ZAE BUC: An Annotated Arabic-English Bilingual Writer Corpus

Nizar Habash† and David Palfreyman‡
†New York University Abu Dhabi, Abu Dhabi, UAE, ‡Zayed University, Abu Dhabi, UAE
nizar.habash@nyu.edu, david.palfreyman@zu.ac.ae

Abstract
We present ZAE BUC, an annotated Arabic-English bilingual writer corpus comprising short essays by first-year university students at Zayed University in the United Arab Emirates. We describe and discuss the various guidelines and pipeline processes we followed to create the annotations and quality check them. The annotations include spelling and grammar correction, morphological tokenization, Part-of-Speech tagging, lemmatization, and Common European Framework of Reference (CEFR) ratings. All of the annotations are done on Arabic and English texts using consistent guidelines as much as possible, with tracked alignments among the different annotations, and to the original raw texts. For morphological tokenization, POS tagging, and lemmatization, we use existing automatic annotation tools followed by manual correction. We also present various measurements and correlations with preliminary insights drawn from the data and annotations. The publicly available ZAE BUC corpus and its annotations are intended to be the stepping stones for additional annotations.

Keywords: Annotated Corpus, Learner Corpus, CEFR, Arabic, English

1. Introduction

Over half the world’s population is estimated to use more than one language every day (Grosjean, 2010); however, language corpora in general tend to focus on specific languages rather than on bilingual writers. Even research on ‘learner corpora’ of writing in English (or another language) tends to compare this writing with a corpus of writing by other, ‘native’ users of the same language.

In this paper, we discuss the development of a new kind of corpus, which focuses on a large set of bilingual writers, and comprises samples of their writing in both languages. The Zayed University Arabic-English Bilingual Undergraduate Corpus (ZAEBUC†) is not a ‘parallel corpus’ of texts with their translations. Instead, ZAEBUC matches comparable texts in different languages written by the same writer on different occasions. The corpus comprises short essays written by 397 first-year university students at Zayed University (ZU) in the United Arab Emirates totaling 388 English essays (87.6K words) and 214 Arabic essays (33.3K words).

We enrich the corpus with a number of layered annotations: (a) anonymized meta-data indicating extralinguistic features of the writers and texts; (b) manually corrected versions of the raw text; (c) automatic and manual annotations to identify morphological tokens, part-of-speech (POS), and lemmas; and (d) writing proficiency ratings using the Common European Framework of Reference (CEFR) (Council of Europe, 2001). The ZAEBUC dataset is an open, publicly available† and extendable research resource, designed with the intention to support empirically driven research in Arabic, English and bilingual development, as well as research and system development in natural language processing (NLP).

The paper is organized as follows: Section 2 presents some related work and background; Section 3 provides an overview of the approach we take in collecting and annotating the data; Section 4 details the data collection process; and Sections 5, 6, 7 present CEFR rating, text correction, and morphological annotations, respectively.

2. Related Work

We present in this section a brief discussion of some the relevant previous research efforts.

Corpora There is an increasing number of text corpora that come in different genres, sizes, and degrees and types of annotations, targeting different tasks. Parallel corpora comprise texts matched with their translations, and are the backbone of machine-learning based approaches to machine translation (Baisa et al., 2016; Tan and Bond, 2011; Rafalovitch et al., 2009; Koehn, 2005; Tiedemann, 2012). Learner corpora are used to study errors and other features of learner production, by comparison with ‘native’ user corpora and/or between sub-corpora produced by learners of different L1’s learning the same target language (Nicholls, 2003; Lee and Chen, 2009; Allafi, 2015). Orthogonally, annotated corpora are typically smaller parts of existing corpora enriched with linguistic annotations, such as manual corrections of spelling and grammar (Dahlmeier et al., 2013; Zaghouani et al., 2014), morphological, syntactic and semantic analyses (Marcus et al., 1993; Maamouri et al., 2004; Pradhan et al., 2007; Nivre et al., 2017), and others.

