



Columbia Arabic Treebank  
Status Report

Nizar Habash and Ryan Roth

Columbia University

`habash@ccls.columbia.edu`



# Overview of Accomplishments

---

- Objectives of the CATiB pilot treebank
  - ✓ – Create a manual for Arabic Lite dependencies
  - ✓ – Convert ATB1, 2&3 into CATiB representation
  - ✓ – Train a team of annotators
  - ✓ – Create a stable end-to-end annotation pipeline
  - ✓ – Produce ~~200K~~ words of annotations → 228K
  - ✓ – Do it fast (6 months) → 7 months

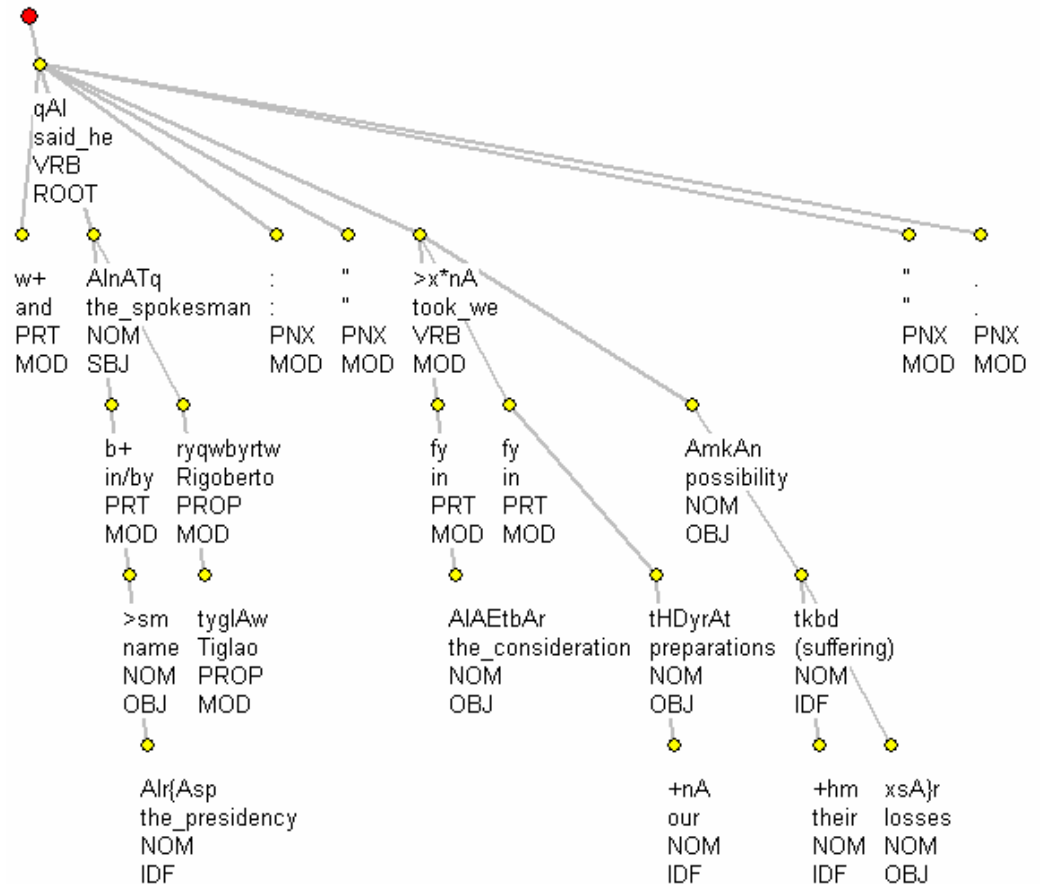
# Overview of Accomplishments

---

- Objectives of the CATiB pilot treebank
  - ✓ – Create a manual for Arabic Lite dependencies
  - ✓ – Convert ATB1, 2&3 into CATiB representation
  - ✓ – Train a team of annotators
  - ✓ – Create a stable end-to-end annotation pipeline
  - ✓ – Produce ~~200K~~ words of annotations → 228K
  - ✓ – Do it fast (6 months) → 7 months
  - ✓ – Achieve high speed of annotation
  - ✓ – Achieve high inter-annotator agreement (IAA)
  - ✓ – Improved parsing accuracy

