Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 79–88
Marseille, 20-25 June 2022

ZAEBUC: 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 ZAEBUC, 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 ZAEBUC 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 are 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)[1] 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 extra-linguistic 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[2] 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; Alfaifi, 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

[1]The Arabic word زئبق *zi'baq* means 'mercury'.

[2]http://www.zaebuc.org

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.

CEFR Annotations Texts in a number of learner corpora are annotated for language proficiency level according to the Common European Framework of Reference for Languages (CEFR) for, among other languages, English (Montgomerie, 2021), Czech, German, and Italian (Boyd et al., 2014). Building on previous efforts on CEFR for English and for Arabic (Mohamed, 2021), all of our corpus essays are rated for CEFR by multiple raters. To our knowledge, the Arabic ZAEBUC CEFR annotations are the first of their kind for Arabic, and for native writing of any language.

Spelling and Grammar Correction In the development of ZAEBUC we took inspiration and followed the lead of spelling and grammar correction decisions made by Zaghouani et al. (2014) for Arabic and Dahlmeier et al. (2013) for English. We opted to annotate grammatical and spelling errors by simply correcting them, and did not tag for error types, unlike Alfaifi (2015). This was driven by cost reduction, as well as simplifying the annotations.

Morphosyntactic Annotations We followed commonly used standards for tokenization, tagging and lemmatizations for Arabic and English to allow the use of the corpus in computational linguistics research and system development (Marcus et al., 1993; Maamouri et al., 2004). In particular, we used the Universal Dependencies part-of-speech standards as they are designed to maximize comparability between languages (Nivre et al., 2017; Taji et al., 2017). We did not work on syntactic annotations, but in a very recent effort, (Habash et al., 2022) presented a new multi-genre Arabic dependency treebank that included portions of ZAEBUC.

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.

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,[3] 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] https://www.ramitechs.com/

Topic	Prompt
وسائل التواصل الاجتماعي Social Media	وسائل التواصل الاجتماعي وتأثيرها على الفرد والمجتمع. How do social media affect individuals and society?
التسامح Tolerance	كيف نعزز ثقافة التسامح في المجتمع؟ How can the UAE promote a culture of tolerance in society?
التطور الحضاري Development	التطور الحضاري الذي تشهده دولة الإمارات العربية المتحدة What do you think are the most important developments in the UAE at the moment?

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.

		Students		Texts	
		397		602	
Gender	Female	353	89%	542	90%
	Male	44	11%	60	10%
High School Type	Government	215	54%	348	58%
	Private	164	41%	229	38%
	Other	18	5%	25	4%
High School Language	English	196	49%	280	47%
	Arabic	183	46%	298	50%
	Other	18	5%	24	4%
Student Language & Topic	Arabic only	9	2%	9	1%
	English only	183	46%	183	30%
	Both	205	52%	410	68%
	Same Topic	*149*	*73%*	*298*	*73%*
	Diff Topic	*56*	*27%*	*112*	*27%*
Text Language & Course & Topic	Arabic	214	54%	214	36%
	Social Media			*171*	*80%*
	Tolerance			*31*	*14%*
	Development			*12*	*6%*
	English	388	98%	388	64%
	Social Media			*330*	*85%*
	Development			*48*	*12%*
	Tolerance			*10*	*3%*

Table 2: Corpus statistics detailing variations across a number of dimensions.

4. Data Collection

We collected the corpus in the Fall of 2019 from the last week of August until mid-September from among first-year students at multiple ZU campuses (Abu Dhabi and Dubai's female and male campuses). We contacted all of the students who took *ENG 140: English Composition I*, *ARA 130: Arabic Concepts* (the primary composition course), or *ARA 030: Arabic Preparedness* (a zero-credit preparatory course) regarding donating their introductory assessment test texts to this project. Only the texts from students who consented in writing were included in the corpus. We anonymized all of the students' private information in the released corpus. All the students were given the same three topics to select from in Arabic and English: *Social Media, Tolerance*, and *Development*. See Table 1 for the prompts associated with the topics.

