

Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank

Mohamed Maamouri, Ann Bies, Seth Kulick
Linguistic Data Consortium
University of Pennsylvania
{maamouri,bies,skulick}@ldc.upenn.edu

Arabic Treebank Newswire Corpora Sizes

Corpus	Source Tokens	Tokens after Clitic Separation
ATB1: AFP	145,386	167,280
ATB2: Umaah	144,199	169,319
ATB3: Annahar	339,722	402,246
ATB123 Total	629,307	738,845

Enhanced and revised Arabic Treebank (ATB)

Preview of key features & results

- ◆ Revised and enhanced annotation guidelines and procedure over the past 2 years. More complete and detailed annotation guidelines overall.
- ◆ **Combination of manual and automatic revisions of existing data to conform to new annotation specifications as closely as possible (ATB123)**
- ◆ Now being applied in annotation production
- ◆ Period of intensive annotator training
- ◆ Inter-annotator agreement f-measure scores improved to 94.3%.
- ◆ Parsing results improved to 84.1 f-measure

What is a Penn-Style Treebank

Penn-Style Treebanks are annotated CORPORA, which include linguistic information such as:

- Constituent boundaries (Clause, VP, NP, PP, ...)
- Grammatical functions of words or constituents
- Dependencies between words or constituents
- Empty categories as place holders in the tree for pro-drop subjects and traces

Syntactic Nodes in Treebank

(S (VP rafaDat رَفَضَتْ
(NP-SBJ Al+suluTAtu السُّلْطَاتُ)
(S-NOM-OBJ
(VP manoHa مَنَحَ
(NP-SBJ *)
(NP-DTV Al>amiyri الأميرِ
(NP-ARIBI AlhAribi الهَارِبِ)
(NP-OBJ (NP jawAza جَوَازَ
(NP safarK سَفَرِ))
(ADJP dyblwmasy~AF دَيْبِلُوْمَاسِيَّأُ))))))

رَفَضَتْ السُّلْطَاتُ مَنَحَ الأميرِ الهَارِبِ جَوَازَ سَفَرِ دَيْبِلُوْمَاسِيَّأُ

The authorities refused to give the escaping prince a diplomatic passport

Choice of Morphological Annotation Style

- ◆ BAMA: Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002)
- ◆ SAMA: LDC Standard Arabic Morphological Analyzer (2009)
- ◆ Input string → Analyzer provides
 - fully vocalized solution (Buckwalter Transliteration)
 - unique identifier or lemma ID
 - breakdown of the constituent morphemes (prefixes, stem, and suffixes)
 - their POS values
 - corresponding English glosses
- ◆ Guidelines available at <http://projects.ldc.upenn.edu/ArabicTreebank/>

Morphological Annotation Tool Screenshot

File

أستاذ عوض كذلك لأسمع رأيك بداية فيما يتعلق بهذا اللقاء

Transliteration: fiymA

Comment:

Annotation File: Y:/rachida/ATB5/ATB5-partc-pass1/xml/submitted-batch2/ALAM_wITHEVENT_ARB_20070206_205801.qtr.xml

Paragraph Number: 76 out of total 457 (word 7)

Selected Analysis: [(fiymA) [mA_1] fiy/PREP+mA/REL_PRON : in, by + what/which

Formatted Selection: fiy+mA PREP+REL_PRON in, by + what/which

fiymA	SUB_CONJ	While, whilst, as, during
fiymA	NOUN_PROP	FEMA (Federal Emergency Management Agency)
fiy+mA	PREP+REL_PRON	in, by + what/which
fiy+mA	PREP+INTERROG_PRON	in, by + what/which
fiy+mA	PREP+SUB_CONJ	in, by + that/if/unless/whether/as long as/as soon as

X-Solution Set Comment No match (<-) Passive form (<-) -u -N -hi NOUN -> ADJ Number (<-) Punctuation (<-)

Left Right MISSING_BEFORE MISSING_AFTER Hamza problem (<-) -a -F WCE ADJ -> NOUN Clitic Grammar Problem

Prev. Paragraph Goto Paragraph Next Paragraph VOC_VAR Gloss problem -i -K NCE NOUN -> Adverb Dialectal Remove Selection

Play Audio Stop Audio ChangeVowel Partial TRANSERR Typo Del CASE DISFL REG ACC NoSpeech FOREIGN

Choice of Syntactic Annotation Style

- ◆ Similar to Penn Treebank II
- ◆ Accessible to research community
- ◆ Based on a firm understanding and appreciation of traditional Arabic grammar principles
- ◆ Guidelines available at <http://projects.ldc.upenn.edu/ArabicTreebank/>

