The Penn Arabic Treebank

Mohamed Maamouri, Ann Bies, Seth Kulick
{maamouri,bies,skulick@ldc.upenn.edu}
Why an Arabic Treebank at Penn?
Current ATB genres and volumes available

10 years of ATB → nearly 2 million words treebanked

<table>
<thead>
<tr>
<th>Genre</th>
<th>Tree tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newswire Text</td>
<td>750K</td>
</tr>
<tr>
<td>Broadcast News</td>
<td>530K</td>
</tr>
<tr>
<td>Broadcast Conversation</td>
<td>200K</td>
</tr>
<tr>
<td>Web Text</td>
<td>250K</td>
</tr>
<tr>
<td>Dialectal Broadcast Conversation</td>
<td>150K</td>
</tr>
</tbody>
</table>
Parallel English-Arabic Treebanks available

English Treebank parallel to ATB → over 1.6 million words

<table>
<thead>
<tr>
<th>Genre</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newswire Text</td>
<td>553K</td>
</tr>
<tr>
<td>Broadcast News</td>
<td>500K</td>
</tr>
<tr>
<td>Broadcast Conversation</td>
<td>150K</td>
</tr>
<tr>
<td>Web Text</td>
<td>157K</td>
</tr>
<tr>
<td>Dialectal Broadcast Conversation</td>
<td>250K</td>
</tr>
</tbody>
</table>
What is a Treebank?

- A bank of syntactic trees
- Running text annotated for syntactic structure, and tokens annotated for POS/morphological information
- Everything in the text must be annotated (unfortunately, you can’t leave out the hard stuff!)
- Need a version of “syntax” that will allow this kind of annotation
- And that will adapt to multiple languages
Goals of Treebanking

- Representing useful linguistic structure in an accessible way
  - Consistent annotation
  - Searchable trees
  - “Correct” linguistic analysis if possible, but at least consistent and searchable
  - Annotation useful to both linguistic and NLP communities
  - Empirical methods providing portability to new languages
  - Structures that can be used as the base for additional annotation and analysis (PropBank or co-reference, for example)
Major characteristics of Arabic Treebank syntactic annotation

SAME AS PENN ENGLISH TREEBANK

1. A sentence is defined as including a subject and a predicate (which may be a verb phrase with a VP-internal subject – thus, VP is often the only child node of an S, if nothing precedes the verb)
2. Node (bracket) labels are syntactic (S, NP, VP, ADJP, etc.)
3. "Dashtags" represent semantic function (-SBJ subject, -OBJ object, -ADV adverbial, -TMP temporal, -PRD predicate, etc.). Dashtags are used only if they are relevant, not on every node
4. Coordination is done as adjunction (Z (Z ) and (Z )); coordination has the same structure at all phrase levels
5. The argument/adjunct distinction is shown via dashtags within VP, and via structure within NP (arguments are sisters of the head noun; adjuncts and all PPs are adjoined to the NP)
6. Same empty categories (representing the same syntactic phenomena)
7. Overall constituency structure and relationships
Major characteristics of Arabic Treebank syntactic annotation

DIFFERENT FROM PENN ENGLISH TREEBANK

1. Trace indices (equivalence class on the node labels only rather than chaining to the empty category markers – same as current LDC English Treebanks, unlike WSJ)

2. Arabic script (bi-directionality and transliteration)

3. New annotation tools necessary to accommodate Arabic script and morphological analysis (now use similar tool for LDC English TB also)

4. Pre-terminal labels are morphological and complex (unless using a reduced tagset)

5. NP-OBJ used to explicitly mark the NP objects of transitive verbs

6. Mostly head-first (adjectives generally follow nouns, e.g.)

7. Arabic subjects are analyzed as VP-internal (following the verb)

8. Pro-drop subjects and topicalized subjects (occur frequently in ATB)

9. No present tense copula (equational sentences)

10. Only one (two) auxiliary verbs, no modals
“POS” tagset differs from English PTB Set

- Compound tags
  - Tags of source tokens include core POS, affixes, clitics
  - Delimited by “+”

- Core parts of speech
  - Differ from both English tagset and Arabic traditional grammar

- Morphological information in addition to POS
  - Richer morphological information for inflected categories (person, gender, etc.)