Our bilingual, writer-matched ZAEBUC corpus stands in contrast to these types of corpora, which are produced by different writers (about whom little is known), and which support research questions about (one or more) languages in general terms, rather than about bilingual writers. ZAEBUC is one of the first writer-matched bilingual corpora (see also Ströbel et al. (2020), and Meunier et al. (2020)), which aim to be representative of
of bilingual writers writing in both their languages, one as natives and one as learners. In the present case, the languages are Arabic and English; to our knowledge, ZAEBUC is the first of its kind for this language pair. Bilingual writer corpora allow researchers to relate L1 with L2 writing in a large number of individuals, in order to address a range of research questions about individual and group variation within and across languages – language dominance, competition or interdependence between individuals’ languages, and multicompetence (Cook, 2016). These issues have begun to be explored cross-linguistically at various linguistic levels, including spelling, vocabulary, grammar and discourse. ZAEBUC provides a resource for such research, drawing on established annotations, guidelines and tools developed with traditional corpora, but within a bilingual frame of reference.

**3.1. Corpus Design and Desiderata**

In designing ZAEBUC, we had the following desiderata in mind.

- **Rich and multilayered annotations** We want to have a corpus of essays written by a cohort of students (to control for variability) in two languages (Arabic and English), ideally with two texts from each writer. We want as many as possible non-private meta-data features associated with these texts, e.g., text topic, writer gender, and language of schooling. And we want text corrections and morphological annotations, all created using comparable principles for the two languages to allow for comparative analysis.

- **High quality annotations** We want the annotations to be done carefully by professional annotators, not crowd sourcing, and with sufficient careful inter-annotator checks to control for quality.

- **Ethical considerations** We want the corpus to be created ethically: consent from the writers is required to include the texts; and any personal information in the texts is redacted.

- **Wide usability** We want ZAEBUC to be usable widely across many communities: researchers in education, sociology and sociolinguistics, as well as NLP researchers and developers. To that end, we want to use accepted tried-and-true conventions and formats.

- **Openness** We want ZAEBUC to be an open resource, available publicly for researchers to use and annotate themselves, with minimal restrictions.

**3.2. Collection and Annotation Processes**

Our corpus creation process consisted of four steps. First was *data collection*, which involved getting approval from the IRB board on ZU campus to collect the data, then contacting and coordinating with the faculty who led the courses we targeted. All students who participated were asked to provide written consent to release their data. The second and third steps happened in parallel independently: *CEFR annotation* and manual *text correction*. The last step is *morphological annotation*, which depends on the output of text correction. This last step was done semi-automatically to increase the efficiency of the annotation and reduce its total cost: automatic annotations were followed by manual corrections.

We worked with a professional data annotation agency, Ramitechs, which employs professional linguists with compatible skillsets and training for the tasks of text correction and morphology annotation. There were three annotators on the Arabic tasks (two linguists and a translator; all native); and three annotators on the English tasks (a teacher, a translator and a linguist; all native, and two bilinguals who speak Arabic). Quality checks were done on a weekly basis to spot inconsistencies and errors, and to update guidelines and educate the annotators about any problematic issues. As for CEFR ratings, we worked with three bilingual researchers who specialize in CEFR rating. All texts were annotated in triplicate.

**3. Approach**

In this section we discuss ZAEBUC’s design and the processes we followed for its collection and annotation.

**3.1. Corpus Design and Desiderata**

In designing ZAEBUC, we had the following desiderata in mind.

---

[https://www.ramitechs.com/](https://www.ramitechs.com/)
Table 1: The prompts given to the essay writers. We pair these here for presentation purposes, but they were used independently for Arabic and English.

| Topic                  | Prompt                                                                 |
|------------------------|------------------------------------------------------------------------|
| Social Media           | How do social media affect individuals and society?                    |
| Tolerance              | كيف نعزز فضيلة التسامح في المجتمع؟                                      |
| Development            | التطوير الاقتصادي الذي تشهده دولة الإمارات العربية المتحدة              |

The meta-data we kept for all the texts include: anonymous student id, school type (government, private, other), language of schooling (Arabic, English, other), city/town of residence, gender, course (ENG 140, ARA 030, ARA 130), chosen topic, date of writing exam, length of exam, and number of days (positive or negative) from their Arabic to their English exam. Table 2 presents some corpus statistics. The following are some of the basic observations: The vast majority of the student contributing to the corpus are females (≈90%). This is consistent with the percentage of female students at ZU. Out of the 397 students, almost all contributed to the English sub-corpus, and about half contributed texts in both English and Arabic. About two-thirds of the 602 texts in the corpus are in English, and the rest in Arabic. Among the Arabic texts, 93% came from the main Arabic course (ARA 130). Finally, Social Media was the most popular topic by far: 80% in Arabic and 85% in English.

