Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2672–2681
Marseille, 20-25 June 2022

Camel Treebank: An Open Multi-genre Arabic Dependency Treebank

Nizar Habash, Muhammed AbuOdeh, Dima Taji,[†]
Reem Faraj,[‡] Jamila El Gizuli,[*] and Omar Kallas
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
[†]Birzeit University
[‡]Columbia University & CUNY Graduate Center
[*]Georgia Institute of Technology
nizar.habash@nyu.edu

Abstract

We present the Camel Treebank (CAMELTB), a 188K word open-source dependency treebank of Modern Standard and Classical Arabic. CAMELTB 1.0 includes 13 sub-corpora comprising selections of texts from pre-Islamic poetry to social media online commentaries, and covering a range of genres from religious and philosophical texts to news, novels, and student essays. The texts are all publicly available (out of copyright, creative commons, or under open licenses). The texts were morphologically tokenized and syntactically parsed automatically, and then manually corrected by a team of trained annotators. The annotations follow the guidelines of the Columbia Arabic Treebank (CATiB) dependency representation. We discuss our annotation process and guideline extensions, and we present some initial observations on lexical and syntactic differences among the annotated sub-corpora. This corpus will be publicly available to support and encourage research on Arabic NLP in general and on new, previously unexplored genres that are of interest to a wider spectrum of researchers, from historical linguistics and digital humanities to computer-assisted language pedagogy.

Keywords: Arabic, Syntactic Dependency Treebank, Multiple Genres, Open Source

1. Introduction

A lot of research and system development in natural language processing (NLP) relies heavily on the existence of data enriched with annotations that represent the specific linguistic features of the text. In the case of Arabic, a morphologically rich and complex language, the creation of the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) has led to the creation of many enabling technologies: tokenization, POS tagging, base phrase chunking and syntactic parsing (Pasha et al., 2014; Shahrour et al., 2016; Obeid et al., 2020), among others.

Since the creation of the PATB, a number of other treebanks for Arabic were created with different texts (as opposed to same texts, but different formalisms), e.g., the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009), the Quran Corpus (Dukes and Buckwalter, 2010), the Arabic Basic Traveling Expressions Corpus (Taji et al., 2018), and most recently the Arabic Poetry Treebank (ArPoT) (Al-Ghamdi et al., 2021). Each of these efforts targets a specific genre, with MSA news getting the lion's share of attention. The news genre comes with an additional problem, namely, copyright restrictions on the original text, which limits access to news-based treebanks.

In this paper we present the Camel Treebank (CAMELTB) a manually annotated large ($\sim$188K words, $\sim$242K tokens) open-source dependency treebank of Modern Standard Arabic (MSA) and Classical Arabic (CA) in the style of CATiB dependencies. We designed CAMELTB to include selections of texts ranging from pre-Islamic poetry to social media online commentaries, and covers a range of genres from religious and philosophical texts to news, novels, and (L1&L2) student essays. All of the selected texts are publicly available (out of copyright, creative commons, or under open licenses), and some were independently previously annotated for other NLP tasks such as spelling and grammar correction, POS tagging, and lemmatization. As part of the creation of this corpus, we extensively extended and updated the CATiB guidelines to accommodate the needs of these new texts. The CAMELTB and its guidelines are publicly available.[1]

Section 2 presents some relevant background and related work. Section 3 presents the various considerations we took in data selection and introduces the CAMELTB sub-corpora. Sections 4 and 5 discuss our annotation guidelines and process, respectively. Section 6 presents our evaluation results. And Section 7 explores some preliminary analysis in lexical and syntactic variation among the CAMELTB genres.

2. Background and Related work

We present some of the relevant Arabic NLP challenges and discuss related efforts on Arabic treebanking.

2.1. Arabic NLP Challenges

The automatic processing of Arabic text faces a number of challenges: orthographic ambiguity, morphological richness, and linguistic variations. First, Arabic is orthographically ambiguous due to the optional writing of its short vowels. This results in around three different core readings per word on average (Habash, 2010). Second,

[1] http://treebank.camel-lab.com/

Arabic is morphologically rich with words expressing a number of features such as gender, number, person, voice, etc., in addition to attachable clitics that include the definite article, some prepositions and possessive pronouns. For example, the single Arabic word ولمعان *wlmςAn*[2] can be interpreted as وَ+لَمَعَانٌ *wa+lamaςaAnũ* 'and glitter [nominative]', or وَ+لِ+مَعَانٍ *wa+li+maςaAnĩ* 'and for some meanings', among others.

Finally, the official form of Arabic today, Modern Standard Arabic (MSA فصحى العصر) coexists with a number of dialectal variants that differ from it phonologically, morphologically, syntactically, and lexically. Additionally, an older form of Arabic, Classical Arabic (CA فصحى التراث) continues to be used and referenced, particularly in religious contexts. CA is generally similar to MSA in terms of syntactic structures; but with many lexical differences.

MSA has historically received the lion's share of attention in developing NLP systems. Dialectal Arabic has increasingly been getting resources and systems built; but Classical Arabic remains relatively impoverished (Habash, 2010; Inoue et al., 2021).

In this work, we use CamelTools (Obeid et al., 2020) for automatic tokenization and POS tagging; and portions of CamelParser (Shahrour et al., 2016) for automatic parsing.

2.2. Arabic Treebanks

There are a number of Arabic Treebanks with different sizes, syntactic formalisms, and focus genres.

