جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

Camel Treebank: An Open Multi-genre Arabic Dependency Treebank

Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas

جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

- **CAMeL Lab**
  Computational **A**pproaches to **M**odeling **L**anguage Lab
  - Established in September 2014

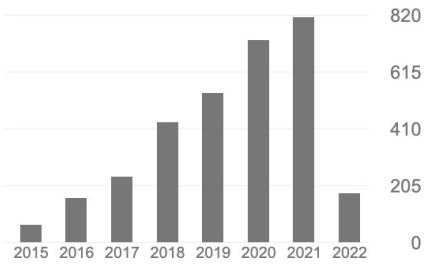
- **Research Areas**
  - Core Arabic & Arabic dialect NLP
  - Resource and tool development
  - Machine translation
  - Pedagogical applications
  - Dialogue systems

- **100+ publications**
- **20 resources**

- **Website:** www.camel-lab.com
- **Twitter:** @CamelNLP

- **Google Scholar:** scholar.camel-lab.com

	All	Since 2017
Citations	3194	2929
h-index	31	29
i10-index	69	66



Introduction

- Treebanks are collections of manually checked syntactic analyses of sentences
  - Important resources for building various NLP tools
  - Tokenization, diacritization, POS tagging, base phrase chunking, etc.
- Arabic is morphologically rich and highly ambiguous
  - Features include gender, number, person, case, state, aspect, mood, voice, and many attachable clitics
  - Optional diacritical marks
- Arabic has many variants
  - **CA**: Classical Arabic فصحى التراث
  - **MSA**: Modern Standard Arabic فصحى العصر
  - **DA**: Dialectal Arabic اللهجات العربية

Introduction

- Some of the Arabic treebanking efforts include
  - PATB: Penn Arabic Treebank (Maamouri, 2004)
  - PADT: Prague Arabic Dependency Treebank (Smrž and Hajič, 2006)
  - CATiB: Columbia Arabic Treebank (Habash and Roth, 2009)
  - Quran Treebank (Dukes et al., 2011)
  - ArPoT: Arabic Poetry Treebank (CATiB style) (Al-Ghamdi et al., 2021)
  - I3rab: A traditional grammatical theory treebank(Halabi et al., 2021)
- Each corpus is in a specific genre and variant
  - Most are in MSA News
  - Some have licensing restrictions
  - Some are not open
- Different representations

Camel Treebank

- **We present the Camel Treebank**
  - An open-source, multi-genre Arabic dependency treebank
  - Focus on MSA and CA
  - A wide range of texts from Pre-Islamic poetry to Social Media commentaries
  - 188K word / 242K tokens
  - CATiB style dependency treebank

- **http://treebank.camel-lab.com/**

Roadmap

- Introduction
- Data Selection
- Guidelines & Extensions
- Annotation Process
- Evaluation & Observations

Data Selection

Sub-Corpus	Text Source	Variant	Century	Genre	#Lines	#Sentences	#Words
Odes	Suspended Odes (Mu'allaqat)	CA	6th	Poetry	784	784	7,465
Quran	Quranic Surahs	CA	7th	Quranic	50	572	11,699
Hadith	Hadiths from Sahih Bukhari	CA	7th	Prophetic Sayings	135	1,190	12,467
1001	One Thousand and One Arabian Nights	CA	12th	Stories	44	1,145	11,831
Hayy	Hayy ibn Yaqdhan (Ibn Tufail)	CA	12th	Philosophical Novel	391	1,198	19,674
OT	Old Testament	MSA	19th	Bible Translation	111	535	9,097
NT	New Testament	MSA	19th	Bible Translation	113	573	9,593
Sara	Sara (Al-Akkad)	MSA	20th	Novel	1,585	1,585	35,356
ALC	Arabic Learner Corpus	MSA	21st	Student Essays (L2)	86	727	9,221
BTEC	Basic Traveling Expressions Corpus (MSA)	MSA	21st	Phrasebook	2,000	2,000	15,935
QALB	QALB Corpus	MSA	21st	Online Commentary	200	923	11,454
WikiNews	WikiNews	MSA	21st	News	393	996	18,314
ZAEBUC	Zayed Bilingual Undergraduate Corpus	MSA	21st	Student Essays (L1)	166	1,109	15,778
					6,058	**13,337**	**187,884**

- 13 sub-corpora
- Texts are out of copyright, creative commons, or under open licenses
- Not a balanced corpus; but representatively diverse of time and genre
- Restricted annotation budget
- Additional public annotations

