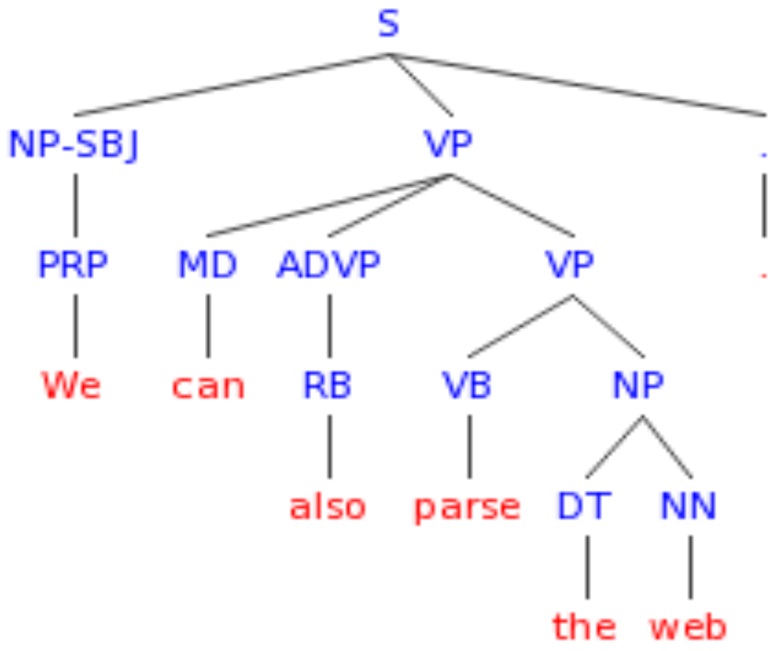
- XLE: “We parse the web.”

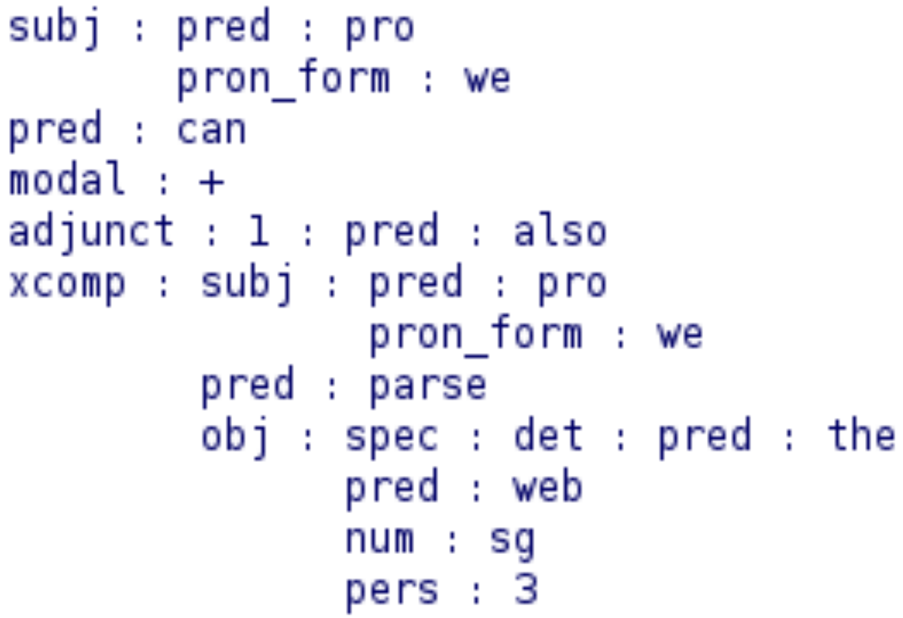


```
"We parse the web."

        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  CHECK      [SUBCAT-FRAME V-SUBJ-OBJ]
  1153
  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?

