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 رَفَضَت
(NP-SBJ Al+suluTAtu السُلُطاتُ)
(S-NOM-OBJ
(VP manoHa مَنْحَ
(NP-SBJ *)
(NP-DTV Al>amiyri الأَميرِ
AlhAribi الهارِبِ)
(NP-OBJ (NP jawAza جَوازَ
(NP safarK سَفَرٍ))
(ADJP dyblwmAsy~AF دِيبلوماسيّاً)))))

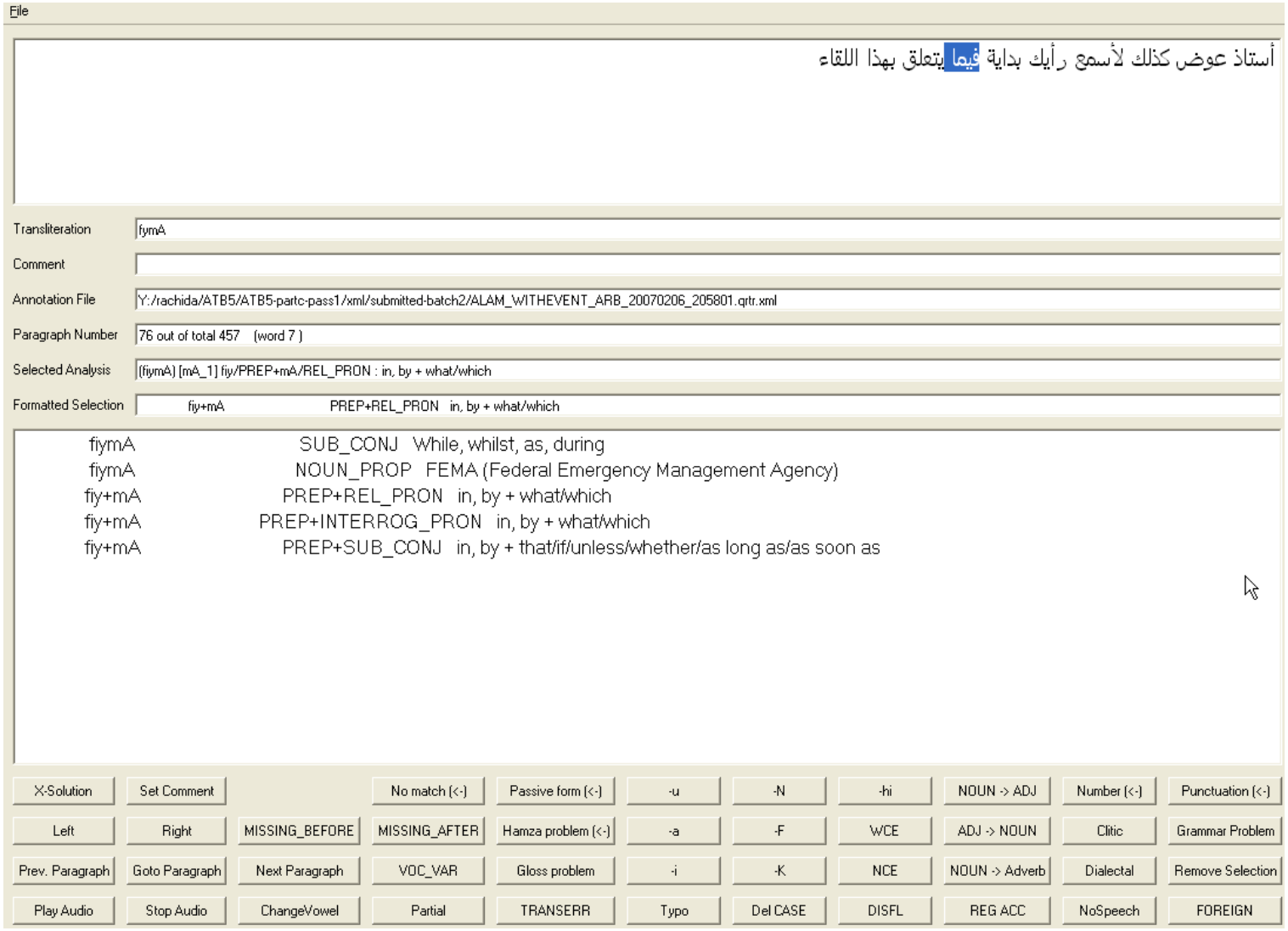
رفضت السلطات منحَ الأمير الهارب جوازَ سفرٍ دِيبلوماسياً