- Mapping table reduction for convenience
  - Number of unique tags is high
  - Map down following linguistic categories
  - Many possible such mappings, one produced at LDC is included with data publications
What are some of the characteristics of the Arabic language that will affect the Arabic Treebank?

- Null present tense copula
- Pro-drop
- Clitics
- Head-initial, in large part
the question is simple
Empty * subjects – Arabic pro-drop

* used for pro-drop subjects in Arabic (never co-indexed)

(S (VP yu+Tamo)in+u يُطمِّمُنُّ
 (NP-SBJ *)
 (NP-OBJ (NP Al+iAji}+iyna اللاجِئِينَ
 (PP-LOC fiy في
 (NP brAzafiyIلَبَرَازَفِيَلّ


(he) reassures the refugees in Brazzaville
Clitics

Clitics separated for treebanking to reflect syntactic function:

(S wa-
  (VP (PRT -sa-
    -tu+$Ahid+uwna-
    (NP-SBJ *))
    (NP-OBJ –hA *))
  ),

and you will observe her

CLUNCH, December 6, 2011
Parallel English and Arabic trees

<table>
<thead>
<tr>
<th>English word order</th>
<th>Arabic word order</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S (NP-SBJ The boy) (VP fed (NP the cat) ) .)</td>
<td>(S (VP fed (NP-SBJ the boy) (NP-OBJ the cat) ) .)</td>
</tr>
</tbody>
</table>
human rights exist within our concern
English compound nouns vs. Arabic complement nouns

- English (NP human rights) is flat
- Arabic is not – “human” is a noun complement of “rights”
  
  (NP-TPC-1 Huquwq+u حُقُوقُ
  
  (NP Al+<inosAn+i الإِنْسَانُ ))

human rights الإِنْسَانُ حُقُوقُ
Modifiers with either noun

(NP (NP رَئِيَّسُ president
     (NP مَدِينَةٌ city)
     (ADJP جَمِيعَةٌ beautiful)))

A beautiful president of a city

(NP رَئِيَّسُ president
     (NP مَدِينَةٌ جَمِيعَةٌ beautiful city))

A president of a beautiful city

CLUNCH, December 6, 2011
MSA vs. Dialectal Arabic

- Diglossia in all Arabic speaking countries
- Modern Standard Arabic (MSA) = nobody’s native or first language
- MSA is mainly the language of written discourse, used in formal communication both written and oral with a well-defined range of stylistic registers.
- Distinct dialects spoken, accommodation in speech between dialects (especially mutually unintelligible dialects)
- Code switching even in highly monitored speech, including broadcast news and broadcast conversation
Diacritics (representing short vowels) not written in most texts

Much of the morphology is in the short vowels, both derivational (noun vs. verb) and inflectional (nominative vs. accusative case)

Lots of ambiguity with most tokens/words/strings of letters, which leads to annotation difficulties if the ambiguity is not resolved
"علم"  
- علم  ‘science, learning’  
- علم’ ‘flag’  
- علم  3rd P. Masc. Sing. Perf. V. (MSA V. I) ‘he learned/knew’  
- علم  3rd P. Sing. Pass. V. (MSA V. I) ‘it/he was learned’  
- علمَ Intensifying, Caus. V. (MSA V. II) ‘he taught’  
- علمَ Causative V. Pass (MSA V. II) ‘he was taught’  
- علمٌ/علمَ NOM Noun -- Definite and Indefinite  
- علمَ ACCU Noun + Definite)  
- علم/علمِ GEN Noun + Definite and Indefinite).
A single string in the text (bAsm, for example) can mean many different things, depending on what the short vowel morphology is. Some strings can be ambiguous among 120 or more possible solutions!

**INPUT STRING:** باسم

**SOLUTION 1:** bAsim  
LEMMNA ID: bAsim_1  
POS: bAsim/NOUN_PROP  
GLOSS: Basem/Basim

**SOLUTION 9:** biAisomi  
LEMMNA ID: {isom_1  
POS: bi/PREP+{isom/NOUN+i/CASE_DEF_GEN  
GLOSS: by/with + name + [def.gen.]
The authorities refused to give the escaping prince a diplomatic passport.
The stages of the annotation process are as follows:

1. The plain Arabic text is acquired from the source text.
2. The text is run through the automatic morphological analyzer, and the initial lexicon possibilities are provided.
3. The POS/morphological annotator’s choice and selection leads to the fully vocalized form, including case endings, etc.
4. Clitics with independent syntactic function are automatically separated.
5. The text and POS information are run through the automatic parser, and the initial parse is provided [Dan Bikel’s Collins-type parser]
6. The treebank annotator’s decisions and annotation lead to the final tree.
7. The treebank annotations are run through a diagnostic process (a series of searches for known potential errors), and the errors found are corrected by hand in a quality control/correction pass to the extent that time allows.