CATiB Guidelines

- **CATiB = Columbia Arabic Treebank**
  - A simplified depenedency representation inspired by Traditional Arabic grammar
  - 6 POS tags + 8 dependency relations

- **Tokenization**
  - conjunction+ particle+ BASE +pronoun

وكبيوتنا	wkbywtnA
و+ ك+ بيوت+ نا	w+ k+ bywt +nA
	and+like+houses+our

CATiB Guidelines

- **Part-of-Speech Tags**
  - **VRB**: Active Verb
  - **VRB-Pass**: Passive Verb
  - **NOM**: noun, adjective, adverb, pronoun, digits, etc.
  - **PROP**: Proper Noun
  - **PRT**: Particles
  - **PNX**: Punctuation
- **Dependency Relations**
  - **SBJ**: Subject (of nominal and verbal sentences)
  - **OBJ**: Object (of verbs, and prepositions)
  - **PRD**: Predicate
  - **TPC**: Topic
  - **IDF**: Idafa (possessive construction)
  - **TMZ**: Tamyiz (specification modifier)
  - **MOD**: Modifier
  - **---**: Flat

CATiB Guideline Extensions

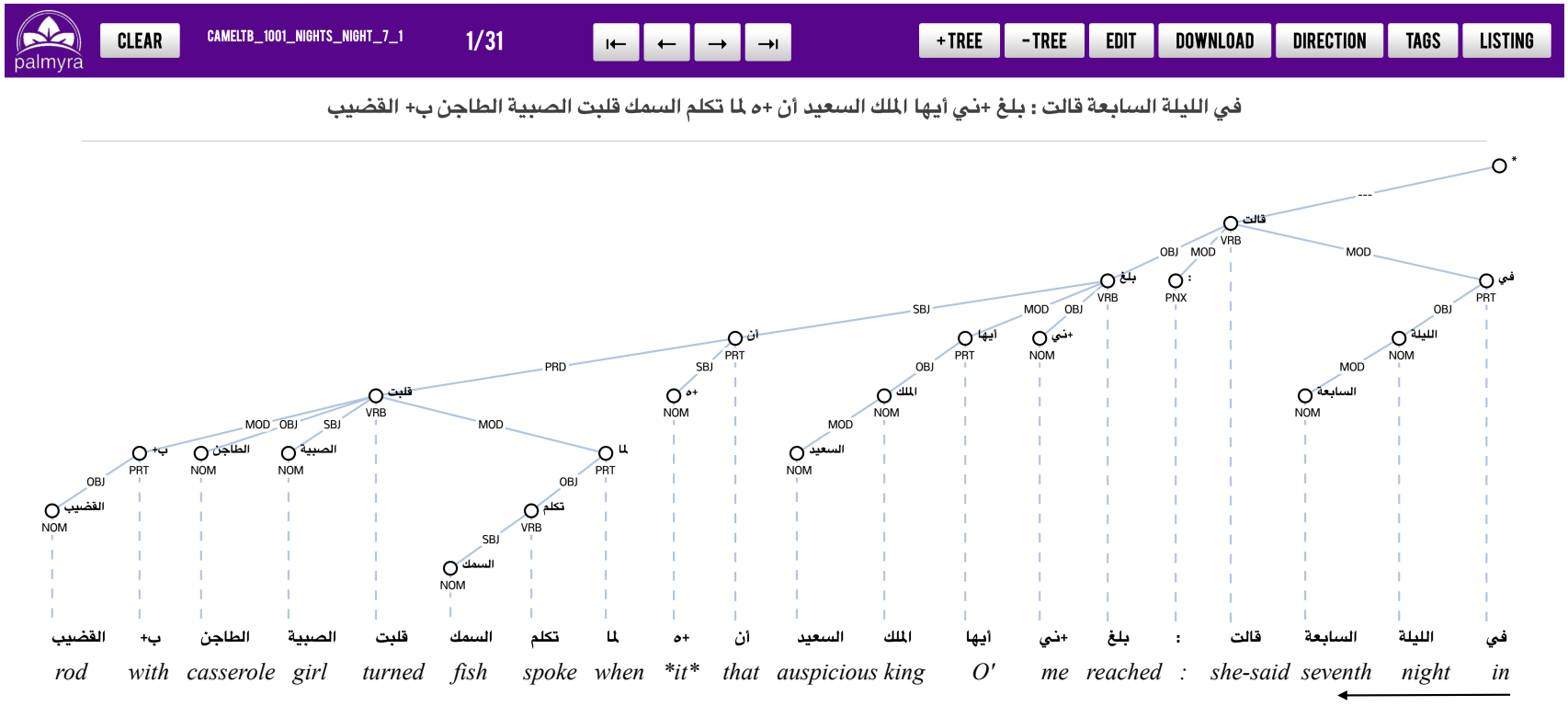
- **Foreign Tokens**
  - POS Foreign
- **Elided Tokens**
  - We allow adding elided tokens marked with a (*) suffix
  - 1 in 10,000, mostly in Quran and Odes
- **New Constructions**
  - 2nd person statements, interrogatives, interjections, so-called frozen verb constructions, and verse numbers in Holy Texts
- **Sentence Segmentation**
  - Break up very long punctuation scarce paragraphs
  - Not applied to Holy texts or Poetry

