Extending the DCU-250 Gold Standard f-structure Bank

H. Béchara
hbechara@computing.dcu.ie

Outline

Outline

Motivation

- Produce an ATB-based LFG gold resource for parsing evaluation similar to DCU's previous work on English, German, Chinese, etc.

- Extend the existing Arabic LFG Gold Standard, from 250 annotated sentences to 500.

- A larger variety of grammatical phenomena
- A more comprehensive reference
- A more general sample for evaluation

Outline

Arabic Grammar

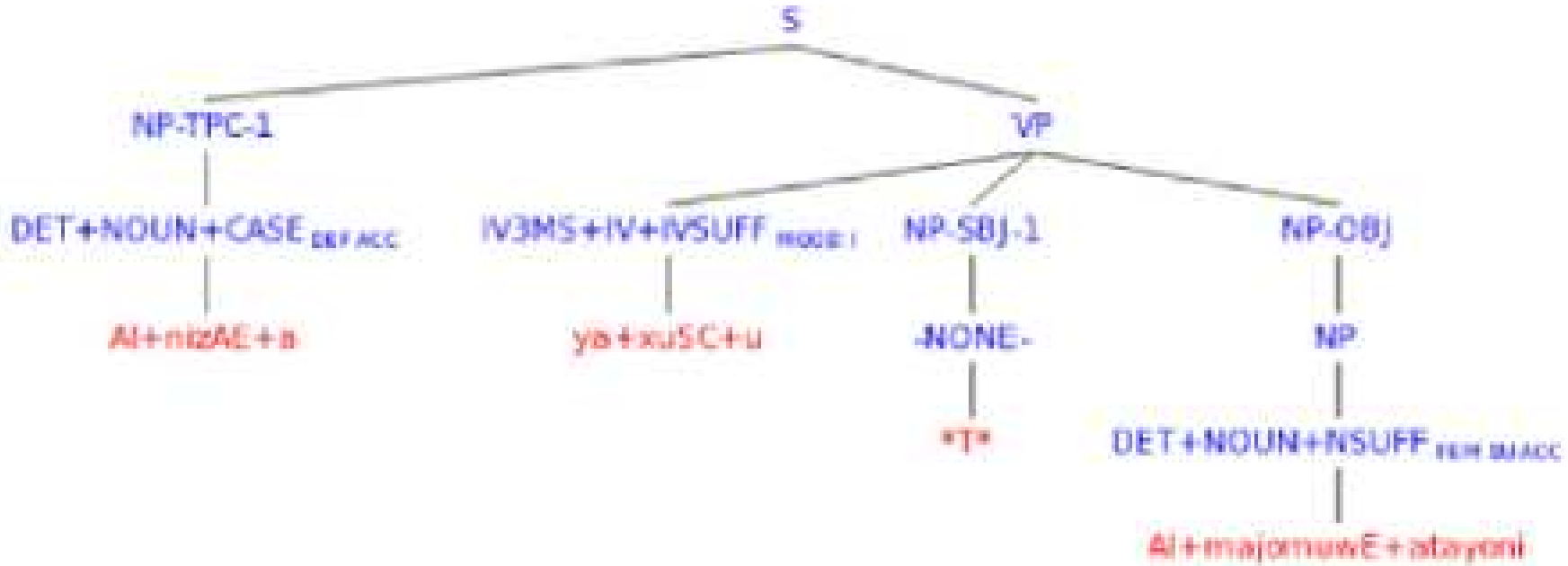
Some Particularities of Arabic Grammar

- Sentences can be very long (longest sentence is 384, average sentence 30)
- Word Order is quite flexible
- Dropping subjects, objects, relative pronouns (pro-drop)
- Word endings can overlap for noun cases

Penn Arabic Treebank (ATB)

- 23,611 parse-annotated sentences in Modern Standard Arabic (Maamouri and Bies 2004)
- Buckwalter Transliteration: Strictly one-to-one transliteration from Arabic to Latin characters (ASCII)
- Part of Speech Tags (Noun, Verb, Prep)
- Phrasal Tags (NP, VP, PP)
- Functional Tags (OBJ, SUBJ, ADJ)