5. CEFR Annotation

The Common European Framework of Reference for Languages (CEFR) is a framework that was published in 2001 by the Council of Europe to describe language learners’ ability in terms of speaking, reading, listening and writing (Council of Europe, 2001). CEFR provides detailed descriptions to classify users according to six ranked levels (A1, A2, B1, B2, C1 and C2) from A1 (Beginner), to C2 (Proficient).

5.1. Annotation Process

Each ZAEBUC text was rated independently by three CEFR-proficient bilingual speakers (Arabic and English), who provided both a CEFR level and a comment to support their assessment. To allow us to average the CEFR levels and compare them in a fine-grained manner, we map the levels to numerical scores such that A1=1, A2=2, B1=3, B2=4, C1=5, and C2=6. The averaged scores are then rounded and converted to corresponding CEFR levels. For instance, if a text received A2, A2, and B1 ratings by our three annotators, the average score is (2+2+3)/3 or 2.33, and the rounded average score is 2.0 which maps to CEFR level A2. Thus, a level difference of 1.0 is equal to the difference between CEFR A1 and A2, or CEFR B2 and C1. If a text is
Social media is a widely controversial subject with various opinions regarding its negative and positive aspects. While social media has made positive impacts on society, it can also impart many negative changes on people worldwide. Social media is widely used as a means of communication between people. Social media lifts boundaries made by distance and creates roads that can easily connect people to one another. However, if not used correctly, those roads can lead to dead ends and cause individuals more harm than good. As the use of social media increased, studies have shown a simultaneous increase of false information or news being spread by people. This can cause harm to societies due to the lack of trust between people. False information or rumors being spread can also stir alliances between countries. Also, bullying on the internet has become much easier to do since it can be anonymously done. Since social media allows people to wear masks, it can be offensive and it affects people because they get influenced by you from different cultures and who has different religions, the content may be offensive. And it affects people because they get influenced by things that are wrong.

A lot of people argue whether social media affect individuals and society or not. Does it? Yes it does both in negative and positive ways. Social media is like a free playground where everyone can say, share, post whatever they like with not much restrictions. Well this can be useful and be beneficial but sometimes it is harmful and negative towards specific groups of people like kids and people who are maybe from different cultures and who has different religions, the content may be offensive and it affects people because they get influenced by you or try to copy what your doing whether it is good or bad.

In my opinion think social media has been the most important thing to everyone. Everyone uses it in the whole part of the earth. It also has a lot of benefits in it, for example knowing about the news and how everything is going on and its also easier for everyone because people had to go out to buy some newspapers and it takes a lot of time and probably half of the people were lazy. What is also good about it is the you also can know news about from the other countries yes also can be at the newspaper but maybe they take few days to write it down and from your phone or laptop you can know it in just one second. One of the best benefits of ghie social media is the you can call people from different countries and that really good and helpful. Best part of that is you can do.

In Arabic Example

C1

Social media is a widely controversial subject with various opinions regarding its negative and positive aspects. While social media has made positive impacts on society, it can also impart many negative changes on people worldwide. Social media is widely used as a means of communication between people. Social media lifts boundaries made by distance and creates roads that can easily connect people to one another. However, if not used correctly, those roads can lead to dead ends and cause individuals more harm than good. As the use of social media increased, studies have shown a simultaneous increase of false information or news being spread by people. This can cause harm to societies due to the lack of trust between people. False information or rumors being spread can also stir alliances between countries. Also, bullying on the internet has become much easier to do since it can be anonymously done. Since social media allows people to wear masks, it can be offensive and it affects people because they get influenced by you from different cultures and who has different religions, the content may be offensive. And it affects people because they get influenced by things that are wrong.