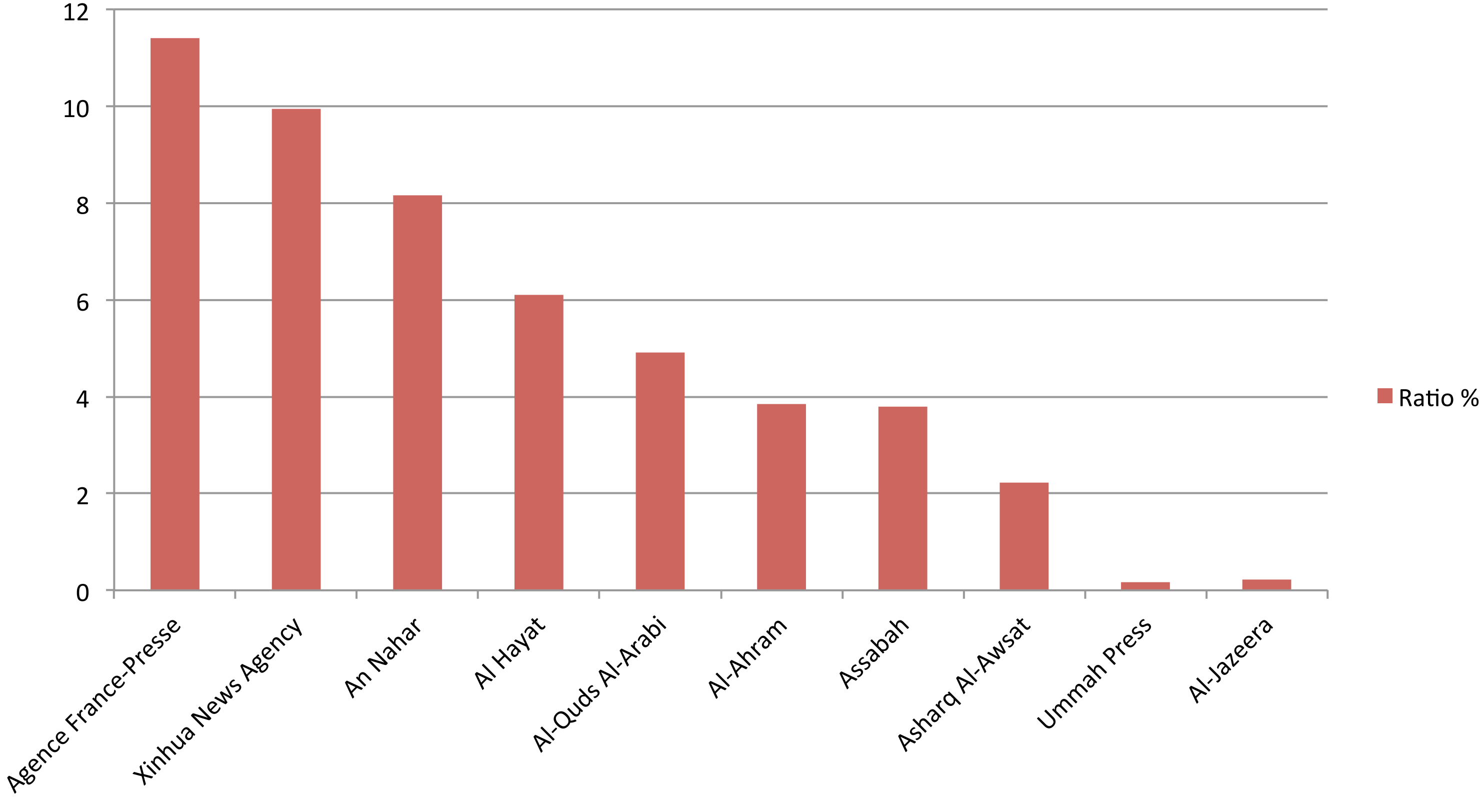
- Statistical grammars are more efficient because:
  - there is a clear separation between the algorithm and the data structure
  - there is a clear division of labour, the linguists fight their battle, and the engineers fight their own battle

Spelling Checking and Correction

Spelling Checking and Correction

How frequent is spelling errors in news web sites?