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

	English Example	Arabic Example
C1	Social media is a widely controversial subject with various opinions regarding its negative and positive aspects. While social media has many positive impacts on society, it can also imprint many negative changes on people worldwide. Social media is widely used as a means of communication between people. Social media lifts boundaries made by long distances 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 simultanious increase of false information or news being shared 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 anonymusly done. Since social media allows people to wear masks, it has been proven by psychology that it allows people to be more comfortable being harsh on others. This causes people to lose their lives and mental health due to bullying through social media. In conclusion, there are many positive and negative impacts of social media. Therefor, it is important to raise awarness on the dangers of social media to help limit the negative impacts it can cause on individuals and societies.	في عصرنا الحالي المبني على التكنولوجيا، تتمتع وسائل التواصل الإجتماعي بأهمية كبيرة، حيث يستصعب على الكثير من الناس العيش من دونها. لدى وسائل التواصل الإجتماعي أثر كبير على حياتنا اليومية و علينا أن نتفادى الوقوع في سلبيات هذه الآثار. لدى وسائل التواصل الإجتماعي إيجابيات وسلبيات عديدة، من إيجابيات هذه الوسائل التعرف على ناس جدد وشعوب مختلفة، بذلك نكون قد كوّنا دائرة جديدة من الأصدقاء، وهذا بحد ذاته يساهم في تطوير شخصياتنا، وقال الله عز وجل: (وخلقناكم شعوباً وقبائل لتعارفوا). من الآثار الجميلة أيضاً هي نشر الوعي لدى الفرد والمجتمع. عندما نستغل هذه الوسائل في تمنية المجتمع ونشر قيم مهمة، نكون قد أجدنا استخدامها. أما إذا استغلينا هذه الوسائل في نشر الفكر الخاطئ، نكون قد وقعنا في فخ وسائل التواصل الإجتماعي. من سلبيات هذه الوسائل هو استغلالها في نشر الإشاعات والتحدث بشكل خاطئ عن الآخرين. الأمر السيء في وسائل التواصل الإجتماعي هو سرعة انتشار الأخبار التي معضمها يكون خاطئا، فنحن كأفراد بالغين وعاقلين، يجب علينا أن نمنع انتشار هذه الإشاعات بتخفيف استخدام وسائل التواصل الاجتماعي. مع الاستخدام الكثير للتكنولوجيا ووسائل التواصل الاجتماعي، يجب علينا أن ندرك أثرها علينا حتى لا تؤثر سلبياً على حياتنا اليومية.
B2	Social media has shown a big effect on individuals and the society. Starting with individuals not only teenagers and adults that are using social media but also kids. It became so important for people to check their social media every hour and before going to bed and the first thing to check when waking up. In old days, people used to visit each other to know what their family or friends are doing but now a days you can check how your family and friends are doing just by social media either through texting or by watching their stories "snapshot" "instagram" etc. …The social media effect on the society in my opinion that people starting hanging out less, families don't meet oftenly but I noticed that it had a positive effect on medical and nutrition awareness people started knowing what is good and bad for their health since it's easier to contact a doctor through social media. To conclude, Social media has a negative and positive affect but its helpful for a lot of things normaly.	للتسامح أهمية واسعة في الدول التي لا تعد ولا تحصى، مع ذلك تفتقرها بعض المجتمعات لسوء العقليات فيها وكذلك اختلاف مستويات الثقافة العامة. في هذا المقال سأذكر طرق تساهم في تعزيز ثقافة التسامح في المجتمعات التي تعد أساس الدولة وركنها. تتشعب طرق تعزيز التسامح بين الأفراد والمجتمعات، ولكن يحتاج المجتمع إلى قلوب صافية للوصول لمستوى مرغوب فيه من التعايش والتسامح مع الآخرين. أولاً التسامح يكون تسامح داخلياً قبل أن تشهده المجتمعات، لأن التسامح شيءٌ عظيم يملكه الإنسان بترك العنصرية الفكرية في المقام الأول و من ثم المضي قدماً نحو تعزيزه في المجتمعات. من سبل تعزيز التسامح نشر المحبة والتعاون والكفاح من أجل محو العنصرية كما يحث الدين الإسلامي. لم يصل التسامح بعد إلى عقول البشر التي لازالت مغلقة ولكنها اليوم تنتشر بشكل فريد مما يساهم في حفظ المجتمعات. في الختام، تقع مسؤولية نشر التسامح وتعزيزة على الأفراد في مالهم دور كبير في المجتمعات. التسامح قيمة وكنز في المجتمعات ويجب أن تتزايد وتصبح منتشره حول العالم. معاً لمحو العنصرية ولنستقبل التسامح بصدر رحب.
B1	A lot of people argue whether social media affect individuals and society or not. Does it? yes it does but both in negative and positive ways .Social media is a like a free playground where everyone can say, share, post whatever they like with not much restrictions, Well this can be useful and benificail 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.	تعد مواقع التواصل الاجتماعي من أشهر الوسائل الموجودة في العصر. أشهر موقع تواصل حالي هو الفيسبوك و من بعده الأنستقرام و يعتبر الناس هذه الوسائل كشيء جيد، و لكن في الحقيقة منذ اننتشرت هذه الوسائل توقف الناس عن التجمع مع بعضهم و قطعوا صلة الرحم و اعتبروا هذه الوسائل كطريقة اسهل لكي يباركوا للناس و يهنوهم بدلا ان يحضورو . سببت هذه الوسائل الأدمان لبعض الناس و اصبح لديهم هوس بلشهرة بسبب هذه الوسائل و لكن هنالك الكثيرون الذين يكسبون رزقهم من هذه الوسائل من خلال التصوير و قيام الدعايات للمتاجر و المطاعم . وسائل التواصل الاجتماعي ممتعة ولكن لها حدود ل يجوز للناس القطاع عن بعضعم البعض بسبب هذه الوسائل و يجب عليهم زيارة البعض حتى لا تنقطع صلة الرحم
A2	In my opinon think socail media has been the most importent 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 esair for everyone because people had to get out to buy some newspapers and it takes a lot of time and probebly half of the pepole were lazy. What is also good about it is the you also can know news about from the other countries yes it also can be at the news paper but maybe they take few days to write it down and from your phone or laptop you can know in just one seconed. One of the best benefits of gthe social media is the your can call people from different countries and thagt really good and helpful. Best part of that is you can do	قام انتشار الوساءل للتواصل الاجتماعية بشكل كبير و هذا اثر على المجتمع بشكل ايجابي و سلبي من الاثار الايجابي للتواصل الاجتماعي هي التواصل مع الناس بشكل اسهل . و من ال الاثار ال سلبية هيه انتشار الكراهية و الفساد بين الناس.