# CATiB Annotation

- Tokenization
  - CONJ PART BASE PRON
- Part of Speech Tag set
  - VRB, VRB-pass, NOM, PROP, PRT, PNx
- Syntactic Annotation
  - SBJ, OBJ, TPC, PRD, IDF, TMZ, MOD and ---



# CATiB Annotation Pipeline

---

- Automatic POS tagging and tokenization
  - MADA+TOKAN (**34K wrd/hr**)
- Hand-correction of tokenization (**5.6K wrd/hr**)
- Automatic parsing (**8.3K tok/hr**)
  - MaltParser trained on converted ATB3 (initially)
  - MaltParser trained on converted ATB3 plus CATiB
- Hand-correction of parsed trees (**540 tok/hr**)
  - Correcting POS, attachment and labels
  - Each annotator worked 6 hr/wk on average
  - Project produced around 15K tok/wk (*400K tok total*)
- IAA Calculation
  - 10% of all annotated data was used
- Packaging

# Data Sets

<b>Annotated Data</b>	<b>Trees</b>	<b>Words</b>	<b>Tokens</b>
LDC2007E46 (nw)	1505	51K	61K
LDC2007E87 (nw)	2357	85K	102K
GALE-DEV07 (nw)	578	18K	21K
MT05 (nw)	957	27K	32K
MT06 (nw)	675	22K	26K
ATB3-Dev	1049	26K	31K
<b>SUB-TOTAL</b>	<b>7121</b>	<b>228K</b>	<b>273K</b>
<b>Converted Data</b>	<b>Trees</b>	<b>Words</b>	<b>Tokens</b>
ATB1-V3.0	5845	138K	166K
ATB2-V2.0	4302	140K	168K
ATB3-V3.0	14051	334K	401K
<b>SUB-TOTAL</b>	<b>24198</b>	<b>613K</b>	<b>735K</b>
<b>TOTALS</b>	<b>31319</b>	<b>841K</b>	<b>1M+</b>

# Inter-Annotator Agreement

*Seeded IAA files annotated by all five annotators; IAA with no punctuation*

IAA Set	Size	Sents Used	POS	ATT	LAB	LABATT
ATB3-Dev	2.4K	ALL	98.4	<b>92.9</b>	94.9	<b>90.0</b>
		$\leq 40$ Toks	98.5	<b>92.8</b>	94.2	<b>89.5</b>
PROD	3.8K	ALL	97.4	<b>90.3</b>	92.6	<b>85.8</b>
		$\leq 40$ Toks	97.4	<b>92.7</b>	93.8	<b>88.5</b>

IAA-ATB3-Dev is a subset of converted-then-annotated ATB3-Dev

IAA-PROD consists of all IAA files extracted from other (non-converted) sources

*The highest and lowest IAA files were used to examine Serial Annotation*

Serial IAA	Size	Toks/hour	POS	ATT	LAB	LABATT
Hi	333	398	96.7	<b>95.7</b>	95.7	<b>91.9</b>
Hi-Serial		956	96.7	<b>98.1</b>	97.7	<b>95.8</b>
Lo	350	476	98.2	<b>89.5</b>	91.2	<b>82.5</b>
Lo-Serial		944	97.5	<b>90.8</b>	93.4	<b>85.3</b>

# Inter-Annotator Agreement – Error Analysis

		Hi	Lo
<b>Average Length (words/sentence)</b>		<b>28</b>	<b>58</b>
<b>Source/Genre</b>		<b>AFP/news</b>	<b>Xinhua/financial</b>
<b>ATT</b>	All agree	<b>90.7</b> Serial: <b>96.0</b>	<b>80.2</b> Serial: <b>83.2</b>
	One dissenting	<b>7.3</b> Serial: <b>2.3</b>	<b>11.9</b> Serial: <b>8.5</b>
<b>LABATT</b>	All agree	<b>82.7</b> Serial: <b>91.0</b>	<b>68.9</b> Serial: <b>73.2</b>
	One dissenting	<b>12.6</b> Serial: <b>5.7</b>	<b>15.2</b> Serial: <b>13.4</b>
<b>POS</b>	NOM/PROP	<b>5.6</b>	<b>2.1</b>
	VRB/VRB-PASS	<b>0.7</b>	<b>0.9</b>
<b>ATT</b>	PP/Mod Attach	<b>4.6</b>	<b>8.2</b>
	Date/Curr/Prop/Num	<b>3.6</b>	<b>4.6</b>
	Subordination	<b>0.7</b>	<b>4.3</b>
<b>LAB</b>	MOD/---/IDF/TMZ	<b>5.0</b>	<b>11.0</b>
	MOD/ARG/ARG	<b>3.0</b>	<b>4.5</b>