The Penn Arabic Treebank (PATB) is the primary treebank for work on Arabic syntactic analysis (Maamouri et al., 2004). It uses a phrase-structure representation, but has been converted to other dependency formalisms (Habash and Roth, 2009; Taji et al., 2017). The PATB contains various parts that come from different domains and resources, but primarily from news or web sources (Maamouri et al., 2010). Other related treebanks were also developed by the Linguistic Data Consortium (LDC) in various dialects such as Egyptian (Maamouri et al., 2012), and Levantine (Maamouri et al., 2006), where the data came from transcribing recorded conversations.

The first Arabic dependency treebank was the Prague Arabic Dependency Treebank (PADT) (Smrž et al., 2002). It employed a multi-level description scheme for functional morphology, analytical dependency syntax, and tectogrammatical representation of linguistic meaning. Another Arabic dependency treebank is the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009). CATiB has around 250K words that were annotated directly in its dependency representation, which is inspired by traditional Arabic grammar. The Quran Corpus is another important Arabic syntactic corpus of the very specific genre of Quranic scripture (Dukes and Buckwalter, 2010). It has its own representation scheme which is a hybrid of dependency and constituency representations.

Most recently, Al-Ghamdi et al. (2021) presented the first CA Arabic Poetry Treebank (ArPoT). They used the CATiB formalism with some extensions. Another notable addition is the *i3rab* treebank, which follows a dependency representation more directly matching traditional Arabic grammatical theory (Halabi et al., 2021). In this paper, we follow the CATiB dependency representation style, with minor extensions.

3. Data Selection

In making the specific selection of the texts we annotated, we wanted to cover a large historical span (from 6th to 21st century), with a large set of genres. But most importantly, we wanted the texts to be publicly available (out of copyright, creative commons, or under open licenses). Also, we wanted to have a large enough selection from any single text genre-period to be able to have data that can be used for fine-tuning and evaluation later on. In this edition of CAMELTB, we do not focus on creating a balanced historical corpus, just one that is representatively diverse. We were restricted by an annotation budget that affected how many data sets we could work with. There were many interesting options that we decided to leave to future annotation follow-up projects. Some of the choices we made were influenced by the fact that other annotations existed for them. For example, the **QALB** (Zaghouani et al., 2014), **ZAEBUC** (Habash and Palfreyman, 2022), **WikiNews** (Abdelali et al., 2016), and **ALC** (Alfaifi, 2015) data sets all have additional non-syntactic annotations targeting NLP tasks such as spelling correction, POS tags, and diacritization. We hope that our annotations will encourage researchers to explore the use of syntactic representations with these tasks. For this edition of the CAMELTB we include the following 13 sub-corpora.

The Suspended Odes (Odes) The full text of the ten most celebrated poems from Pre-Islamic Arabia (المعلقات Mu'allaqat). All texts were extracted from Wikipedia.[3]

Quran The first three and last 14 Surahs from the Holy Quran. We selected the text from the Quran Corpus Project (Dukes et al., 2013).[4]

Hadith The first 134 Hadiths from Sahih Bukhari (al Bukhari, 846). We selected the text from the LK Hadith Corpus[5] (Altammami et al., 2019).

One Thousand and One Nights (1001) The opening narrative and the text of the first eight nights from the Arabian Nights (Unknown, 12th century). We extracted the text from an online forum.[6]

[2] The transliteration scheme is HSB (Habash et al., 2007).

[3] https://ar.wikipedia.org/wiki/المعلقات

[4] https://corpus.quran.com/

[5] https://github.com/ShathaTm/LK-Hadith-Corpus

[6] http://al-nada.eb2a.com/1000lela&lela/

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**

Table 1: The 13 sub-corpora of CAMELTB 1.0. #Words counts white-space and punctuation tokenized words.

Hayy ibn Yaqdhan (Hayy) The full text of the philosophical novel and allegorical tale written by Ibn Tufail (Tufail, 1150). We extracted the text from the Hindawi Foundation website.[7]

Old Testament (OT) The first 20 chapters of the Book of Genesis (Smith and Van Dyck, 1865).[8]

New Testament (NT) The first 16 chapters of the Book of Matthew (Smith and Van Dyck, 1860).[8]

Sara The full text of *Sara*, a novel by Al-Akkad first published in 1938 (Al-Akkad, 1938). We extracted the text from the Hindawi Foundation website.[9]

QALB 200 online comments from the Qatar Arabic Language Bank (QALB) Corpus (Zaghouani et al., 2014; Mohit et al., 2014).

ZAEBUC 100 student-written articles from the Zayed University Arabic-English Bilingual Undergraduate Corpus (Habash and Palfreyman, 2022).

BTEC The MSA translation of the Basic Traveling Expression Corpus (Eck and Hori, 2005; Takezawa et al., 2007). This portion of the corpus revises a previously reported effort by the authors (Taji et al., 2018).

WikiNews 70 Arabic WikiNews articles covering politics, economics, health, science and technology, sports, arts, and culture (Abdelali et al., 2016).

ALC 20 L2 articles from the *Arabic Learner Corpus* (Alfaifi, 2015).

Table 1 lists these sub-corpora with some additional information. About 40% of the text words come from 21st century sources, 35% from 19th and 20th century sources, and the rest from 6th to 12th century sources. About one-third of the selections by word count are fiction (novels, stories) and one-quarter religious texts.

4. Treebanking Guidelines

We followed the Columbia Arabic Treebank (CATiB) annotation scheme (Habash et al., 2009), but with some extensions. We chose CATiB because its relational labels and dependency structure are inspired by traditional Arabic grammar, making it intuitive for Arabic speakers, and allowing for faster annotation. CATiB representations can be automatically enriched with more morphological features (Alkuhlani et al., 2013), and converted into other dependency formats such as Universal Dependencies (Taji et al., 2017). Being a functional head dependency representation, CATiB is similar to Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2021).