Syntactic Annotation Tool Screenshot

python

File

وَزَيْرُ النَّاجِيَةِ الْفلسطينِيَّةِ نُصِرَ يُوسُفَ لِحْجَمَاعاً طَرَفًا لِلْقِيَادَاتِ الْاِثْنِيَّةِ فِي ظِلِّ نَدْوَى الْاَوْضَاعِ فِي قِطَاعِ غَزَّةَ وَذَلِكَ مَعَ اقْتِرَابِ مَوْجِدِ الْاِتِّخَاذَاتِ التَّشْرِيعِيَّةِ الْفلسطينِيَّةِ

Prev Font Save Next

Trans Arabic Ref Gap Inde Gloss

Annotation File
E:/news/ATBS/en/ALIC_NEWS15_ARE_20060104_085000.xml
23 out of total 138 paragraphs
Information on action:

S	NP	VP
PP	SBAR	SBARQ
SQ	SNV	LST
NX	PRN	PRT
QP	ADJP	ADVP
FRAG	WHNP	WHPP
WHADJP	WHADVP	CONJP
INTJ	NAC	PRC
UCP	X	EDITED

SBJ	TRC	PRD	OBJ
PRP	CLR	LOC	DIR
MNR	TMP	ADV	LSS
NOM	DTV	VOC	BNF
EXT	CLF	HLN	TTL
ETC	IMP	SEZ	UNF

Up	Down	Redo	Undo
OpenAll		Print	UnAll
Trace	Gap	Coef	Print
NP*	Emp W	WHN O	Copy W

Rin tag	Rin func	
Rin Empty	Rin Coef	Rin Gap

Annotation tree:

- S
 - VP
 - va+Eqoid+u بعد 0 he/it + hold/convene/concili
 - NP-SBJ
 - NP
 - NP
 - waziy+u وزير 1 minister + [def.nom.]
 - NP
 - Ah+dAvilyy--+ap+i الداخلية 2 the + internal/domestic + [f
 - ADJP
 - Ah+filasoTiyiny--+u الفلسطيني 3 the + Palestrian + [def nor
 - NP
 - naSor نصر 4 Nasr
 - yusif يوسف 5 Yousif/Yusif/Joseph
 - NP-OBJ
 - NP
 - {jotimAE+AF اجتماعا 6 meeting/gathering/society +
 - TAnj)+AF طرفا 7 emergency/unscheduled/ur
 - PP
 - l- ل 8 to/for
 - NP
 - A+qiyAd+At+i لقيادات 9 the + leaders/commanders
 - Ah+>amoniy--+ap+i الأمنية 10 the + security/safety + [fer
 - PP
 - fiy في 11 in
 - NP
 - Zi+-i ظل 12 patronage/shelter/shade + [
 - NP
 - NP
 - tadahowur+i تدهور 13 deterioration/decline + [def
 - NP
 - Ah+>awoDAE+i الأوضاع 14 the + conditions/situation/st
 - PP-LOC
 - fiy في 15 in

Change W
fu -> hi
Change G
Spl W
Merge W
PSEUDO_VERB -> SUB_CONJ
SUB_CONJ -> PSEUDO_VERB
CONNEC -> CONJ
CONJ -> SUB_CONJ
vCASE_DEF_ACC
vCASE_DEF_NOM
vCASE_DEF_GEN
vCASE_INDEF_NOM
vCASE_INDEF_GEN
vCASE_INDEF_ACC
ADJ -> NOUN
NOUN -> ADJ
ADV -> NOUN
PREP -> NOUN
IV -> IV_PASS
PV -> PV_PASS
TYPO
TRANSERR

Start desktop - Notepad Inbox for fatma@ic... 1040New :: Online T... Shortcut to at5-TREE python answersforansA7B... 20062009appraisalf... 2:55 PM

Revision Process

- ◆ Motivation
 - Examination of inconsistencies in annotation
 - Lower than expected initial parsing scores
- ◆ Complete revision of annotation guidelines, both morphological and syntactic
- ◆ Combined automatic and manual revision of annotation in existing corpora: ATB1 (AFP), ATB2 (Umaah), ATB3 (Annahar)

Stages of Correction

Stage	Type
1. Complete manual revision of trees according to new guidelines	Human only
2. Limited manual correction of targeted POS tags	Human , based on automatic identification
3. Revision of targeted tokenization and POS tags according to new guidelines, based on purely lexical information	Automatic only
4. Revision of targeted tokenization and POS tags according to new guidelines, based on tree structure information	Automatic , based on human trees
5. Corrections based on targeted error searches	Human , based on automatic identification

Manual and Automatic Revision

- ◆ Stage 1 focused on a human revision of all of the trees.
- ◆ Stages 2 , 3 & 4 focused on revising lexical information, based in part on the new tree structures, using a combination of automatic and manual changes.
- ◆ Stage 5 focused on error searches targeting both lexical information and tree structures.