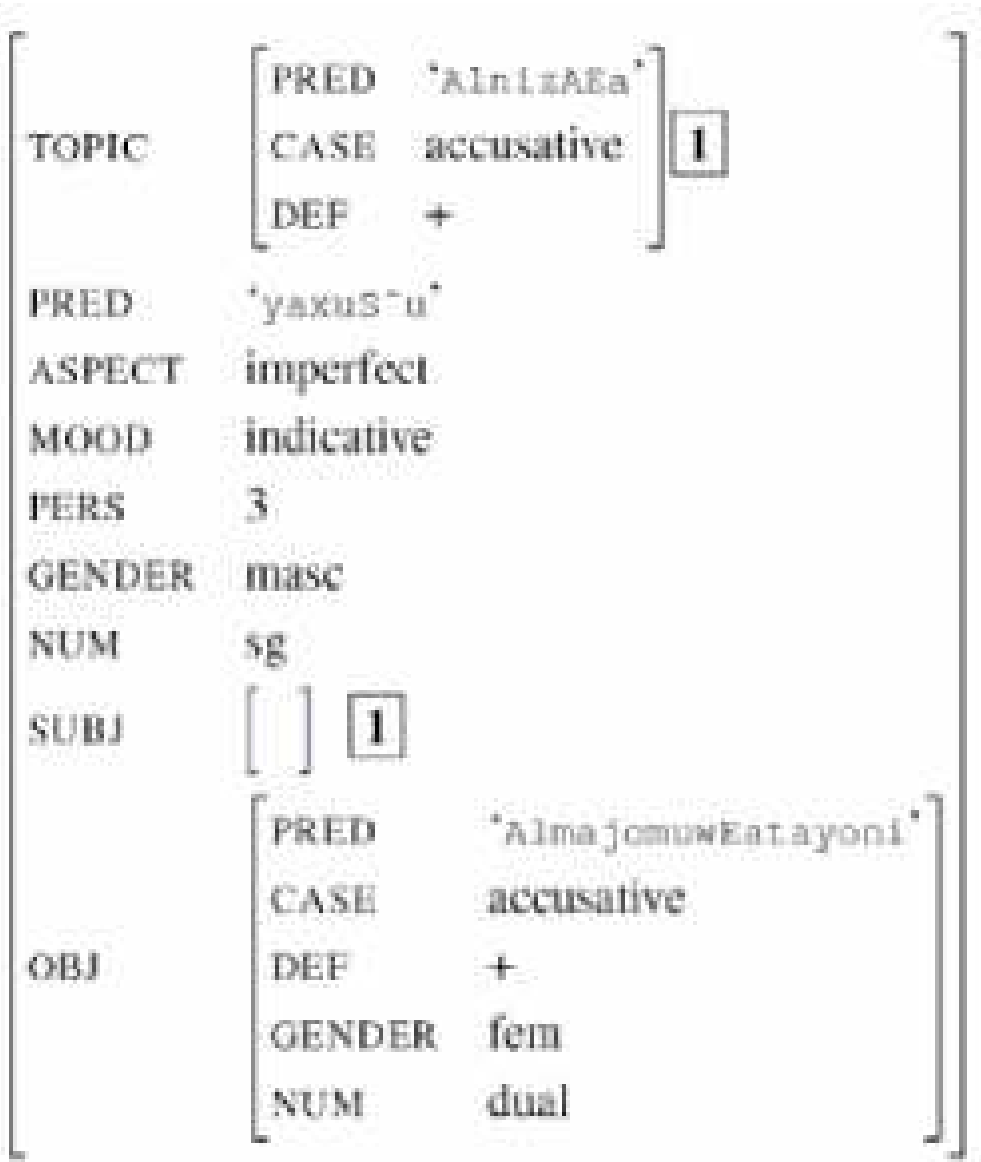
Penn Arabic Treebank (ATB)



Arabic Annotation Algorithm

- The Arabic Annotation Algorithm aims to convert the c-structure provided by the Penn Arabic Treebank into an f-structure.
- It is a recursive process which annotates eah node of a tree with f-structure information used to generate proper f-structures

Arabic Annotation Algorithm



Outline

Methodology

- Random Selection of 250 new sentences from the Penn Arabic Treebank
- Application of the Arabic Annotation Algorithm
- Combination of old and new Sets for Full Evaluation.

Methodology

Correction Method

- Surface Improvements (manual, semi-automatic and automatic)
  - Noun Cases
  - Functional Tags
  - Improper Constructions
- Annotation Improvements (manual, semi-automatic and automatic)
  - Adjunct Tags
  - Pro-Drop
  - Resolving Clashes

Surface Changes

Noun Case Ambiguity

Arabic has three noun cases which are generally differentiated morphologically based on word endings.

Generally:

Nominative (NOM): -u

Accusative (ACC): -a

Genitive (GEN): - i

However, there are particular instances where both the genitive and accusative endings are the same.

Case	Female Plurals	Male Plurals	Duals
Nominative	-AtN	-uwon	-An
Genitive	-AtK	-iyon	-ayon
Accusative	-AtK	-iyon	-ayon

The morphological analyser assigns these words the tag: ACCGEN

This Tag occurs 162 times in the 500 sentences.