Annotation Process

- **Semi-automatic Sentence Segmentation**
  - Initial segmentation on !؟؛.
  - Arabic commas (،) are used for phrase and clause boundary
  - Manual segmentation by merging and splitting
- **Automatic Annotation**
  - Modified Camel Parser (Shahrour et al., 2016)
  - Camel Tools (Obeid et al., 2020)
  - MALT parser (Nivre et al., 2006) trained on CATiB converted PATB.
- **Manual Annotation**
  - Four Arabic native speakers
  - Extensive experience in treebanking or linguistic training

Annotation Interface

- **Palmyra 2.0**
  - We used the Palmyra 2.0 interface for manual annotation.
  - Palmyra allows modifying the tokenization, POS, and relations.



Evaluation Metrics

- **Tokenization F-1 (TOK)**
  - F-1 score of the precision and recall of correctly tokenized aligned tokens
- **POS Accuracy (POS)**
  - The percentage of gold tokens with correct POS
- **Label Score (LS)**
  - The percentage of gold tokens with correct dependency labels
- **Unlabeled Attachment Score (UAS)**
  - The percentage of gold tokens with correct dependency arcs
- **Labeled Attachment Score (LAS)**
  - The percentage of gold tokens with correct dependency labels & arcs

Evaluation

- **Annotation Validation**
  - TOK and POS scores are 99.9% and 99.7% on average, respectively
  - LS is 97.3% on average, from 99.0% (**ZAEBUC**) to 93.2% (**Quran**)
  - UAS is 95.5% on average, from 98.7% (**ZAEBUC**) to 91.6% (**NT**)
  - LAS is 94.5% on average, from 98.2% (**ZAEBUC**) to 90.2% (**NT**)
  - For **NT**, over half of the disagreements involved PNX, and PRT
- **Automatic Parsing Evaluation**
  - TOK is 97.1% on average, from 99.0% (**WikiNews**) to 90.8% (**Odes**)
  - POS is 93.4% on average, from 95.3% (**Hayy**) to 85.6% (**Odes**)
  - LS is 82.0% on average, from 90.9% (**WikiNews**) to 68.4% (**Odes**)
  - UAS is 75.5% on average, from 83.6% (**WikiNews**) to 66.3% (**Odes**)
  - LAS is 69.1% on average, from 79.7% (**WikiNews**) to 56.1% (**Odes**)
  - Gold Tokenization and POS increase the dependency scores by ~4.8%