A lot of people argue whether social media affect individuals and society or not. Does it? Yes it does both in negative and positive ways. Social media is like a free playground where everyone can say, share, post whatever they like with not much restrictions. Well this can be useful and be beneficial but sometimes it is harmful and negative towards specific groups of people like kids and people who are maybe from different cultures and who has different religions, the content may be offensive and it affects people because they get influenced by you or try to copy what your doing whether it is good or bad.

In my opinion think social media has been the most important thing to everyone. Everyone uses it in the whole part of the earth. It also has a lot of benefits in it, for example knowing about the news and how everything is going on and its also easier for everyone because people had to go out to buy some newspapers and it takes a lot of time and probably half of the people were lazy. What is also good about it is the you also can know news about from the other countries yes also can be at the newspaper but maybe they take few days to write it down and from your phone or laptop you can know it in just one second. One of the best benefits of ghie social media is the you can call people from different countries and that really good and helpful. Best part of that is you can do.

B2

A. The prevalence of social media in the Arabic and English language in the past two decades has been enormous. For example, the term "social media" was introduced in Arabic in the early 2000s. In English, the term gained widespread usage in the late 2000s. This increase in usage is largely due to the rise of social media platforms such as Facebook, Twitter, and Instagram.

B. Social media has become an integral part of daily life for many people. People use social media to communicate with friends and family, to share news and information, and to participate in online communities. However, the use of social media has also led to some negative consequences, such as the spread of false information and the rise of cyberbullying.

C1

Social media is a widely controversial subject with various opinions regarding its negative and positive aspects. While social media has made positive impacts on society, it can also impart many negative changes on people worldwide. Social media is widely used as a means of communication between people. Social media lifts boundaries made by distance and creates roads that can easily connect people to one another. However, if not used correctly, those roads can lead to dead ends and cause individuals more harm than good. As the use of social media increased, studies have shown a simultaneous increase of false information or news being spread by people. This can cause harm to societies due to the lack of trust between people. False information or rumors being spread can also stir alliances between countries. Also, bullying on the internet has become much easier to do since it can be anonymously done. Since social media allows people to wear masks, it can be offensive and it affects people because they get influenced by you from different cultures and who has different religions, the content may be offensive. And it affects people because they get influenced by things that are wrong.

A lot of people argue whether social media affect individuals and society or not. Does it? Yes it does both in negative and positive ways. Social media is like a free playground where everyone can say, share, post whatever they like with not much restrictions. Well this can be useful and be beneficial but sometimes it is harmful and negative towards specific groups of people like kids and people who are maybe from different cultures and who has different religions, the content may be offensive and it affects people because they get influenced by you or try to copy what your doing whether it is good or bad.

In my opinion think social media has been the most important thing to everyone. Everyone uses it in the whole part of the earth. It also has a lot of benefits in it, for example knowing about the news and how everything is going on and its also easier for everyone because people had to go out to buy some newspapers and it takes a lot of time and probably half of the people were lazy. What is also good about it is the you also can know news about from the other countries yes also can be at the newspaper but maybe they take few days to write it down and from your phone or laptop you can know it in just one second. One of the best benefits of ghie social media is the you can call people from different countries and that really good and helpful. Best part of that is you can do.

Figure 1: Examples of essays and their CEFR levels.
### Table 3: ZAEBUC CEFR level distributions.

| Level       | Arabic | English |
|-------------|--------|---------|
| Advanced    | C1     | 5%      | 3%      |
| Upper Inter | B2     | 37%     | 21%     |
| Intermediate| B1     | 51%     | 50%     |
| Pre Inter   | A2     | 3%      | 24%     |
| Beginner    | A1     | 0%      | 2%      |
| Unassessable|        | 3%      | 0%      |

fuzzy match is 16/36 (44%), leading to a Cohen’s Kappa of 0.84 (almost perfect) for Arabic and 0.73 (substantial) for English. It is clear that the CEFR assignment task is hard, but the raters were quite close to each other, around one level of difference on average.