We describe next the basic CATiB guidelines; and then summarize our extensions. The new updated guidelines are publicly available.[1]

4.1. Basic CATiB Guidelines

Tokenization We followed the PATB/CATiB tokenization scheme, which tokenizes all the clitics, except for the definite article ال+ *Al+* 'the' (Maamouri et al., 2004).

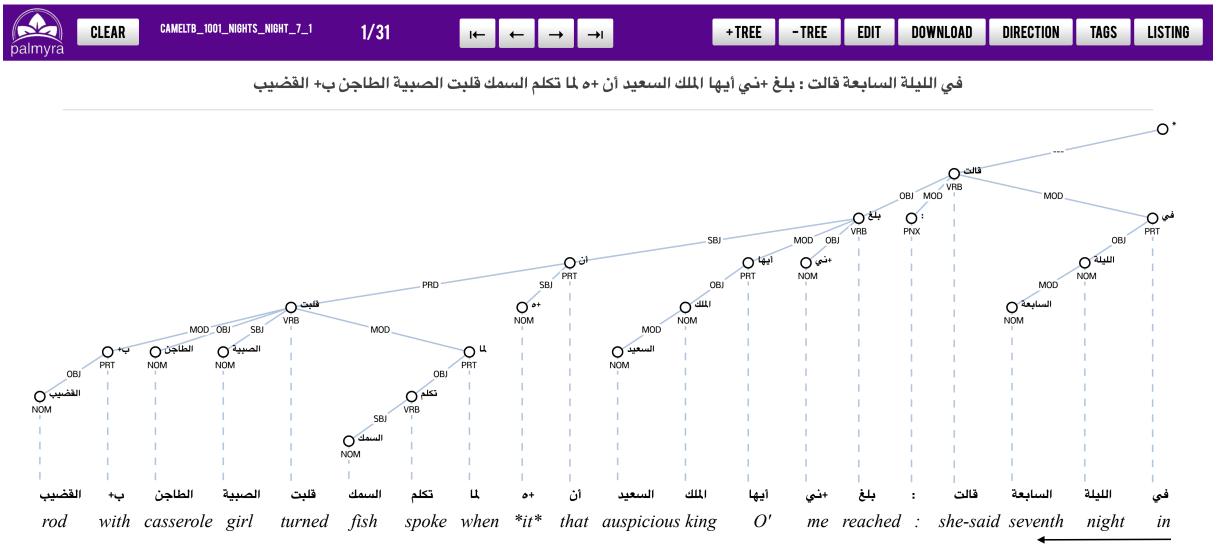
POS Tags CATiB uses six POS tags: **NOM** for all nominals excluding proper nouns, **PROP** for proper nouns, **VRB** for active-voice verbs, **VRB-PASS** for passive-voice verbs, **PRT** for particles, which include prepositions and conjunctions, and **PNX** for punctuation marks.

Relations CATiB uses eight relations: **SBJ** (subjects of verbs and the topics of simple nominal sentences); **OBJ** (objects of verbs, prepositions, or deverbal nouns); **TPC** (topics of complex nominal sentences containing explicit pronominal referents); **PRD** (complements of the extended copular constructions); **IDF** (marking *Idafa*, the possessive nominal construction); **TMZ** (marking *tamyiz*, the *specification* nominal construction); **MOD** (general modification of verbs or nominals);

[7] https://www.hindawi.org/books/90463596/

[8] https://www.arabicbible.com/

[9] https://www.hindawi.org/books/72707304/


palmyra
CLEAR
CAMELTB_1001_NIGHTS_NIGHT_7_1
1/31
+TREE
-TREE
EDIT
DOWNLOAD
DIRECTION
TAGS
LISTING
rod with casserole girl turned fish spoke when *it* that auspicious king O' me reached : she-said seventh night in

في الليلة السابعة قالت : بلغ +ني أيها الملك السعيد أن +ه لما تكلم السمك قلبت الصبية الطاجن ب+ القضيب

Figure 1: An example of a CAMELTB syntactic analysis in the PALMYRA interface.The sentence translates as 'On the seventh night, she said: It has reached me, O auspicious King, that when the fish spoke, the young girl turned the casserole with the rod' (*The Tale of the Prince and the Ogress — The Book of the Arabian Nights*). The English glosses are aligned with the Arabic words and shown from right to left.

and, **—** (marking flat constructions such as first-last proper name sequences).

4.2. Guideline Extensions

Based on a pilot study we carried out before starting our annotation campaign, and as annotation progressed, we identified a number of issues that required specification or clarification in the CATiB guidelines, which were primarily created for news genre texts. This led to many extensions that became part of the guidelines manual for CAMELTB.[1] We highlight some of these issues below.

Foreign Tokens We extended the POS tag set by including a *FOREIGN* tag for non-Arabic script words. These appear occasionally in modern news texts.

Elided Tokens We extended the CATiB guidelines to allow the specification of elided tokens explicitly in the tree. This extension was especially needed for the older poetic texts. The previous guidelines allowed children of elided nodes to attach to the parent of the elided nodes with the same relation they would have to the elided node. In the new guidelines, the annotators are allowed to add an elided token and mark it with a (*) suffix. It should be noted that elided tokens were quite infrequent, accounting for 24 instances out of 242K tokens (1 in 10,000). 10 instances were in the **Quran** text, and 4 in the **Odes**.

New Constructions We clarified and extended the guidelines regarding first and second-person statements, interrogatives, interjections, so-called frozen verb constructions, and verse numbers in Holy Texts (**Quran**, **NT**, **OT**).