Some Cross-Genre Observations

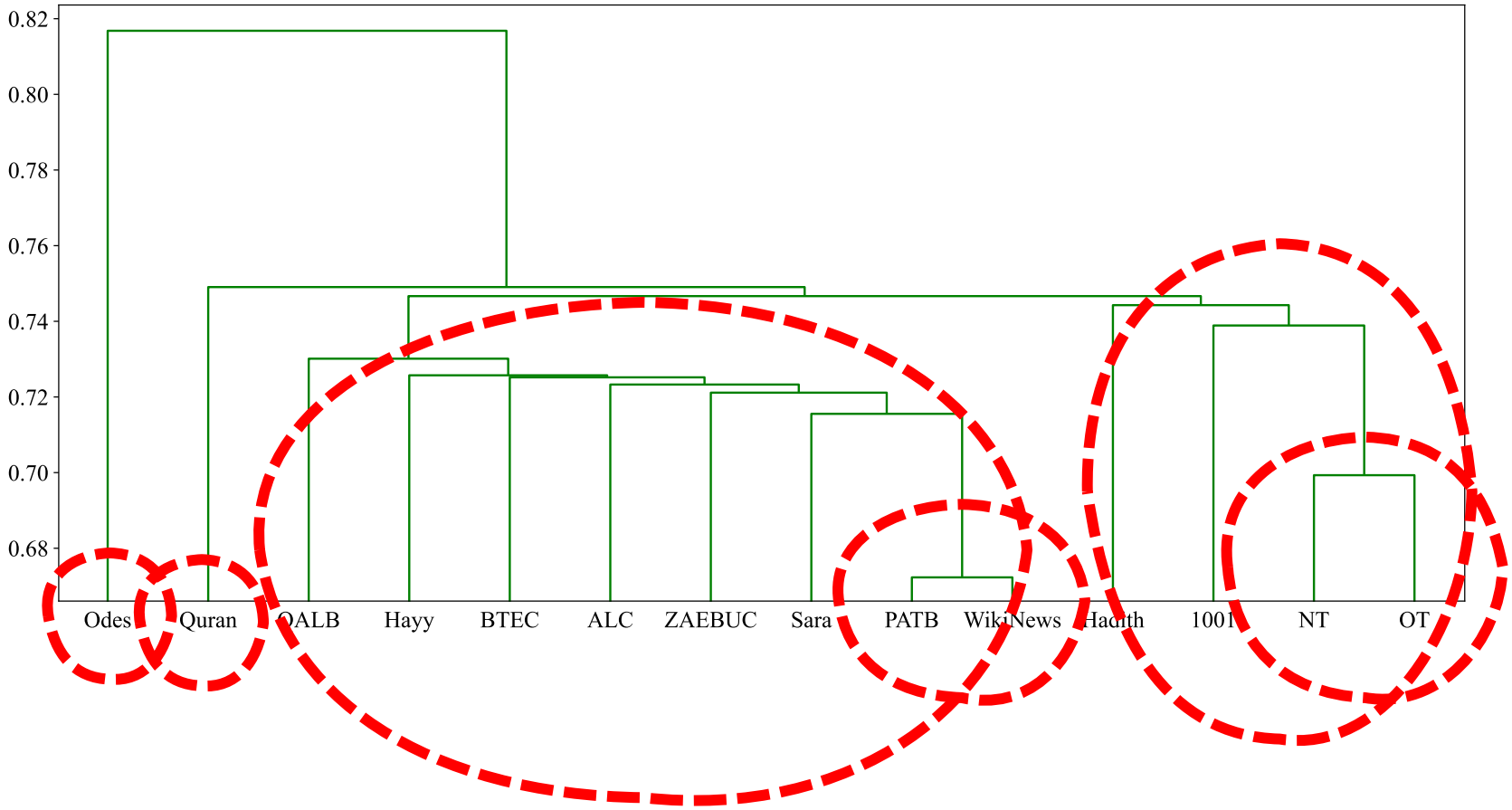
- **Tokenization Variations**
  - The tokenization ratio (tokens/word) varies widely from 1.45 tokens/word in **1001** and 1.36 in **Odes** to 1.17 in **BTEC** and **WikiNews**, with an average of 1.29
  - The correlation between the token/word ratio and text century is -59.4% – *the older the text, the higher the token/word ratio*
- **Idafa (Possession) Relation Variations**

  رئيس مجلس إدارة شركة مايكروسوفت

  Chair *of* board *of* direction *of* company *of* Microsoft = Microsoft's CEO
  - 6-7% in **Hadith**, **Quran** to 16.4% in **WikiNews**
  - Quran has max of length 2 Idafa chains, while **WikiNews** reaches 5
  - The correlation between text century and Idafa chain length is 70.5% – *the older the text, the shorter the Idafa chain length*

Some Cross-Genre Observations

- **Lexical Similarity**



Conclusion & Future Work

- **We presented Camel Treebank**
  - ~188K word/ ~242K token manually annotated open-source dependency treebank
  - MSA and CA texts from different historical periods and genre
  - Some interesting insights about syntax and different genres in Arabic
- **Our future plans include**
  - Additional texts from other periods and genres
  - Additional texts from Arabic dialects
  - Developing improved genre-aware parsing models
  - Analysis of genre differences in syntactic dimensions

جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

Thank you!

treebank.camel-lab.com

nizar.habash@nyu.edu