### 5.3. CEFR Level and Corpus Variables

Table 4 presents the average CEFR scores (i.e., average over the average rater scores per text) across different corpus variables. The columns show the scores by specific text languages, and for all texts. We exclude all texts with unassessable CEFR. The following are some of the basic observations about this corpus. First, the average CEFR level for all texts, Arabic, and English, is B1 (3.0 rounded average, Intermediate). However, the average CEFR score for Arabic texts is 3.5 as opposed to English 2.9, a difference of half a CEFR level. The difference is statistically significant at p<.001 using a two-tailed paired T-test on the paired texts by the writers except for closed-class terms such as prepositions, pronouns and articles, as well as correcting the use of punctuation marks.

The texts were edited directly by the annotators in Google Docs to create a parallel spelling and grammar corrected version of the texts. Automatic character and word-level alignment was then used to pair raw words with their corrections. Table 5 (columns Raw, Corrected and Edit) exemplify the results of this process for an Arabic text and an English text, respectively.

### 6. Text Correction

We present next the text correction guidelines and process we followed, statistics and observations, and a discussion of the different classes of errors.

#### 6.1. Annotation Guidelines and Process

For spelling and grammar correction, we followed a set of guidelines inspired by Zaghouani et al. (2014) for Arabic, and Dahlmeier et al. (2013) for English. The instructions provided to the annotators who did the correction specifically required that they focus on spelling correction and grammatically informed changes such as proper inflection in context. The annotators were instructed to avoid changing the lexical choices made by the writers except for closed-class terms such as prepositions, pronouns and articles, as well as correcting the use of punctuation marks.

The instructions provided to the annotators who did the correction specifically required that they focus on spelling correction and grammatically informed changes such as proper inflection in context. The annotators were instructed to avoid changing the lexical choices made by the writers except for closed-class terms such as prepositions, pronouns and articles, as well as correcting the use of punctuation marks.

The texts were edited directly by the annotators in Google Docs to create a parallel spelling and grammar corrected version of the texts. Automatic character and word-level alignment was then used to pair raw words with their corrections. Table 5 (columns Raw, Corrected and Edit) exemplify the results of this process for an Arabic text and an English text, respectively.

**Inter-annotator Agreement** We calculated the text correction inter-annotator agreement scores using two corrected versions of 26 pairs of texts in English and in Arabic. For Arabic, the Dice Similarity Coefficient between the two corrections is 97.1%, and for English, it is 96.7%. The vast majority of differences, 95.6% in Arabic and 92.8% in English, are non-erroneous disagreements, such as punctuation choice, or valid but unnecessary corrections. These results give us confidence in the correction quality.

#### 6.2. General Statistics and Observations

Table 6 (b,c,d) summarizes the high-level spelling and grammar correction patterns. The Arabic text average word count is about two-thirds the English text word count. This is most likely connected to Arabic’s morphology and orthography: Arabic is a pro-drop language, with no indefinite articles, and numerous cliticized particles and pronouns. Corrections to English hardly affect the total word count, whereas in Arabic we see a drop of about 5% in word count. Finally, the
Table 5: Two examples of B1 (CEFR) text segments in English and Arabic. Examples align the raw sentences with their corrections, marked edit points, white-space tokens (WS Tokens), morphological tokens (M Tokens), parts-of-speech (POS) and lemmas. The Arabic and English sentences are not parallel.

| Raw | Corrected | Edit | WS Tokens | M Tokens | POS | Lemma |
|-----|-----------|------|-----------|----------|-----|-------|
| the | The       |      | DET       | NOUN     |     |       |
| social | social | EDIT | ADJ       | the       |     |       |
| media | media |      | NOUN      | media     |     |       |
| didn't | didn't | do+not | AUX+ADV | in       | ADV | in    |
| affect | affect |       | VERB     | affect   |     |       |
| one | one |       | NUM       | one       |     |       |
| country | country |      | NOUN      | country   |     |       |
| or | or |       | CONJ  or  |          |     |       |
| specific | specific | EDIT | ADJ+specific |          |     |       |
| group | group | EDIT | NOUN+group |          |     |       |
| of | of | EDIT | ADP       | of        |     |       |
| people, | people, | EDIT | people    | people   |     |       |

| Arabic | English |
|--------|---------|
| 214    | 388     |

Table 6: Corpus Statistics for Arabic and English texts. The edit percentages in section (d) are calculated against the Raw text total count.

| Arabic | English |
|--------|---------|
| 33,376 | 87,602  |
| 156    | 226     |
| 31,661 | 87,621  |
| 148    | 226     |
| 34,235 | 97,478  |
| 160    | 251     |
| 42,927 | 98,452  |
| 201    | 254     |
| 51,609 | 241     |

6.3. Text Error Analysis

We conducted a detailed manual error analysis in 10 randomly selected texts in each language.

