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

<table>
<thead>
<tr>
<th>INPUT_STRING</th>
<th>POS</th>
<th>ONS</th>
<th>VOCALIZED</th>
<th>UNVOCALIZED</th>
<th>GLOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>INPUT_STRING</th>
<th>POS</th>
<th>ONS</th>
<th>VOCALIZED</th>
<th>UNVOCALIZED</th>
<th>GLOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### STATUS 3 (PENDING SAMA SOLUTION): Solution not SAMA solution, manually entered, but with vocalization, as a ‘pending’ SAMA solution

<table>
<thead>
<tr>
<th>INPUT_STRING</th>
<th>POS</th>
<th>ONS</th>
<th>VOCALIZED</th>
<th>UNVOCALIZED</th>
<th>GLOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>INPUT_STRING</th>
<th>POS</th>
<th>ONS</th>
<th>VOCALIZED</th>
<th>UNVOCALIZED</th>
<th>GLOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>STATUS</th>
<th>MEANING</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INCLUDED IN SAMA</td>
<td>287,282</td>
</tr>
<tr>
<td>2</td>
<td>LIMITED SOLUTION</td>
<td>939</td>
</tr>
<tr>
<td>3</td>
<td>PENDING SAMA SOLUTION</td>
<td>4323</td>
</tr>
<tr>
<td>4</td>
<td>EXCLUDED FROM CHECK WITH SAMA</td>
<td>47,156</td>
</tr>
</tbody>
</table>

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 $kbh$ → tree tokens VOCALIZED: (1) $kbh$ and (2) $hi$
  - tree tokens INPUT_STRING: (1) $kbh$ and (2) $h$

- **Sometimes not:**
  - source token: $EmA$ → tree tokens VOCALIZED: (1) $Ea$ and (2) $mA$
  - tree tokens INPUT_STRING: (1) $E$ and (2) $m$

  - 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: $kbh$** → tree tokens VOCALIZED: (1) $kbh$ and (2) $hi$
  - tree tokens INPUT_STRING: (1) $kbh$ and (2) $h$
  - tree tokens UNVOCALIZED: (1) $kbh$ and (2) $h$

- **source token: $EmA$** → tree tokens VOCALIZED: (1) $Ea$ and (2) $mA$
  - tree tokens INPUT_STRING: (1) $E$ and (2) $m$
  - tree tokens UNVOCALIZED: (1) $Ea$ 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