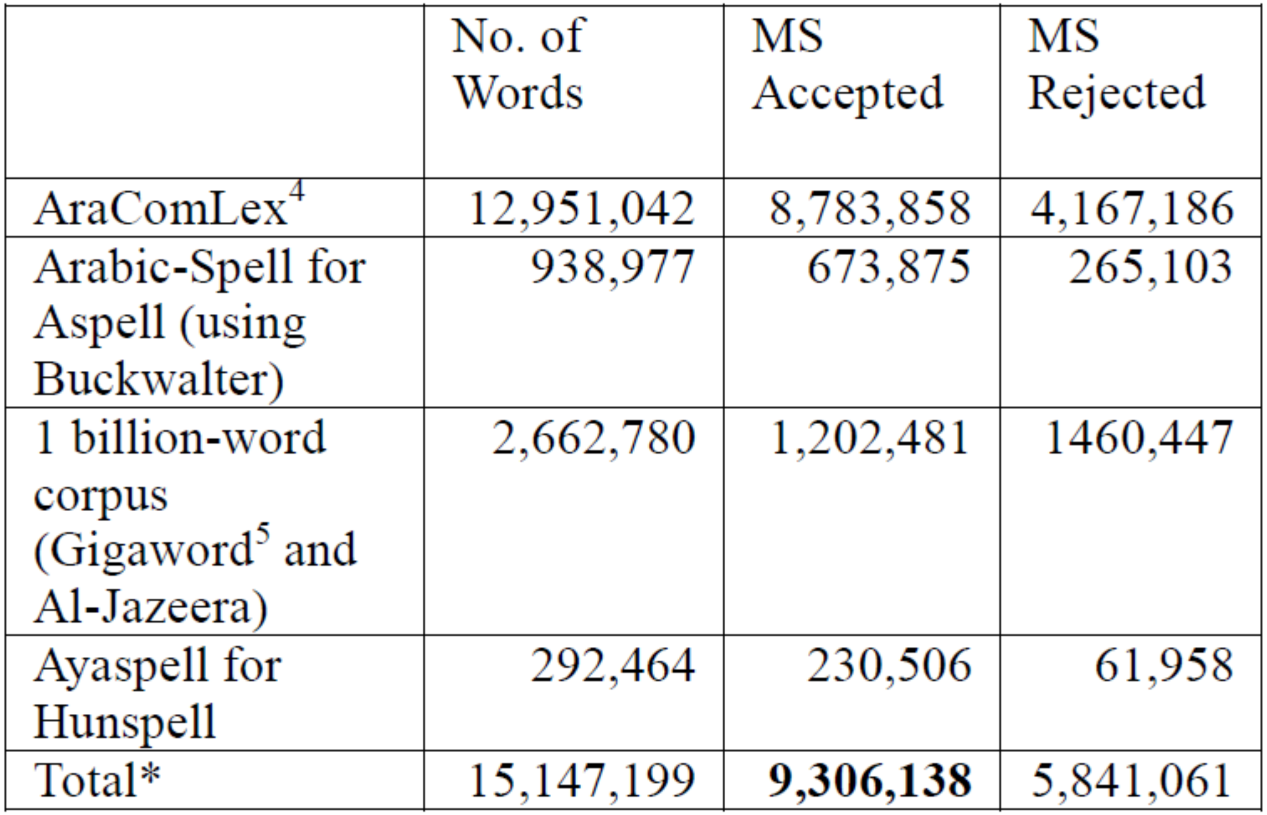
Spelling errors in News Web Sites



Spelling Checking and Correction

Creating a wordlist

English: 708,125, French: 338,989, Polish: 3,024,852

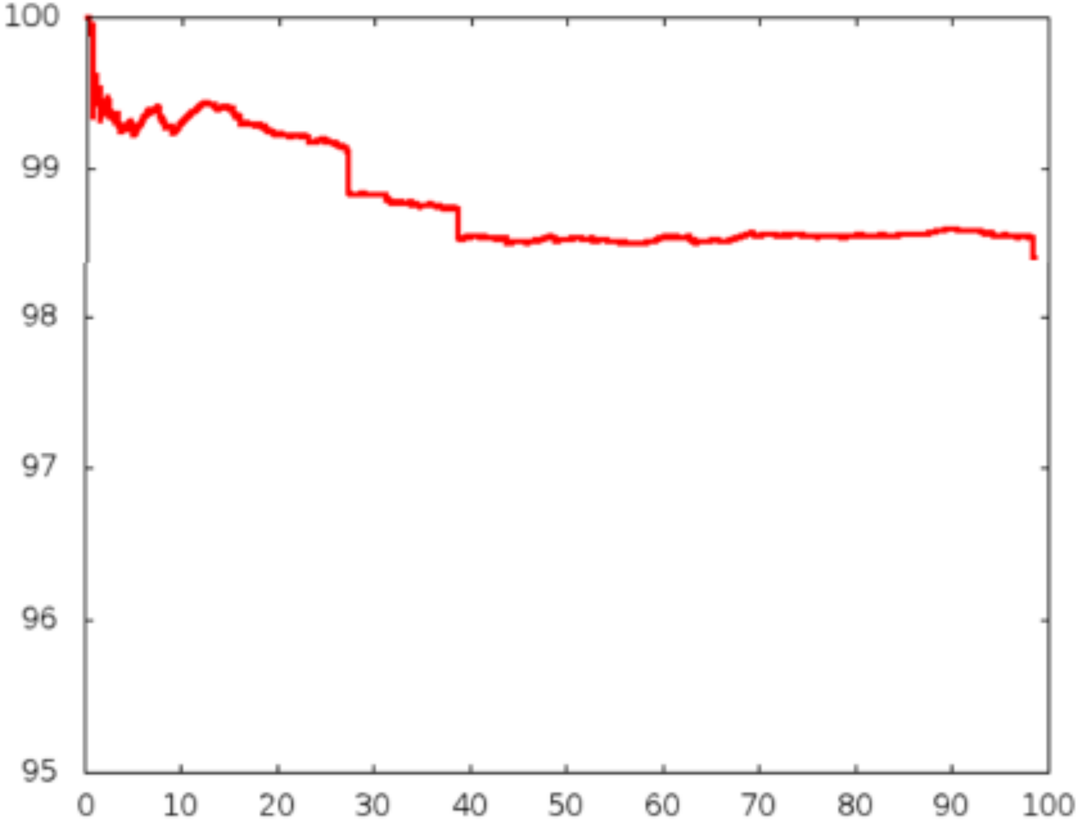


	No. of Words	MS Accepted	MS Rejected
AraComLex[4]	12,951,042	8,783,858	4,167,186
Arabic-Spell for Aspell (using Buckwalter)	938,977	673,875	265,103
1 billion-word corpus (Gigaword[5] and Al-Jazeera)	2,662,780	1,202,481	1460,447
Ayaspell for Hunspell	292,464	230,506	61,958
Total*	15,147,199	**9,306,138**	5,841,061

Spelling Checking and Correction