Stage 1: Manual Revision of Trees

- ◆ Introduction of iDAfa structure, e.g. (formerly flat NPs)

(NP كتاب kitaAbu book
(NP نحو naHowK grammar))
كتاب نحو
(a) *grammar book*

(NP every -kul~u - كُلُّ
(NP collection majomuwEapK مَجْمُوعَةٌ))
كُلُّ مَجْمُوعَةٍ
every collection

Stage 2: Manual correction of targeted POS tags

- ◆ Specific tokens ambiguous with respect either to multiple POS tags or to tokenization were revised by hand (about 13 passes deemed important include such tokens as *wa-*, *fa-* , *laysa* , *<il~A*, *Hat~aY* etc.)
- ◆ **Example: mA values in SAMA**
 1. mA/REL_PRON *what/which*
 2. mA/NEG_PART *not*
 3. mA/INTERROG_PRON *what/which*
 4. mA/SUB_CONJ *that/if/unless/whether*
 5. mA/EXCLAM_PRON *what/how*
 6. mA/NOUN *some*
 7. mA/VERB *not be*
 8. mA/PART [*discourse particle*]

mA: Relative Pronoun vs. Negative Particle

mA=REL_PRON

لِيَحْصُلَ عَلَى مَا يَسُدُّ رَمَقَهُ

li+yaHoSula ElaY **mA** yasud~u ramaqa+hu

for+gets (he) **what** fill breath of life+his

in order for him to get what he really craves

mA=NEG_PART

مَا زَالَ حَيًّا إِلَى الْآنَ

mA zAla Hay~AF <ilaY Al|na

not finished (he) alive until the+now

He doesn't cease to be alive now

Ma SUB_CONJ vs. mA REL_PRON

◆ بَعْدَمَا أَظْهَرْتَهُ لَهُ

after she showed (it) (to) him

◆ بَعْدَ مَا أَظْهَرْتَهُ لَهُ

After what she showed (it) (to) him

◆ بَعْدَمَا أَظْهَرْتَهُ لَهُ مِنْ حُبِّ ِ [ungrammatical]

after she showed (it) (to) him of love

◆ بَعْدَ مَا أَظْهَرْتَهُ لَهُ مِنْ حُبِّ ِ

after what she showed (it) (to) him of love

Stage 3: Automatic revision of targeted tokenization and POS tags based on lexical information only

- ◆ Use lexical information in revised guidelines and new SAMAs for “function words” as in PREP → NOUN
- ◆ Create a version of the corpus associating each original token from the source text file with the one or more Treebank tokens that together make up that original token
- ◆ Use this characterization of all original tokens to modify the tokenizations to match the new guidelines
 - Example: “limA*A” لَمَّاذَا → single token in new guidelines, from both single token and two token forms (“li” and “mA*A”) in pre-revision corpus

Stage 4: Automatic revision of targeted tokenization and POS tags based on lexical and tree information

Original unvocalized token	Possible vocalization/POS alternatives	Count in ATB123
<nmA or AnmA إِذَا إِذَا	<in~amA/RESTRIC_PART	138
	< i n - a / P S E U D O _ V E R B + m A / R E L _ P R O N	2
f y m A فِيْمَا	f i y / P R E P + m A / R E L _ P R O N	1 4
	f i y m A / S U B _ C O N J	2 5 6
k m A كَمَا	k a / P R E P + m A / R E L _ P R O N	2 3 3
	k a / P R E P + m A / S U B _ C O N J	1 2 5
	k a m A / C O N J	3 9 8
b m A بِمَا	b i / P R E P + m A / R E L _ P R O N	2 3 2
	b i / P R E P + m A / S U B _ C O N J	1 5

Stage 5: Manual corrections of automatic search results

- ◆ Searches targeting several types of potential inconsistency and annotation error
- ◆ Increased the number of error searches threefold during the revision process
- ◆ Run searches after annotation is complete
- ◆ Hand-correct all errors detected

Not Revised

- ◆ A certain residual type of correction is not possible in this context
 - Corrections that require too much human decision to be made automatically
 - But that are too frequent or otherwise too time-consuming to be made manually
- ◆ Example: highly complex and very frequent noun (NOUN) vs. adjective (ADJ) distinction in Arabic
- ◆ Time and funding allowing, a manual revision of these cases in the Arabic Treebank will be undertaken in the future, using an appropriate combination of automatic and manual means.