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/

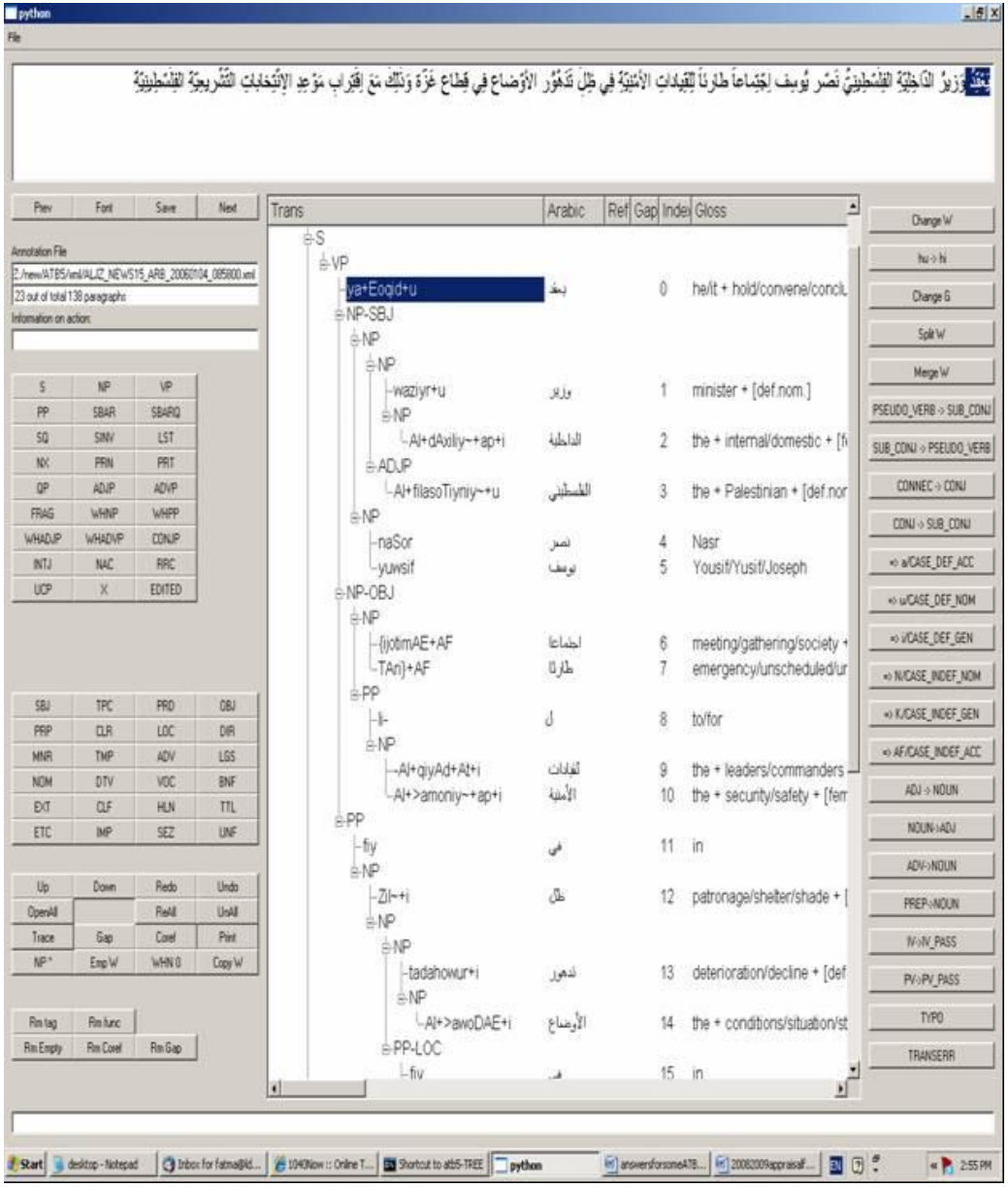
Morphological Annotation Tool Screenshot