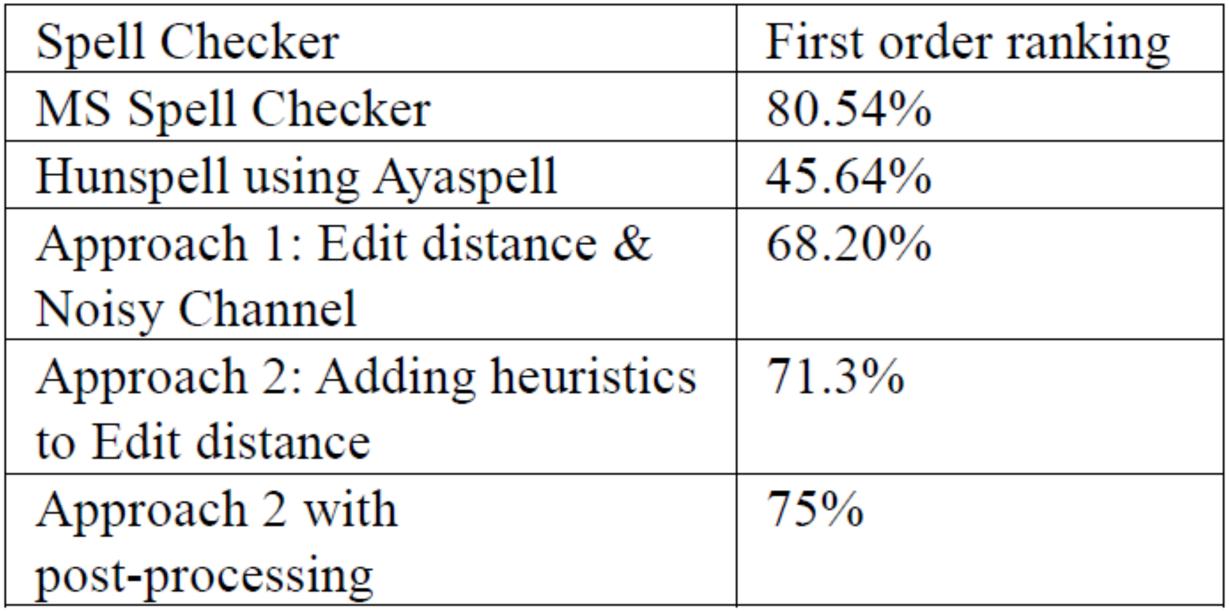
Spelling error detection

1. Matching against a word list 2. Character-based LM

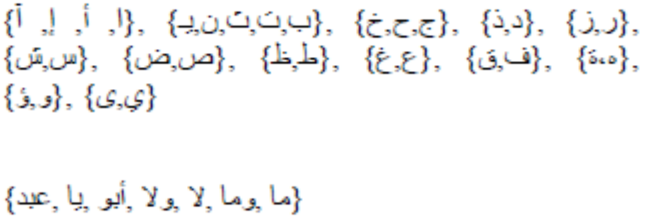


Spelling Checking and Correction

Automatic correction of spelling errors



Spell Checker	First order ranking
MS Spell Checker	80.54%
Hunspell using Ayaspell	45.64%
Approach 1: Edit distance & Noisy Channel	68.20%
Approach 2: Adding heuristics to Edit distance	71.3%
Approach 2 with post-processing	75%

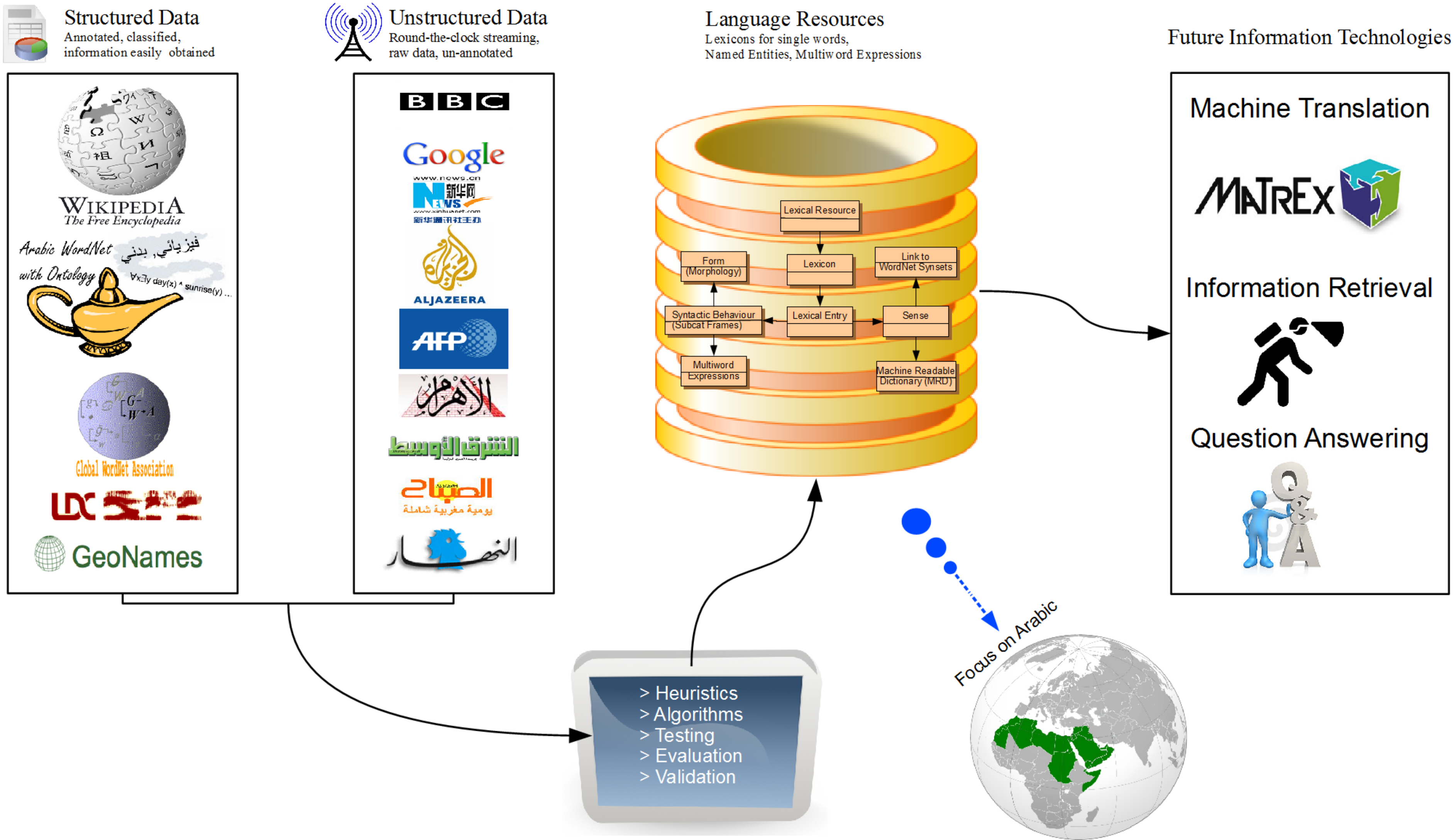


{ا, أ, إ, آ}, {ب,ت,ث,ن,يـ}, {ج,ح,خ}, {د,ذ}, {ر,ز}, {س,ش}, {ص,ض}, {ط,ظ}, {ع,غ}, {ف,ق}, {ه,ة}, {و,ؤ}, {ى,ي}

{ما, وما, لا, ولا, أبو, يا, عبد}

Integration with Applications

Applications of Arabic Language Technologies



Structured Data
Annotated, classified,
information easily obtained

Unstructured Data
Round-the-clock streaming,
raw data, un-annotated

Language Resources
Lexicons for single words,
Named Entities, Multiword Expressions

Future Information Technologies

Machine Translation

Information Retrieval

Question Answering

> Heuristics
> Algorithms
> Testing
> Evaluation
> Validation

Focus on Arabic