Sentence Segmentation To address the challenge of very long sentences, we defined guidelines for manual sentence segmentation. The challenge stems from the dearth of punctuation marks in general, and the dual use of the Arabic comma (،) for phrase and clause boundaries. We make use of simple automatic sentence segmentation using (!؟؛. .*;?!*), and define the guidelines around merging and splitting the automatically segmented units. Manual splitting is required only between two complete independent sentences that are not connectable. Splitting is never allowed before or after a dependent or incomplete clause. Manual merging took place after a punctuation mark that splits two parts of one sentence. We opted against splitting sentences in Holy Texts (**Quran**, **NT**, **OT**) and CA poetry (**Odes**) to respect verse boundaries; however, the guidelines specify that trees containing multiple sentences should link the sentences directly to the root, thus making the sentences sibling sub-trees.

Figure 1 presents an example CATiB syntactic tree from CAMELTB as it appears in the PALMYRA interface (Taji and Habash, 2020) used by the annotators.

5. The Treebanking Process

Our annotation process consisted of the following steps: (a) semi-automatic sentence segmentation, (b) automatic tokenization, POS tagging, and parsing, and (c) manual correction of tokenization, POS tagging, and parsing errors.

5.1. Semi-automatic Sentence Segmentation

After applying simple regular expressions to segment the text using a number of punctuation marks (!؟؛. .*;?!*), the lead annotator (fourth author) read every single line and merged and split sentences following the sentence segmentation guidelines.

5.2. Automatic Annotation

For tokenization and POS tagging, we used the open-source python library CamelTools (Obeid et al., 2020). We opted for CamelTools as opposed to Farasa (Abdelali et al., 2016) or MADAMIRA (Pasha et al., 2014), because CamelTools produces the CATiB-style tags among its large set of features, and gives us more control over the output forms.[10] For parsing we used the MALT parser model (Nivre et al., 2006) used within the CamelParser (Shahrour et al., 2016). For the Quran and ZAEBUC, we did not use automatic tokenization and POS tagging since we had access to gold tokenization and POS tags (Dukes and Habash, 2010; Habash and Palfreyman, 2022).

5.3. Manual Annotation

The manual annotation was done by four Arabic native speakers, with extensive experience in treebanking and/or linguistic training. The annotation was done using PALMYRA, a configurable platform independent graphical dependency tree visualization and editing software (Taji and Habash, 2020) (Figure 1). PALMYRA allows annotators to make corrections in tokenization in addition to POS tagging and dependencies.

5.4. CAMELTB 1.0

In the first release of the CAMELTB (version 1.0) we include the raw texts, their sentence-segmented versions, and the dependency trees in CoNLL-X style (Buchholz and Marsi, 2006). We also include recommended Train-Dev-Test document splits targeting roughly a 70-15-15 distribution at the word level. We tried when possible to follow the recommendations by Diab et al. (2013) for data divisions in experimental setups. Further details are included in the public release.[1]

6. Evaluation

In this section we present two evaluations of the annotation effort in order to validate its quality, and to quantify the number of changes needed from the initial automatic processes.

6.1. Metrics

We report our evaluations in terms of five metrics that consider the tokenization, POS, and dependency tree differences between gold and predicted trees.

Token Alignment We follow Habash (2010)'s distinction between tokenization and segmentation, where the latter refers to string breakup into smaller units without any modification, as opposed to the former which includes orthographic and morphological regularization. We do not evaluate on segmentation under this definition, but acknowledge that the term is sometimes used interchangeably. We take inspiration from previous efforts that addressed the challenge of evaluating tokenization and joint tokenization and tagging for morphologically rich languages where different word analyses can lead to different tokenizations (and segmentations) (Shao et al., 2018; More et al., 2019). In this effort, however, we faced a second more complex challenge, where new words (not sub-word tokens) may be introduced by editing the automatically annotated tree to (a) correct spelling errors through splitting, merging, or rewriting, or (b) add elided words. To address this issue, we utilize a word alignment technique that uses character-level edit-based alignment to align the characters, and then group them into words that minimize the overall edit distance (Khalifa et al., 2021; Belkebir and Habash, 2021). This alignment step maximizes pairing of tokens, including reasonable substitutions. Inserted and deleted tokens are paired with null tokens. Overall, there were 4,173 instances of inserted tokens (39%) and deleted tokens (61%) in the whole corpus (comparable in size to 1.73% of the total token count).

After acquiring the gold-predicted token alignments, we align the gold and predicted trees by inserting the null tokens as needed and adjusting the parent indices. We then proceed to the evaluation metrics.

- **Tokenization F-1 (TOK)** is calculated as the F-1 score of the precision and recall of correctly tokenized aligned tokens in a similar manner to Shao et al. (2018) and More et al. (2019).

For the rest of the metrics, unlike More et al. (2019), we do not consider inserted tokens in the predicted tree, and we evaluate on accuracy against the gold tree without reference to the token forms.[11] All predicted tree null tokens aligned with gold tree tokens, i.e. deleted tokens, are counted as errors.

- **POS Accuracy (POS)** is the percentage of gold tokens with correct POS.
- **Label Score (LS)** is the percentage of tokens with correct dependency labels.
- **Unlabeled Attachment Score (UAS)** is the percentage of tokens with correct dependency arcs (correct parent).
- **Labeled Attachment Score (LAS)** is the percentage of tokens with correct dependency labels and arcs.

All results include punctuation tokens. The results are not macro-averaged over the separate tree files for any

[10] One challenge with using MADAMIRA, which we tried first, was that it hallucinated back-off analyses for some CA and literary MSA words that were not in its lexicon. This made it harder for the annotators to correct such cases.