**Arabic Errors** The most common error type, occurring 28.9% of the time, had to do with the spelling of the Hamza (glottal stop), which can be spelled in seven ways, depending on phonological context and morphological derivation. It is not particularly surprising to see this error. The next very common error in this data set (28.7%) is the incorrect separated spelling of the conjunction wa ‘and’. This error is responsible for two-thirds of all DELETE edits and one-sixth of all EDIT errors, as it involves a DELETE of wa, and an edit of the word it cliticizes to. Punctuation errors are about one-sixth of all errors. The next error (8.9%) is the misspelling of the feminine ending Ta Marbuta (ئ) without its dots (أ) – a common spelling error (Zaghouani et al., 2014). Other typos account for almost 8% of all errors. Many of these are the result of dialectal pronunciation. Errors involving morphological case, state, gender, and feature agreement, are infrequent.

**English Errors** The most common error type in the English texts involves punctuation marks (31.2%). These errors are twice as common in English as they are in Arabic. Misspellings (e.g., around for around) are responsible for one-sixth of all errors. English has more grammar and morphology errors than Arabic, which makes sense given that it is the students’ second language. Some English-specific phenomena are not possible to consider in Arabic such as capitalization (6.7% of all errors).

It is rather hard to compare the errors between English and Arabic as they are the result of different linguistic phenomena (e.g., verb agreement or determiner use) and orthographic rules (e.g., Hamza spelling or capitalization). One interesting aspect is that while the percentage of exact matches in Arabic is much lower...
than in English, the average CEFR is higher in Arabic than English. Many of the Arabic errors (Hamza, Wa spelling and even Ta Marbuta) may be the results of shallow orthographic technicalities that do not affect readability or understanding; in fact many of these errors are widely tolerated, even in public signage. This, together with Arabic’s more compressed spelling, resulting in a lower total word count, may be inflating the ratio of errors overall.

7. Morphological Annotation

In this section, we present our morphological annotation guidelines and process, as well as some general statistics and observations.

7.1. Annotation Guidelines

Our final set of annotations focused on morphological tokenization, part-of-speech (POS) tagging and lemmatization. For tokenization and POS tagging, we followed the guidelines of the Universal Dependency (UD) project (Nivre et al., 2017).

Tokenization

UD follows the morphological tokenization choices made in the PTB (Marcus et al., 1993) for English and PATB for Arabic (Maamouri et al., 2004). For English, this includes separating contractions such as can’t into can+not. For Arabic, all clitics are separated except for the definite article, e.g., the word ُوَالْقَمْرِ ‘and like the moon’ is tokenized as وَا لْقَمْرِ. For Arabic, all clitics are separated except for the definite article, e.g., the word ُوَالْقَمْرِ ‘and like the moon’ is tokenized as وَا لْقَمْرِ.

POS

UD defines 17 POS categories: Open class (ADJ, ADV, INTJ, NOUN, PROPN, VERB), Closed class ADP (adposition), AUX, CCONJ, DET, NUM, PART, PRON, SCONJ and other (PUNCT, SYM, X ‘other/unclassifiable’). We made extensive use of the UD guidelines, PTB and PATB guidelines.

Lemmatization

The lemma is an abstraction that represents the various inflectional forms of a particular lexical item with a specific derivation and POS. For example, the English verb forms eat, eats, eating, eaten are all lemmatized to eat; and the Arabic verb forms َكَتَبَ ‘he wrote’, َسَيَكِبْتَ ‘he will write it’, and َوَلْنَكِبَ ‘we will write’ are all lemmatized to َكَتَب. For Arabic, we use undiacritized lemmas, which were easier and cheaper to annotate.