CLUNCH, December 6, 2011
"سرية تجاوزها الاجبارية غير المرسم مع تطاعن الشعوب". كما اعتبرت أن من مقومات إصلاح دور الآية "الترفع عن المشاركة في معالجات نهجين لم تكن على علم الإتحاذ وعدم تغلب قافلة اجتماعية على أخرى". يدعب الحركة أن ذات الآية التي "تفرج بين البطلة الصالحة ونذل جلساء السوء، وحل الخلافات بين الاقتراحات من دون إعطاء المجال للانهيارين والتصديق والإعراض الخارجي.

"تداخل". وأكدت أن على رمز الآية "البعض من منافسة العامة في مجال المال والوظائف العامة".

Transliteration: waAlmtSydya
Comment: waAlmtSydya NOT FOUND
Annotation File: UMAAH_UM ARB_20020317-a.0020.xml
Paragraph Number: 9 (word 88)
Selected Analysis: (waAlmtSydya) wa/CONJ+AI/DET+mtSyd+yona/NSUFF_MASC_PL_ACCGEN : and + the + NOT_IN_LEXICON + [masc.pl]
POS: Follow Arabic traditional grammar

- List of prepositions strictly limited to traditional grammar list (most lexical items previously PREP now categorized as NOUN, or “prepositional nouns”)

- Particles given several POS alternatives: \( \text{فااء} \)
  - CONJ for \( \text{فااء الـعطف} \) فاء الـعطف for coordination: ‘and’
  - CONNEC_PART for \( \text{فااء الـربط} \) فاء الـربط comment after focus particle \( >\text{امـأ} / \) ا : ‘well (then)’
  - RC_PART for \( \text{فااء الجزء} \) فاء الجزء as a Response Conditional to introduce result of preceding conditional clause: ‘then’ or ‘so’
  - SUB_CONJ for \( \text{فااء السببية} \) فاء السببية to introduce subordinate result clause: ‘so that’
TB: Follow Arabic traditional grammar

- Constructions including
  - Comparatives
  - Numbers and numerical expressions
  - Several pronominal constructions such as
    - Separating pronouns/Damiyr Al-faSl
    - Anticipatory pronouns/Damiyr Al$a>n
  - More careful and complete classification of verbs and their argument structure
  - Thorough treatment of gerunds, participles and verbal nouns

- Intensive annotator training focused on agreement and consistency to (evalb) f-measure 94.3%
Dialectal Arabic (DA)

- DA is mostly spoken and rarely written because of Arabic diglossia
- Scarcity of existing written data compared to other target languages
- Lack of orthographic standard leads to inconsistency
- Undiacritized collected texts or transcribe diacritics
  - Arabic script-based DA is difficult to vocalize and understand because it is not usually diacritized. Only native speakers of a given Arabic dialect can provide the diacritics needed for reading comprehension.
  - Knowledge of missing diacritics is vital to WSD (Word Sense Disambiguation) and while this is true in MSA, it is even more important for the dialects.
  - Much more than in MSA, missing diacritics (short vowels and germination ‘shadda’) increase word level ambiguity
- Lack of NLP tools to help annotation tasks (taggers, parsers, morph analyzers)

CLUNCH, December 6, 2011
Future challenges for dialectal ATB (EA twitter feed)

hwa scorek fi IQ kam? Yaret t2oli eh l IQ da.xD aywaaa, Enti nazla 2moro emta?

We7sha awi lama neb2a fa ir m3 l nas, w manla2ish 7d fa ir m3ana

Ramadhan da begad 3'areeb, Msh zay eli fat 5ales

He's gd n she's gr8, but he thought she was superior so he didn't take it 2 da expected step n then he lost her 4eva. wats rong? Wer did U go? Enti bored leh, mdam 3ndk net w l donia msh 7ar, So what!