[11] The intuition for this decision comes from the experience of annotators who first correct tokens, then correct POS and dependency. If one thinks of the metric as measuring "distance" of transformation, then deleting tokens that are incorrect makes looking at their POS and dependency later meaningless. Of course this means our metric will produce different values if the gold and predicted trees are flipped, and there is a higher reliance on optimizing the initial alignment.

			Predicted TOK & POS					Gold TOK & POS		
Sub-Corpus	**#Words**	**#Tokens**	**TOK**	**POS**	**UAS**	**LS**	**LAS**	**UAS**	**LS**	**LAS**
1001	11,831	17,109	96.8	92.1	77.7	80.2	70.0	83.3	85.8	77.3
ALC	9,221	12,047	97.4	94.7	81.0	87.1	76.2	83.5	89.7	79.4
BTEC	15,935	18,602	95.3	94.5	81.3	84.7	75.6	85.4	89.1	80.9
Hadith	12,467	15,745	96.6	88.8	73.6	77.1	61.4	78.5	86.2	71.3
Hayy	19,674	26,583	98.5	95.3	76.8	84.1	70.7	80.4	87.6	75.5
NT	9,593	12,293	97.1	93.0	69.4	79.8	63.1	72.8	83.2	67.2
Odes	7,465	10,170	90.8	85.6	66.3	68.4	56.1	74.3	78.0	66.2
OT	9,097	11,788	96.4	89.1	71.3	80.9	65.0	77.6	86.0	72.1
QALB	11,454	14,139	98.8	94.8	75.3	84.5	70.5	77.5	87.0	73.6
Quran	11,699	15,791	*99.2	*98.8	70.9	76.2	64.6	71.3	76.6	65.0
Sara	35,356	46,375	97.3	94.9	72.2	82.6	66.6	75.6	86.3	71.1
WikiNews	18,314	21,481	99.0	93.2	83.6	90.9	79.7	86.1	92.7	82.5
ZAEBUC	15,778	19,787	*99.5	*99.0	81.9	90.1	79.0	82.3	90.8	79.7
Average	**14,453**	**18,608**	**97.1**	**93.4**	**75.5**	**82.0**	**69.1**	**79.1**	**86.1**	**74.0**
Total	**187,884**	**241,910**								

Table 2: Evaluation of Automatic tokenization, POS tagging, and parsing. For **Quran** and **ZAEBUC**, we started with gold tokenization and POS tags from previous projects.

sub-corpus. The code to these metrics is available from the CAMELTB website.[1]

6.2. Annotation Validation

To validate the quality of the annotations, we carefully checked and corrected a large sample of the text annotations ($\sim$8% of the full corpus). The TOK and POS scores are quite high, 99.9% and 99.7% on average, respectively. LS is 97.3% on average, and ranging from 99.0% (**ZAEBUC**) to 93.2% (**Quran**). UAS is 95.5% on average, and ranging from 98.7% (**ZAEBUC**) to 91.6% (**NT**). LAS is 94.5% on average, and ranging from 98.2% (**ZAEBUC**) to 90.2% (**NT**). For **NT**, over half of the disagreements involved PNX, and PRT.

We conducted additional rounds of quality checking where we automatically flagged possible error cases and shared them with the annotators. We plan to release further improved versions of the corpus in the future.

6.3. Automatic Parsing of Different Genres

Next we evaluate the quality of the automatic tokenization, POS tagging and parsing, which was given to the annotators. This can be seen as a measure of the amount of changes made by the annotators to the automatic parses they started. The results are in Table 2. For reference, the CamelParser's **PATB** parsing accuracy as reported by Shahrour et al. (2016) is 86.4%, 93.2%, snd 83.8%, for UAS, LS, and LAS, , respectively.

On average the TOK and POS scores are decent for this task at 97.1% and 93.4%, respectively. When we exclude the **Quran** and **ZAEBUC**, which have exceptionally high TOK and POS accuracies because the annotators were given previous annotations for tokenization and POS (Dukes and Habash, 2010; Habash and Palfreyman, 2022), the scores drop slightly to 96.7% and 92.4%. Despite that, there were some minor corrections in both.

Overall, the annotators had to change about one-fifth of all labels; and one-quarter of all parent attachments.

The best performance (in LAS) is on **WikiNews**, followed by **ZAEBUC** and **ALC**. The worst performance (in LAS) is on **Odes**, followed by **Hadith** and **NT**. The high degree of variation among the different genres suggests more research is needed in targeting specific genres.

In Table 2, we also include the UAS, LS and LAS results of parsing starting with gold tokenization and POS. Naturally the results are better. Excluding the **Quran** and **ZAEBUC**, the average increase is 4.2%, 4.6% and 5.6% for UAS, LS and LAS, respectively. These increases correspond to 17%, 26% and 18% error reduction rates, respectively.

7. Analysis of Genre Variations

The common annotation scheme and tokenized word representation in our data set allow us to study the genre variation in terms of syntactic and lexical similarity. What we present here is only an initial analysis. We leave further detailed structural comparisons to future work.

7.1. Syntactic Variations

Table 3 presents the distributions of the POS tags and relation labels per sub-corpus. We also include the respective distributions in the **PATB** training corpus. The sub-corpus rows are ordered by the degree of similarity to the **PATB** (as indicated by the correlation in the last column). Following are some interesting observations about this data. Cells connected with the discussed notes are highlighted in Table 3 for readability.