7.2. Annotation Process

Automatic Annotation

All three annotation aspects were automatically produced using the corrected text version of our corpus. All texts were automatically white-space-and-punctuation tokenized. For English, we used Stanza (Qi et al., 2020) to generate an initial version of the tokenizations, POS, and lemmatization.

For Arabic, we used Madamira (Pasha et al., 2014) to do the same. The Madamira POS tagset was mapped (many to one) to the UD tagset, as was done by Taji et al. (2017).

Manual Annotation Correction

Three annotators then went through the full automatically annotated corpus and manually corrected it. The effort was completed on Google Sheets. For English, the accuracy of the automatic process, measured on the full corpus, was 96.9%, 95.1%, and 96.7% for tokenization, POS tagging and lemmatization, respectively. The Arabic accuracies were 99.5%, 90.2%, and 93.6%, respectively. The lower results for Arabic are not surprising given the higher degree of complexity and ambiguity in Arabic. The automatic processes produced very good starting points for the manual correction task, which helped its efficiency. We expect that the topics and genre (university-level essays) helped a lot in having a strong automatic starting point since the tools we used were mostly trained on news text with similar style to the essays.

Inter-annotator Agreement

We calculated inter-annotator agreement using 26 texts for English (averaging 196 words/text) and another 26 documents for Arabic (averaging 120 words/text). The results from two annotators were compared. In English the degree of inter-annotator agreement is 99.98%, 99.57%, and 99.86% for tokenization, POS, and lemmatization, respectively. In Arabic, the respective inter-annotator agreement figures are 99.94%, 98.11%, and 99.68%. These are very high levels of agreement.

The last four columns (WS Tokens, M Tokens, POS, and Lemma) in Table 5 exemplify the results of the morphological annotation process. WS Tokens refer to white-space-and-punctuation tokenization results, whereas M Tokens refer to morphological tokenization results.

7.3. General Statistics and Observations

Tokenization

There is a noticeable difference in the number of words per text between Arabic and English texts, where Arabic texts had around 69% of the number of raw words in English texts on average. The numbers became lower (66%) once corrections were made. In the white-space tokenization versions of the texts, where punctuation is separated from words, the ratio of Arabic to English becomes even smaller (64%), which is consistent with the higher use of punctuation in English compared to Arabic. However, in terms of morphological tokenization, where Arabic token count increases by about 25%, Arabic tokens are almost 80% of English’s comparable count. If we split the Arabic definite article َلْا لْثِ ‘the’, which occurs in 20% of all words, the difference between English and Arabic in word count diminishes to less than 5%; see Table 5 (g).

Parts-of-Speech

The number of unique POS tags in tokenized words is 91 and 29 for Arabic and English, respectively. The tokenized words’ tags are 17 for both languages, of course. Arabic clitics, e.g.,

For more details on Arabic computational morphology, see Habash (2010).
**Lemma Clouds** Figure 2 contrasts the lemma clouds of the most frequent 100 verb, noun, adjective, or adverb lemmas in the Arabic and English texts. These lemma account for 56% and 62% of all verbs, nouns, adjectives, and adverbs in the Arabic and English texts, respectively. The figure’s caption includes the top ten Arabic and English lemmas. It is not surprising that lemmas related to social media, communication, society, and individuals dominate since the topic of social media had the lion’s share among the student’s essay topics.

8. **Conclusion and Future Work**

We presented in detail the process of collecting and annotating the ZAEBUC corpus, a writer-matched bilingual corpus. We discussed the various meta-data and annotations we provided in qualitative and quantitative terms. We also presented a host of insights and notable patterns that (we hope) will encourage and excite other researchers to use and extend ZAEBUC.

In the future we plan to extend ZAEBUC in a number of directions: adding full syntactic representations, adding deeper morphological features such as person, gender, and number, and doing a follow up round of essay collections from the same students in a later stage of their university experience to support diachronic analysis. All of the corpus and its annotations are publicly available at [http://www.zaebuc.org](http://www.zaebuc.org).