Surface Changes

Noun Case Ambiguity (Automatic)

Habash and Rambow, 2007: Determining Case in Arabic: Learning Complex Linguistic Behaviour Requires Complex Linguistic Features.

We explore the local subtree's current node, mother node, and sister nodes.

- ACC: ADJ, CONJ, OBJ, TPC, PRD of subordinating conjunction
- GEN: ADJ, CONJ, PP, NP-adjuncts (Idafa construction)

Surface Changes

Missing Functional Tags (Semi-automatic)

When the word is unreadable, the analyser fails to assign a part of speech tag.

A word becomes unreadable when it is improperly alliterated, usually due to missing vowels.

Examples:

- fsTynyA
- xTAb
- AstrAtyjyA

The morphological analyser assigns these words the tag:
NO_FUNC This Tag occurs 82 times in the 500 sentences.

Surface Improvements

Improper Sentence Construction (Manual)
Problems that arise from the Parser's confusion and/or tokenisation.
Example: fa+sa+nalEab+u (then+will+we+play)
Example: Helping the elderly **and** the poor **and** the handicapped **and** feeding the hungry.

Annotation Improvements

- Specifying Adjuncts
  - Appositions
  - Adjective Types: attributive, predicative.
  - Adverbs
  - Prepositional Phrases: temporal, directional, locative, etc.
  - Titles: Lexicalising 52 Titles (Mr, Miss, Dr, Sir, Prince, Queen, President, etc)

Annotation Improvements

- Appositions (ATB Guidelines)
  Names in apposition are an exception to the 'all adjuncts on same level' rule: an extra NP level is added in the tree
  (NP (NP (NP head noun)
  (XP any adjunct))
  (NP appositive name)

Annotation Improvements

- Demonstrative Pronouns
  h‘*ihi + Al+tagoyiyrAt+i + tata$Abak+u + *Akirat+u+hA + fiy + h‘*A + Al+faDA’+i + Al+HaDAriy +i
  these + the+changes + be interwoven + remembering+its + in + this + the+space + the+cultural
  Remembering these changes is interwoven with this cultural space
- NP modified by quantificational NP
  akalot+u + Al+dajAjap+a + niSofa+hA
  ate + the+chicken + half +its
  I ate half of the chicken
- NP modified by numerical NP
  qaraot+u + Al+kitAb+a + Ei$rina + SafoHap+F + min+hu
  read + the+book + twenty + page from+it
  I read twenty pages of the book

Annotation Improvements

- Specifying Adjuncts
  - Appositions
  - Adjective Types: attributive, predicative.
  - Adverbs
  - Prepositional Phrases: temporal, directional, locative, etc.
  - Titles: Lexicalising 52 Titles (Mr, Miss, Dr, Sir, Prince, Queen, President, etc)

Annotation Improvements

- Resolving Clashes
  - A problem of Heads: Predicates preceding subjects in nominal sentences.
  - A problem of Traces: Phonetically Empty WHNP
  - A problem of Subjects: Every Sentence needs a subject.
- Resolving Traces: Passive constructions

  (S (VP *uhila
  (NP-SBJ-1 Aljumhuwru)
  (NP-OBJ-1 *)))

  *uhil+a + Al+jumohuwr+u

  shocked + the+audience

  The audience was shocked
- Pro-drop

Outline

Interannotator Agreement

Calculating Agreement

An evaluation set of 50 sentences including all the problems outlined earlier and annotated using the Arabic Annotation Algorithm has been selected.

The automatic annotations were corrected by two separate annotators and agreement was calculated based on Artstein and Poesio's coefficients for Pi, S, and Kappa.

Calculating Agreement

- S: All Categories are equally likely (Bennett, Alpert, and Goldstein 1954)
- π: Random assignment of categories to items is governed by the distribution of items among categories in the actual world. (Scott 1955)
- κ: If coders were operating by chance alone, we would get a separate distribution for each coder. (Cohen 1950)

Interannotator Agreement

Results

- Agreement for the Evaluation Set

S	0.98608303
Pi	0.9843478
Kappa	0.98124266

- Agreement for Specific Cases

Traces	0.866666
Noun Case	1.0

Outline

Conclusion

Summing Up

- Selected 250 new parsed sentences from the Penn Arabic Treebank
- Applied the Arabic Annotation Algorithm to the 250 new sentences
- Merged the 250 with the existing gold standard
- Isolated clashes and particularities that the Annotation Algorithm missed
- Improved the Gold Standard and the Annotation Algorithm

Future Work

- Solve coordination instances with no apparent conjunction or punctuation.
- Standardising how to deal with numbers.
- Extending the Gold Standard even further