		POS							Relation								
Sub-Corpus	**Tok/Sen**	**NOM**	**PROP**	**VRB**	**VRB-PASS**	**PRT**	**PNX**	**FOR-EIGN**	**SBJ**	**OBJ**	**TPC**	**PRD**	**MOD**	**IDF**	**TMZ**	**---**	***R***
Hadith	13.2	28.1	16.2	19.1	0.7	24.0	11.9	0.0	9.7	28.5	0.1	2.2	38.8	6.1	0.1	14.6	*86.2*
1001	14.9	45.5	1.2	17.7	0.2	33.0	2.4	0.0	7.5	31.6	0.1	2.4	40.3	10.5	0.1	7.6	*93.0*
Odes	13.0	50.4	3.3	14.2	1.4	30.1	0.5	0.0	8.3	26.4	0.1	2.5	38.4	13.3	0.1	10.9	*94.9*
Quran	27.6	42.2	4.1	14.7	1.0	30.8	7.2	0.0	7.9	28.3	0.3	2.8	42.3	7.1	0.1	11.4	*94.9*
NT	21.5	39.7	4.6	14.0	0.9	26.8	14.1	0.0	7.4	26.7	0.2	2.7	45.3	7.2	0.1	10.4	*96.1*
Hayy	22.2	48.8	0.9	12.0	0.5	32.6	5.2	0.0	7.6	26.6	0.2	4.0	45.1	11.0	0.1	5.2	*96.9*
Sara	29.1	45.3	1.4	12.9	0.3	30.4	9.8	0.0	6.8	28.9	0.1	2.6	47.0	10.3	0.0	4.2	*97.1*
BTEC	9.3	49.5	2.1	11.0	0.4	21.7	15.4	0.0	7.5	21.2	0.2	2.0	43.4	11.5	0.3	14.0	*97.6*
ZAEBUC	17.8	54.7	2.0	8.9	0.3	27.6	6.5	0.0	7.4	24.5	0.4	2.7	43.2	14.4	0.1	7.3	*98.3*
OT	22.0	41.5	8.2	12.1	0.5	24.1	13.6	0.0	7.2	23.4	0.1	2.4	44.1	11.7	0.5	10.6	*98.4*
ALC	16.6	46.8	3.8	12.5	0.4	27.6	8.9	0.0	6.2	26.1	0.2	3.3	44.0	13.6	0.1	6.6	*98.4*
QALB	15.3	45.5	5.9	10.3	0.4	25.6	12.3	0.0	8.0	24.5	0.2	2.9	44.4	11.4	0.1	8.5	*98.9*
WikiNews	21.6	49.7	9.9	8.0	0.7	23.3	8.2	0.2	6.6	21.8	0.1	2.4	45.4	16.4	0.4	7.0	*99.7*
PATB	37.4	51.2	7.5	7.7	0.5	23.0	10.1	0.0	5.8	22.0	0.1	1.9	49.9	14.5	0.2	5.5	

Table 3: Variations among the different genre in terms of POS and relation labels. *R* is the Pearson correlation coefficient between each sub-corpus row of {POS, Relation} and the **PATB** row of {POS, Relation}. Highlighted items are discussed in the text.

Most and least similar to PATB As expected, **WikiNews** is the most similar to **PATB** (news genre); **QALB** (online commentaries on news) follows closely. The most different from **PATB** are **Hadith**, **1001**, and **Odes**, also unsurprising.

Proper name distribution **Hadith** has a very high ratio of PROP (16.2%) compared with **Hayy**'s (0.9%). This is most likely due to **Hadith** text's inclusion of the supporting transmittal record (*sanad*) which consists of series of names. As for **Hayy**, the vast majority of the book is about a young man, raised by an antelope, isolated from people, trying to make sense of the world. There are very few other named individuals in it.

Passive verb distribution The usage of VRB-PASS is much higher in **Odes** than in other genres. The usage of the passive verb is generally less in modern texts.

Punctuation distribution There is a lot of variation in the percentage of PNX. **BTEC** is the highest, possibly because it has the shortest lengths of sentences. The **Odes** have almost no PNX. For the **Quran**, **OT** and **NT**, verse number notation contributes to higher percentages of PNX.

Foreign word distribution Words written in a foreign script are not common; they mostly appear in **WikiNews** in reference to names of English movies and the like.

Predicate relation distribution The relatively higher ratio of the PRD relation in **Hayy** seems connected to the more than average use of the copular verb كان *kAn* 'to be'.

Idafa relation distribution The IDF relation also varies from 6-7% (**Hadith**, **Quran**, **NT**) to 16.4% (**WikiNews**). This seems to be connected with the heavier use of Idafa chains (sequences of possessives) in news text, e.g., رئيس مجلس إدارة شركة مايكروسوفت *rŷys mjls ǍdArħ šrkħ mAykrwswft* 'lit. Chairman *of* the board *of* the directing *of* the company *of* Microsoft / Microsoft's CEO'. The **Quran** maximally has Idafa chains of length 2, while **WikiNews** has 5, and **PATB** 6. In the **Quran**, single Idafa constructions constitute 95% of all Idafa constructions. The respective numbers for **WikiNews**, **ZAEBUC**, and **1001** are 74.5%, 80.7%, 87.8%. The correlation between the sub-corpus century (as listed in Table 1) and the average Idafa chain length is 70.5% — the older the text, the shorter the Idafa chain length. Another interesting related observation is that the percentage of pronominal clitics, which can only end Idafa chains, out of all words in Idafa relations varies widely from 56.8% and 54.7% for **1001** and the **Quran**, respectively down to 13.3% and 16.2% for **WikiNews** and **PATB**, respectively. The Idafa pronominal clitic ratio and sub-corpus century correlate at -67.4% — the older the text, the more likely it is to end with a pronominal clitic.