**Acknowledgments**

The ZAEBUC project was funded by a Zayed University Research Incentive Fund (RIF #R19068). We would like to thank Ramy Eskander for helpful discussions and his work as part of Ramitech.
9. Bibliographical References

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

Baisa, V., Michelfeit, J., Medved’, M., and Jakubíček, M. (2016). European union language resources in Sketch Engine. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 2799–2803.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wiesniewski, K., Abel, A., Schöne, K., Stúdlová, B., and Vettori, C. (2014). The MERLIN corpus: Language learner and the CEFR. In LREC, pages 1281–1288. Reykjavik, Iceland.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychosocial Measurement, 20:37–46.

Cook, V. (2016). Premises of multi-competence. The Cambridge handbook of linguistic multi-competence, pages 1–25.

Council of Europe, C. o. E. (2001). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Grosjean, F. (2010). Bilingual. Harvard university press.

Habash, N., AbuOdeh, M., Taji, D., Faraj, R., Gizuli, J. E., and Kallas, O. (2022). Camel Treebank: An open multi-genre Arabic dependency treebank. In Proceedings of the Language Resources and Evaluation Conference (LREC), Marseille, France.

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

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Machine Translation Summit (MT Summit), pages 79–86, Phuket, Thailand. AAMT.

Lee, D. Y. and Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by chinese learners. Journal of Second Language Writing, 18(4):281–296.

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.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. Computational Linguistics, 19(2):313–330.

Meunier, F., Hendrikx, I., Bulon, A., Van Goethem, K., and Naets, H. (2020). MultiTINCo: multilingual traditional immersion and native corpus. better documented multiliteracy practices for more refined sla studies. International Journal of Bilingual Education and Bilingualism, pages 1–18.

Mohamed, S. (2021). The development of an Arabic curriculum framework based on a compilation of salient features from CEFR level descriptors. The Language Learning Journal, pages 1–15.

Montgomerie, A. (2021). CEFR levelled English texts: A dataset of English texts labelled with CEFR reading levels. [https://www.kaggle.com/amontgomerie/cefr-levelled-english-texts](https://www.kaggle.com/amontgomerie/cefr-levelled-english-texts)

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and elt. In Proceedings of the Corpus Linguistics 2003 conference, volume 16, pages 572–581.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Candito, M., Cebriáñez Eryijít, G., Celano, G. G. A., Chalúp, F., Choi, J., Čoltínek, Č., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovic, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Erjavec, T., Farkas, R., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökkürmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grützits, N., Guillaume, B., Habash, N., Hajić, J., Hà Mỷ, L., Haug, D., Hladká, B., Hohle, P., Ion, R., Irímia, E., Johannsen, A., Jørgensen, F., Kaškura, H., Kanayama, H., Kanerva, J., Kotsyba, N., Krek, S., Laippala, V., Lê Hồng, P., Lenci, A., Ljubešić, N., Lyashevskaya, O., Lynn, T., Makazhanov, A., Manning, C., Máránduc, C., Mareček, D., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Mori, S., Moskalevskyi, B., Müschnek, K., Mustaffa, N., Mührisep, K., Nguyễn Thị L., Nguyễn Thị Minh, H., Nikolova, V., Nurmí, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalnina, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Real, L., Rituma, L., Rosa, R., Saleh, S., Sanguinetti, M., Saulté, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakhurova, L., Shen, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Tsrarfaty, R., Tyers, F., Uematsu, S., Uria, L., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.0.
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.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In International Conference on Semantic Computing (ICSC 2007), pages 517–526. IEEE.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082.

Rafalovitch, A., Dale, R., et al. (2009). United Nations general assembly resolutions: A six-language parallel corpus. In Proceedings of the Machine Translation Summit (MT Summit), volume 12, pages 292–299, Ottawa, Canada.

Ströbel, M., Kerz, E., and Wiechmann, D. (2020). The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. Language Learning, 70(3):732–767.

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.

Tan, L. and Bond, F. (2011). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In Proceedings of the Pacific Asia Conference on Language, Information and Computation, pages 362–371.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Language Resources and Evaluation Conference (LREC), volume 2012, pages 2214–2218, Istanbul, Turkey.

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.