Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank

Seth Kulick, Ann Bies, Mohamed Maamouri

Linguistic Data Consortium

- **Treebank Annotation Issue: Multiple Levels of Annotation**
  - Annotation not on the source text, but more abstract representation
  - How to maintain annotation consistency and relation between different levels?
  - How to make available the multiple levels of representation for the user?
- **Arabic Treebank as a case study**:
  - Mapping between two levels of annotation:
    - Morphological analysis of source text
    - Further tokenization for treebank annotation
  - Mapping between morphological analysis and separate morphological analyzer

Source Tokens and Tree Tokens

- Source text broken into whitespace-delimited "source tokens"
- Each source token receives a solution from SAMA (Standard Arabic Morphological Analyzer)
  - Sequence of segments, each having [vocalization, part-of-speech tag, gloss]
- Syntactic annotation uses "source token" SAMA analysis divided into "tree tokens" appropriate for syntactic analysis
- All syntactic annotation refers only to the vocalized form of the tree tokens, resulting from SAMA

- Example: source token ktbh receives SAMA analysis:
  - [kutub,NOUN,books]
  - [i,CASE_DEF_GEN,def.gen.]
  - [hi,POSS_PRON_3MS, its/his]

- Results in tree tokens for syntactic analysis:
  - [kutub+i, NOUN+CASE_DEF_GEN, books+def.gen.]
  - [hi, POSS_PRON_3MS, its/his]

Token Information in Data Publication

- **"POS-level"** information for each source token

INPUT_STRING	كتبه
IS_TRANS	ktbh
INDEX:	P22W10
OFFSETS:	42-46
TOKENS:	P22W13-P22W14
STATUS:	1
LEMMA:	[kitAb_1]
UNSPLITVOC	(kutubihi)
POS:	NOUN+CASE_DEF_GEN+ POSS_PRON_3MS
VOC	kutub+i+hi
GLOSS	books+[def.gen.]+its/his

- **Treebank-level** information for each tree token

INPUT_STRING	كتب
IS_TRANS	ktb
INDEX:	P22W13
OFFSETS:	42-45
UNVOCALIZED:	Ktb
VOCALIZED:	kutub+i-
POS:	NOUN+CASE_DEF_GEN
GLOSS	books+[def.gen.]

INPUT_STRING	ه
IS_TRANS	h
INDEX:	P22W14
OFFSETS:	45-46
UNVOCALIZED:	h
VOCALIZED:	-hi
POS:	POSS_PRON_3MS
GLOSS	Its/is

Trees and Alternate Tree Token Forms Provided in Integrated Format

- 1) with tree tokens (leaves) arising from the word index value (the index value includes all information from tree token tables):

```
(NP W13
    (NP W14))
```

- 2) with tree tokens arising from the VOCALIZED field:

```
(NP (NOUN+CASE_DEF_GEN kutub+i-)
    (NP (POSS_PRON_3MS -hi)))
```

- 3) with tree tokens arising from the UNVOCALIZED field:

```
(NP (NOUN+CASE_DEF_GEN ktb)
    (NP (POSS_PRON_3MS h)))
```

New and Improved Aspects of Token Information

- **New Information for source tokens:**
  - Explicit mapping between source and tree tokens (OFFSETS)
  - Relationship with SAMA (STATUS)

- **Improved information for tree tokens:**
  - UNVOCALIZED form and INPUT_STRING
  - Provided for users wishing to experiment with differing degrees of vocalization
  - These forms are created after annotation is finished. Syntactic annotation is on the VOCALIZED form.

Relationship Between Source Tokens and SAMA Solutions

- **STATUS 1 (INCLUDED IN SAMA):** Solution (POS,VOC) exactly matches a SAMA solution for the source token INPUT_STRING

INPUT_STRING	جندياً
IS_TRANS	jndyAF
INDEX:	P1W2
OFFSETS:	4-11
TOKENS:	P1W2-P1W2
STATUS:	1
LEMMA	[junodiy~_1]
UNSPLITVOC	(junodiy~AF)
POS:	NOUN+CASE_DEF_ACC
VOC:	junodiy~+AF
GLOSS:	soldier+[acc.indef.]