Flat relation distribution Both **Hadith** and **BTEC** have a higher than average use of the flat relation, but for different reasons. In **Hadith**, it is due to the larger number of multi-part PROP constructions. **BTEC**'s shorter sentences explain the higher ratio of flat relations, as they are used by default as the relation to the root (node 0).

A note on tokenization Finally, we note from Table 2 that the tokenization ratio (tokens/word) also 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.

7.2. Lexical Variations

Next, we consider the lexical variations among our sub-corpus genres. Figure 2 presents a dendrogram of lexical

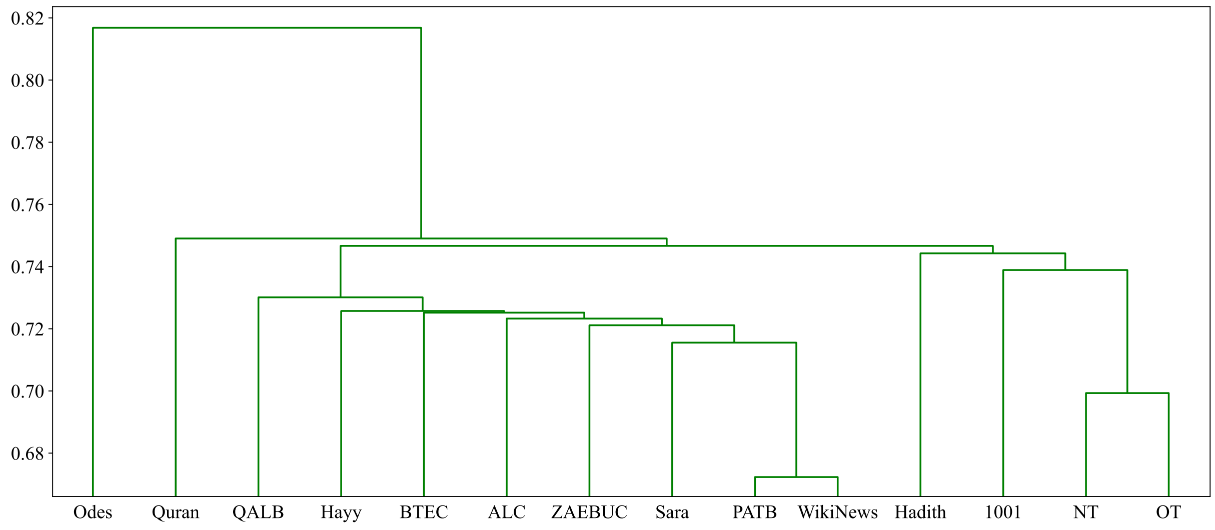


Figure 2: A dendrogram of cosine-based lexical dissimilarity among the 13 CAMELTB sub-corpora, and the **PATB**.

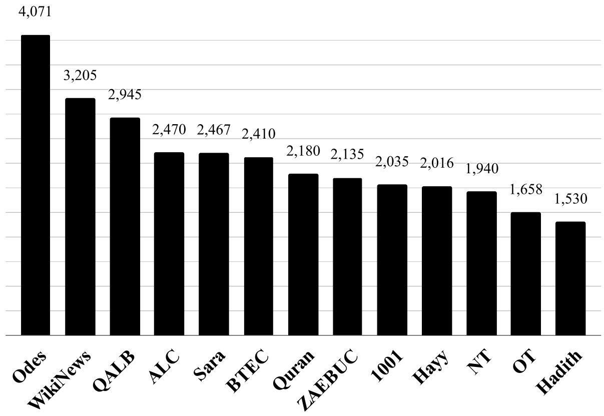


Figure 3: Number of unique types for the first 10,000 tokens in each of the 13 CAMELTB sub-corpora.

dissimilarity among the 13 CAMELTB sub-corpora as well as the **PATB**. To create this dendrogram, we use cosine similarity across token-based distributions over the union of all the token vocabulary in the studied corpora (49,661 unique tokens overall),[12] and perform hierarchical agglomerative clustering on the distributions.[13]

Unsurprisingly, **WikiNews** and **PATB** (news genre) cluster together, and so do **OT** and **NT** (religious text). Both clusters are distinctly different from the larger groups they are clustered with.

The **Odes** stands apart from all other sub-corpora, which is reasonable given that they are different in many ways: Classical Arabic and poetry. The **Quran** also stands separately as a unique genre in Arabic, but still closer to the rest of the other sub-corpora than the **Odes**.

All of the 20th and 21st century texts cluster together. Oddly **Hayy** (12th century) also clusters with them. **Hadith** and **1001** cluster with **OT** and **NT**, which may reflect their style and themes.

We also consider lexical diversity measured as the number of unique types in the same number of tokens (10,000 tokens to allow us to compare all of the sub-corpora). The **Odes** have the highest number (4,071), over 2.5 times the **Hadith** (1,530, the least diverse sub-corpus). See Figure 3. The variation most likely reflects the number of authors and topics as well as the genres.

8. Conclusion and Future Work

We presented CAMELTB, a large ∼188K word, ∼242K token manually annotated open-source dependency treebank of MSA and CA texts from different historical periods and genres. We presented some interesting insights about syntax and different genres in Arabic. We hope this will inspire others to explore this corpus and add to it.

In the future, we plan on extending CAMELTB with additional texts from other periods and genres. We also plan on using it to develop improved genre-aware parsing models (Müller-Eberstein et al., 2021). As parsing models improve in quality, we hope new pathways of research in digital humanities will make use of them. We also plan on extending the PALMYRA system to allow sentence merging and splitting within the same interface to minimize extra preprocessing steps.