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



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 نحوٍ
(a) grammar book

(NP every -kul~u - كُلُّ
(NP collection majomuwEapK مَجْمُوعَةٍ))
كُلّ 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 علَى ما 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

ما زالَ حَيّاً إِلَى الآنَ

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

- بَعدَما أظهرته لَه

  after she showed (it) (to) him

- بَعدَ ما أظهرته لَهُ

  After what she showed (it) (to) him

- بَعدَما أظهرته لَه مِن حُبٍّ [ungrammatical]

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

- بَعدَ ما أظهرته لَهُ مِن حُبٍّ

  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 SAMA 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" لِمَاذَا → 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 إذ ما اذ ما	<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 فِيما	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 ك م ا	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 بِما	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

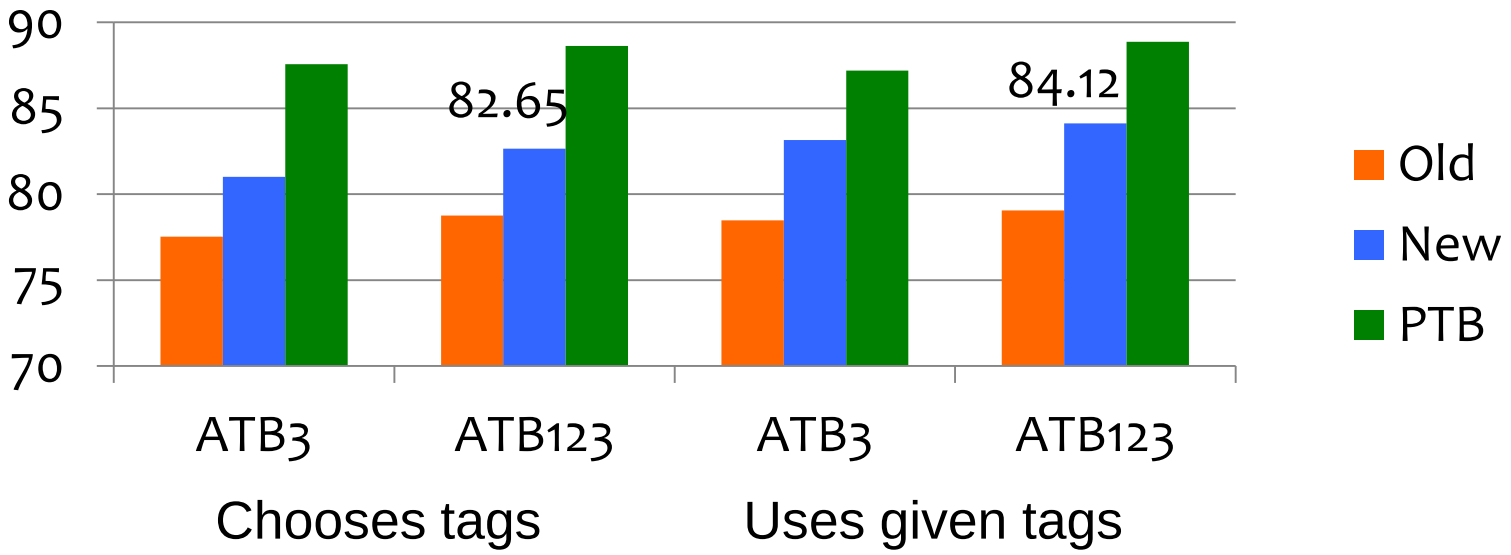
Not Revised

- 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 ATB1,2,3 separately and ATB123 together
  - Mona Diab's train/dev/test split (<=40 words)
  - Using gold tokenization andtags
  - 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>