Parsing Experiment: Significant Improvement using Revised Data

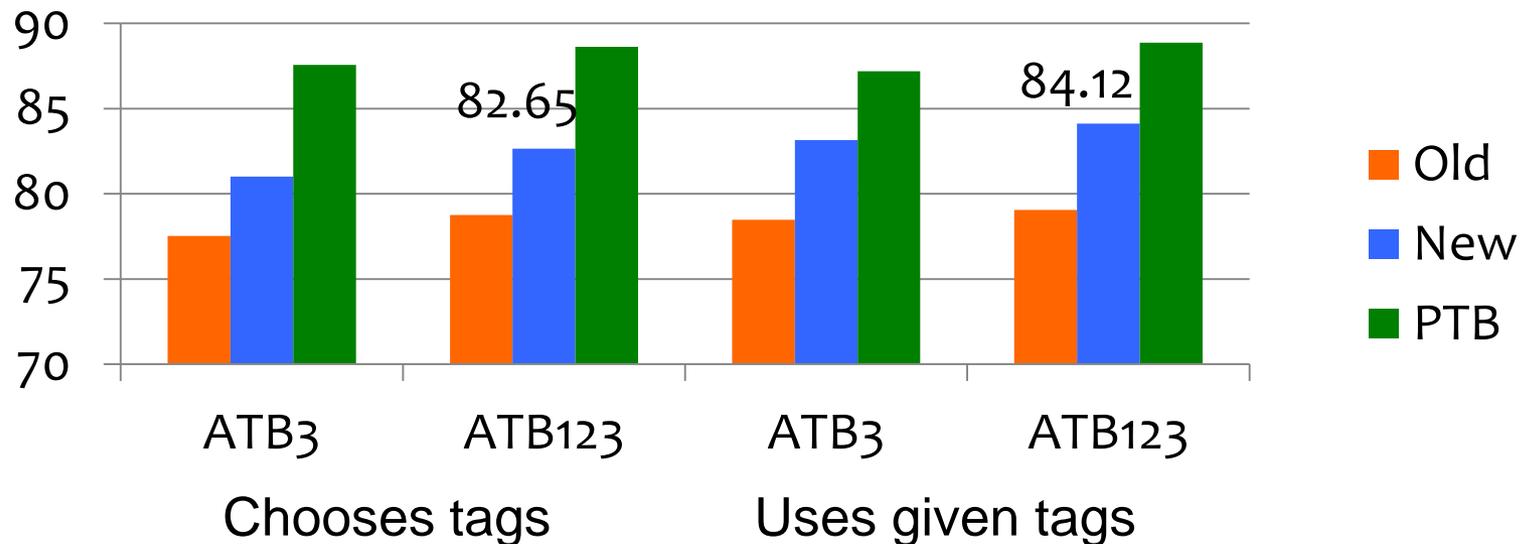
◆ New ATB and old ATB:

- Parsed ATB_{1,2,3} separately and ATB₁₂₃ together
- Mona Diab's train/dev/test split (≤ 40 words)
- Using gold tokenization and tags
- Two modes
 - Parser uses its own tags for “known” words
 - Parser forced to use given tags for all words
- LDC reduced TAG set (+DET)

◆ Penn (English) Treebank

- Made up training, test sets same size as ATB₃, 123

Parsing Improvement



- ◆ Nice improvement, not at PTB level yet, but closer
- ◆ Results not as good for test section
- ◆ Dependency Analysis shows:
 - Improvement in recovery of core syntactic relations
 - Problem with PP attachment!

(Kulick, Gabbard, Marcus TILT 2006, Gabbard & Kulick 2008 ACL)

Concluding Remarks

- ◆ Revised and enhanced guidelines
- ◆ Revised annotation in existing data
- ◆ Increased consistency
- ◆ Improved parsing results
- ◆ Combined manual and automatic corrections crucial to the revision process

THANK YOU FOR YOUR ATTENTION

For more information or if you have any
questions please contact

Dr. Mohamed MAAMOURI

<maamouri@ldc.upenn.edu>