Acknowledgments

The work on this project was funded by a New York University Abu Dhabi Research Enhancement Fund grant. We thank Ramy Eskander and the team of annotators at Ramitechs for their hard work on portions of CAMELTB. We also thank the anonymous reviewers for their helpful feedback.

[12] Excluding **PATB**, we have 26,695 unique tokens in CAMELTB 1.0.

[13] The matrix dimensions are 14 x 49,661, where the rows represent the sub-corpora, and the columns represent the unique tokens. The vector describes the existence of tokens in each sub-corpus. In other words, a 1 signifies that the given token exists in the sub-corpus, while a 0 signifies its absence.

9. Bibliographical References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 11–16, San Diego, California.

Al-Ghamdi, S., Al-Khalifa, H., and Al-Salman, A. (2021). A dependency treebank for classical Arabic poetry. In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 1–9, Sofia, Bulgaria, December.

Alfaifi, A. Y. G. (2015). *Building the Arabic Learner Corpus and a System for Arabic Error Annotation*. Ph.D. thesis, University of Leeds.

Alkuhlani, S., Habash, N., and Roth, R. (2013). Automatic morphological enrichment of a morphologically underspecified treebank. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 460–470, Atlanta, Georgia.

Belkebir, R. and Habash, N. (2021). Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online, November.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June.

Diab, M., Habash, N., Rambow, O., and Roth, R. (2013). LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Dukes, K. and Buckwalter, T. (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings of the Conference on Informatics and Systems (INFOS)*, Cairo, Egypt.

Dukes, K. and Habash, N. (2010). Morphological Annotation of Quranic Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Valetta, Malta.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2021). Starting a new treebank? Go SUD! Theoretical and practical benefits of the Surface-Syntactic distributional approach. In *Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, Sofia, Bulgaria.

Habash, N. and Palfreyman, D. (2022). ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Habash, N. and Roth, R. (2009). CATiB: The Columbia Arabic Treebank. In *Proceedings of the Joint Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 221–224, Suntec, Singapore.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Habash, N., Faraj, R., and Roth, R. (2009). Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Halabi, D., Fayyoumi, E., and Awajan, A. (2021). I3rab: A new Arabic dependency treebank based on Arabic grammatical theory. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–32.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual), April.

Khalifa, S., Obeid, O., and Habash, N. (2021). Character Edit Distance Based Word Alignment. `https://github.com/CAMeL-Lab/ced_word_alignment`.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and Using a Pilot Dialectal Arabic Treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Bies, A., and Kulick, S. (2010). Ldc standard Arabic morphological analyzer (sama) version 3.1.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., and Ciul, M. (2012). Developing aret: an nlp-based educational tool set for Arabic reading enhancement. In *Proceedings of the Workshop on Building Educational Applications using NLP (BEA)*, pages 127–135.

More, A., Seker, A., Basmova, V., and Tsarfaty, R. (2019). Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from Modern Hebrew. *Transactions of the Association for Computational Linguistics*, 7:33–48.

Müller-Eberstein, M., van der Goot, R., and Plank, B. (2021). Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic, November.

Nivre, J., Hall, J., and Nilsson, J. (2006). MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Language Resources and*

Evaluation Conference (LREC), pages 2216–2219, Genoa, Italy.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Shahrour, A., Khalifa, S., Taji, D., and Habash, N. (2016). CamelParser: A system for Arabic syntactic analysis and morphological disambiguation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 228–232.

Shao, Y., Hardmeier, C., and Nivre, J. (2018). Universal word segmentation: Implementation and interpretation. *Transactions of the Association for Computational Linguistics*, 6:421–435.

Smrž, O., Šnaidauf, J., and Zemánek, P. (2002). Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. In *Proceedings of the International Symposium on Processing of Arabic*, pages 147–155, Manouba, Tunisia.

Taji, D. and Habash, N. (2020). PALMYRA 2.0: A configurable multilingual platform independent tool for morphology and syntax annotation. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 168–177, Barcelona, Spain (Online), December.

Taji, D., Habash, N., and Zeman, D. (2017). Universal dependencies for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Valencia, Spain.

Taji, D., El Gizuli, J., and Habash, N. (2018). An Arabic dependency treebank in the travel domain. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, Miyazaki, Japan.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

10. Language Resource References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 11–16, San Diego, California.

Al-Akkad, A. M. (1938). *Sarah*. Hindawi.

al Bukhari, I. M. (846). *Sahih al-Bukhari*. Dar Ibn Khathir.

Alfaifi, A. Y. G. (2015). *Building the Arabic Learner Corpus and a System for Arabic Error Annotation*. Ph.D. thesis, University of Leeds.

Altammami, S., Atwell, E., and Alsalka, A. (2019). The arabic–english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology-IJASAT*.

Dukes, K., Atwell, E., and Habash, N. (2013). Supervised collaboration for syntactic annotation of quranic arabic. *Language resources and evaluation*, 47(1):33–62.

Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Habash, N. and Palfreyman, D. (2022). ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 39–47, Doha, Qatar.

Smith, E. and Van Dyck, C. (1860). *New Testament (Arabic Translation)*.

Smith, E. and Van Dyck, C. (1865). *Old Testament (Arabic Translation)*.

Taji, D., El Gizuli, J., and Habash, N. (2018). An Arabic dependency treebank in the travel domain. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, Miyazaki, Japan.

Takezawa, T., Kikui, G., Mizushima, M., and Sumita, E. (2007). Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Tufail, I. (1150). *Hayy ibn Yaqdhan*. Hindawi.

Unknown. (12th century). *One Thousand and One Nights*.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.