# Conversion

*Reported over two 3K tok sets from ATB3-Train (no punctuation)*

	IAA (3)	Head Percolation	Rules Jun'08	Rules Jan'09	Stat Model + Rules	Ratio to IAA
A: ATT	94.5	86.0	90.9	<b>92.9</b>	<b>93.3</b>	<b>98.3%</b>
A: LABATT	91.3	55.4	87.3	<b>90.0</b>	<b>90.3</b>	<b>98.6%</b>
B: ATT	93.0	82.4	88.7	<b>90.0</b>	<b>91.0</b>	<b>96.8%</b>
B: LABATT	89.1	52.7	83.0	<b>85.9</b>	<b>86.6</b>	<b>96.4%</b>

*Conversion was done using an earlier release of ATBs  
We plan to re-visit conversion with new ATBs*

# Parsing with CATiB

*Parsing evaluated against the annotated ATB3-D (31K tokens; no punctuation)*

Training Set	Training Size	Sents Used	ATT	LABATT
CATiB-E	160K	ALL	<b>83.8</b>	<b>79.0</b>
		$\leq 40$ Toks	<b>84.3</b>	<b>79.5</b>
ATB3-T-Converted-Sized	160K	ALL	82.6	77.3
		$\leq 40$ Toks	83.7	78.5

*CATiB outperforms Converted ATB3 (for same size)*

# Parsing with CATiB

*Parsing evaluated against the annotated ATB3-D (31K tokens; no punctuation)*

Training Set	Training Size	Sents Used	ATT	LABATT
CATiB-E	160K	ALL	83.8	79.0
		$\leq 40$ Toks	84.3	79.5
ATB3-T-Converted-Sized	160K	ALL	82.6	77.3
		$\leq 40$ Toks	83.7	78.5
ATB3-T-Converted	339K	ALL	83.6	78.6
		$\leq 40$ Toks	84.6	79.7
CATiB-E+ ATB3-T-Conv'd	499K	ALL	<b>84.8</b>	<b>80.1</b>
		$\leq 40$ Toks	<b>85.7</b>	<b>81.1</b>
(2xCATiB-E)+ ATB3-T-Conv'd	659K	ALL	<b>84.9</b>	<b>80.2</b>
		$\leq 40$ Toks	<b>85.8</b>	<b>81.2</b>

*Combining CATiB and Converted ATB gives improved results*

## Other Activities

---

- CATiB Package
  - Available through LDC: LDC2009E06
  - Includes all data (annotated and converted)
  - English parallel data included
  - Phrase structure version of CATiB
  - Manual
- Helped IBM launch and test their CATiB-*lite*+ effort
  - Successfully installed the end-to-end pipeline
- Developed tools to generate alternative representations automatically
  - Separate particles (A1+, s+)
  - Reconstructed NPs and VPs to mimic English
  - Predict case and state from the dependency tree

## Future Directions

---

- Revisit Training
  - Focus on longer sentences, special constructions, proper name tagging and punctuation
  - Make serial annotation a part of the pipeline?
- Annotate more data
  - Different genres/sources
  - Dialectal Arabic
- Improve automatic conversion
  - Use the new and improved ATBs
  - Explore statistical methods for learning conversion from limited annotated data
  - Annotate additional converted data
- Continue the effort on CATiB automatic enrichment
- Continue working on parsing improvement



**CATiB**

Columbia Arabic Treebank

*Questions?*

Nizar Habash

habash@ccls.columbia.edu