- **STATUS 2 (LIMITED SOLUTION):** Solution not SAMA solution, manually entered with no vocalization (TYPO, FOREIGN, DIALECT – not expected to be in SAMA)

INPUT_STRING	بتقوم
IS_TRANS	btqwm
INDEX:	P15W7
OFFSETS:	36-42
TOKENS:	P15W8-P15W8
STATUS:	2
LEMMA	None
UNSPLITVOC	None
POS:	DIALECT
VOC:	btqwm
GLOSS:	nogloss

- **STATUS 3 (PENDING SAMA SOLUTION):** Solution not SAMA solution, manually entered, but with vocalization, as a "pending" SAMA solution

INPUT_STRING	باتنه
IS_TRANS	bAnh
INDEX:	P6W15
OFFSETS:	68-73
TOKENS:	P6W18-P6W20
STATUS:	3
LEMMA	[bi>an~a_1]
UNSPLITVOC	(bi>an~ahu)
POS:	PREP+SUB_CONJ+PRON_3MS
VOC:	bi+>an~a+hu
GLOSS:	by/with+that+it/he

- **STATUS 4 (EXCLUDED FROM CHECK WITH SAMA):** Punctuation or other (non-Arabic script) token that by intent is not included in SAMA

INPUT_STRING	650
IS_TRANS	650
INDEX:	P1W1
OFFSETS:	0-4
TOKENS:	P1W1-P1W1
STATUS:	4
LEMMA	[DEFAULT]
UNSPLITVOC	(650)
POS:	NOUN_NUM
VOC:	650
GLOSS:	nogloss

Source Token to SAMA Relationship Over Entire Corpus

- Releases now have complete and explicit information relating each source token to SAMA

- ATB3-v3.2 – 339,710 source tokens

STATUS	MEANING	COUNT
1	INCLUDED IN SAMA	287,282
2	LIMITED SOLUTION	939
3	PENDING SAMA SOLUTION	4323
4	EXCLUDED FROM CHECK WITH SAMA	47,156
Total		339,710

Improved INPUT_STRING Form of Tree Token

- INPUT_STRING for a source token is simply the string characters in the source text

- INPUT_STRING for a tree token relates that tree token to a subsequence of characters for the source token it comes from

- Algorithm developed:
  - Input: source token INPUT_STRING and VOCALIZED tree tokens
  - Output: source token INPUT_STRING split up appropriately to correspond to the VOCALIZED tree tokens

- Sometimes simple:
  source token: ktbh →
  tree tokens VOCALIZED: (1) kutub+i and (2) hi
  tree tokens INPUT_STRING: (1) ktb and (2) h

- Sometimes not::
  source token: EmA →
  tree tokens VOCALIZED: (1) Ean and (2) mA
  tree tokens INPUT_STRING: (1) E and (2) mA

- Requires accounting for all possible types of SAMA normalization that might occur in the vocalized tree token

Improved UNVOCALIZED Form of Tree Token

- UNVOCALIZED form of tree tokens used in parsing and other work

- Previously had an inconsistent definition

- Now clean – simply the VOCALIZED form without diacritics

- source token: ktbh →
  tree tokens VOCALIZED: (1) kutub+i and (2) hi
  tree tokens INPUT_STRING: (1) ktb and (2) h
  tree tokens UNVOCALIZED: (1) ktb and (2) h

- source token: EmA →
  tree tokens VOCALIZED: (1) Ean and (2) mA
  tree tokens INPUT_STRING: (1) **E** and (2) mA
  tree tokens UNVOCALIZED: (1) **En** and (2) mA

Conclusions

- **Arabic Treebank and SAMA now a more tightly integrated unit**
  - Each release has explicit characterization of source tokens with SAMA
- **Arabic Treebank more easily usable for tagger experiments**
  - Explicit mapping between source tokens and tree tokens
- **More easily usable for experimentation with different input forms for parsing**
  - VOCALIZED,UNVOCALIZED,INPUT_STRING forms all provided for tree tokens
- **Multiple levels of annotation cleaner**
  - Changes at one level of annotation reflected in all the others