يا رب ارزقنا حبك و حب من أحبك و حب كل عمل يقربنا لحبك

Always seeking polar lights dream, Fighting for it. -Allah is enough for me

قال ابن تيميه" العبرة ليست بنقص البدايات و إنما العبرة بكمال النهايات." فاصبح فيما بقى يغفر لك ما سلف و إجتهد فلا تعلم متى تدرك رحمته

يغني لو شخص يبعلم حاجة غلوط و ينتج منه مفسده، نحاول أبدا نعمل نفس الحاجة غلوط عشان مجرد تقليل الفاسد؟

الثوواalletكلوا الساحة الخضراء يا رجاءااااااااااااااااااااالله أكبر
Two levels of annotation

- **“Source tokens”** (whitespace/punc-delimited)
  - source token text, vocalized form with POS and gloss

- **“Tree tokens”** (1 source token -> 1 or more tree tokens)
  - needed for treebanking
  - partition of source token vocalization/POS/gloss

### ATB analysis for one source token:

<table>
<thead>
<tr>
<th>S.TEXT:</th>
<th>ktbh</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>kutub</td>
</tr>
<tr>
<td>POS:</td>
<td>NOUN</td>
</tr>
<tr>
<td>GLOSS:</td>
<td>books</td>
</tr>
</tbody>
</table>

### Results in two ATB tree tokens:

<table>
<thead>
<tr>
<th>T.TEXT:</th>
<th>ktb</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>kutub+u</td>
<td>hu</td>
</tr>
<tr>
<td>POS:</td>
<td>NOUN+CASE.DEF.NOM</td>
<td>POSS_PRON.3MS</td>
</tr>
</tbody>
</table>
How design of ATB impacts on NLP pipeline

- Start with whitespace/punc-delimited source tokens
- End up with…
  - Morphological Analysis (useful for stuff)
    - Include vocalized form? (not all do)
    - Which POS tagset? (there are many)
    - How is tokenization even defined? (different forms of a token)
  - Trees (useful for stuff, like making more trees)
    - Parsing input depends on choices of output of morph analysis
    - Lots of tagset modifications.
    - To what extent can it all be integrated into one? (lattice-parsing, etc.)
    - dependency, phrase structure
    - How to take advantage of morphology, etc.
Morphological disambiguation

- **Analysis** – The SAMA (BAMA) analyzer gives a list of possibilities for an input source token.

- **Machine disambiguation** – select the \{pos,morph,lemma,vocalization\} for the given input word.
  - or maybe some subset of all of these?
  - roughly analogous to POS tagging
  - And maybe a different tagset than used in SAMA?

- **Roughly two approaches**
  - Use the SAMA tables as a set of possible solutions for each input word. Return everything that SAMA does
    - And for words not in SAMA?...
  - Don’t use the SAMA tables.
Morphological disambiguation

- **MADA** (Habash & Rambow 05…) – most established tagger – uses SAMA tables, produces SAMA solutions.
  - Separate (mostly SVM) classifiers for different features, (determiner? clitic?), assembled to decide on a SAMA solution.
  - Roughly 96% accuracy on lemmatization, tokenization, full tags except for noun case and verb mood.

- **SAMT** (Shah et al, 2010) – same input/output, different technology

- Also different approaches, without SAMA
  - **AMIRA** – “data-driven” pipeline (no SAMA) (Diab, 2004…)
  - Kulick, 2011 – weird hybrid, no SAMA
    - pos and tokenization simultaneously.
Early work (Bikel, 2004) – assume gold tree tokens, gold POS tags for input.

- But so many pos tags – map them down
  
  e.g. DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN -> NN

- Informally known as the “Bies” tagset.

Slightly later work (Kulick et al, 2006) – parsing improves if the determiner is kept in.

- DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN -> DT+NN

- Along with some other things, parsing went from 73 to 79, compared to English 87.5 on same amount of data

- Augmented with more and revised data, now up to 82.7 using gold tags (although parser not forced to use them, 84.1 if tags forced.)
Tagset Reduction Industry

From Table of Contents of Habash Arabic NLP book

5.4.1 The Buckwalter Tag Set

5.4.2 Reduced Buckwalter Tag Sets: Bies, Kulick, ERTS . . . .

From Stanford parser Arabic FAQ page

The parser uses an "augmented Bies" tag set. The so-called "Bies mapping" maps down the full morphological analyses from the Buckwalter analyzer to a subset of the POS tags used in the Penn English
More parsing: dependency

- CoNLL Shared task 2007 – Arabic still relatively poor.
- More recent work uses CATiB, the Columbia dependency version of the ATB. (using MaltParser)
  - interesting because using morphological features in a better way than tagset games
  - Also augmenting SAMA/ATB annotation to include more morphological information (e.g., broken plurals – functional instead of surface features)
  - Using predicted features from morph/pos step
  - Best score is 80.52 LAS, 83.66 UAS
- Stanford reports 77.4 phrase structure 84.05 unlabeled (?) dependency. (using gold tokens/tags (?))
Joint tokenization & parsing

- Usual idea – Avoid cascading errors
- Only one experiment I know of for Arabic - (Green and Manning, 2010) - Lattice Parsing
  - tokenization: MADA 97.67, Stanford Joint system – 96.2
  - parsing: gold 81.1, MADA 79.2, Joint: 76.
  - But “MADA is language specific and relies on manually constructed dictionaries. Conversely the lattice parser requires no linguistic resources.”

- I’m dubious
  - tokenization can be as high as 99.3% now, and I think it can go higher.
  - Is the joint model worth it?
Playing with morphological features is more convenient with dependency parsing
  • But we’re doing phrase structure treebanking
  • And what exactly is the relationship between phrase structure/dependency?

Goal:
  • Convert ATB to dependency (not necessarily same as Coumbia)
  • Parse with that, automatically convert back to phrase structure

Testing with Penn Treebank
  • Convert from dependency to phrase structure, with a 96.5 evalb.
  • Labelled score using Libin parser with 5 key function tags 90.9
  • Converts back to phrase structure with 87 evalb score
Questions?

استفسار؟

http://projects.ldc.upenn.edu/ArabicTreebank/
Example source token -> 2 tree tokens

<table>
<thead>
<tr>
<th>S_TEXT:</th>
<th>ktbh</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>kutub</td>
</tr>
<tr>
<td>POS:</td>
<td>NOUN</td>
</tr>
<tr>
<td>GLOSS:</td>
<td>books</td>
</tr>
</tbody>
</table>

Results in two ATB tree tokens:

<table>
<thead>
<tr>
<th>T_TEXT:</th>
<th>ktb</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>kutub+u</td>
</tr>
<tr>
<td>POS:</td>
<td>NOUN+CASE_DEF_NOM</td>
</tr>
</tbody>
</table>

- partition based on the reduced POS tags NOUN and POSS_PRON
- trivial for VOC,POS,GLOSS, not for S_TEXT->T_TEXT
Example source token -> 1 tree token

<table>
<thead>
<tr>
<th>S_TEXT:</th>
<th>Alktb</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>Al</td>
</tr>
<tr>
<td>POS:</td>
<td>DET</td>
</tr>
<tr>
<td>GLOSS:</td>
<td>The</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alktb</th>
</tr>
</thead>
<tbody>
<tr>
<td>kutub</td>
</tr>
<tr>
<td>i</td>
</tr>
<tr>
<td>NOUN</td>
</tr>
<tr>
<td>CASE_DEF_GEN</td>
</tr>
<tr>
<td>The books</td>
</tr>
<tr>
<td>[def.gen.]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T_TEXT:</th>
<th>Alktb</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC:</td>
<td>Al+kutub+i</td>
</tr>
<tr>
<td>POS:</td>
<td>DET+NOUN+CASE_DEF_GEN</td>
</tr>
</tbody>
</table>

- partition based on the reduced POS tag DET+NOUN
Morphological tagging other possibilities

- MADA (Habash & Rambow, 2005), SAMT (Shah et al., 2010) pick a single solution from the SAMA possibilities
  - tokenization, POS, lemma, vocalization all at once
- AMIRA – “data-driven” pipeline (no SAMA) (Diab, 2004…)
  - tokenization, then (reduced) POS – no lemma, vocalization
  - not entirely clear which form of the data is used for input
- Kulick, 2011 – weird hybrid
  - Like MADA, SAMT – simultaneous tokenization/POS-tagging
  - Like AMIRA – no SAMA