Figure 1: Examples of essays and their CEFR levels.

considered unassessable by at least one rater, we consider it unassessable on average, and exclude it from the overall calculations; but not from inter-rater agreement. Figure 1 showcases some example texts and their associated average CEFR levels. Table 3 presents the overall distributions of the averaged CEFR levels assigned to the Arabic and English texts.

5.2. Inter-rater Agreement

The average pairwise exact agreement among our three raters (R1, R2, and R3), i.e., of R1=R2, R1=R3, and R2=R3, is 47% for Arabic texts and 30% for English texts. Assuming a simple random agreement of 1/6 (or 17%, for the six possible CEFR levels), Cohen's Kappa (Cohen, 1960) is then 0.36 (fair agreement) for Arabic, and 0.16 (slight agreement) for English. We note that the average maximum difference between the highest and lowest assigned CEFR levels per text is 0.9 in Arabic, and 1.3 in English. A fuzzy match allowing a difference of up to 1 level maximum, leads to average pairwise fuzzy agreement of 91% for Arabic texts, and 85% for English texts. The random agreement for this

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

Table 3: ZAEBUC CEFR level distributions.

		Arabic	English	All
All Students		3.5	2.9	3.1
Gender	**Female**	3.5	3.0	3.2
	Male	3.4	2.6	2.8
High School Language	**Arabic**	3.5	2.6	3.0
	English	3.4	3.3	3.3
High School Type	**Government**	3.5	2.6	3.0
	Private	3.4	3.4	3.4
Topic	**Social Media**	3.5	3.0	3.2
	Development	3.4	2.5	2.7
	Tolerance	3.5	3.0	3.4

Table 4: CEFR statistics for Arabic and English texts across corpus variables.

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 200 students,[4] whose corresponding averages are 3.5 and 2.9 for Arabic and English. For 50% of these 200 students, the Arabic CEFR level was better than their English CEFR level; for 15%, the English CEFR was better; and for the rest, 35%, the two levels were the same. Second, female students received higher CEFR scores on average than male students by 0.4 level. The difference in English is higher than the difference in Arabic. Third, students who went to English-medium schools performed about the same in English and Arabic. However, students who went to Arabic-medium schools did slightly better on Arabic, but much worse on English, than students who went to English-medium schools: there is a difference of 0.9 level between their Arabic and English texts; and they are lower on English than their English-medium school peers by 0.7 level. This pattern repeats almost exactly for government vs. private schools; this is not a surprise since in our data, 94% of private schools use English primarily, and 82% of government schools use Arabic primarily. Finally, in terms of topics, there is no difference among the Arabic texts, but texts about Development in English are lower by 0.5 level compared to the other topics.

[4]There are 205 students with Arabic and English texts, but 5 of them received unassessable CEFR scores in Arabic.

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 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

Raw	Corrected	Edit	WS Tokens	M Tokens	POS	Lemma
the	The	EDIT	The	The	DET	the
social	social		social	social	ADJ	social
media	media		media	media	NOUN	media
didnt	didn't	EDIT	didn't	did+not	AUX+ADV	do+not
affect	affect		affect	affect	VERB	affect
one	one		one	one	NUM	one
country	country		country	country	NOUN	country
or	or		or	or	CCONJ	or
	a	INS	a	a	DET	a
specific	specific		specific	specific	ADJ	specific
group	group		group	group	NOUN	group
of	of		of	of	ADP	of
people,	people;	EDIT	people	people	NOUN	people
			;	;	PUNCT	;
...	...	...	...	...	...	...

Raw	Corrected	Edit	WS Tokens	M Tokens	POS	Lemma
التسامح	التسامح		التسامح	التسامح	NOUN	تسامح
شي	شيء	EDIT	شيء	شيء	NOUN	شيء
مهم	مهم		مهم	مهم	ADJ	مهم
في	في		في	في	ADP	في
الحياه	الحياة	EDIT	الحياة	الحياة	NOUN	حياة
منه	منه		منه	من+ه	ADP+PRON	من
نتعلم	نتعلم		نتعلم	نتعلم	VERB	تعلم
كيف	كيف		كيف	كيف	ADV	كيف
ان	أن	EDIT	أن	أن	SCONJ	أن
نصبح	نصبح		نصبح	نصبح	VERB	أصبح
اكثر	أكثر	EDIT	أكثر	أكثر	ADJ	أكثر
تعطفن	تعاطفا،	EDIT	تعاطفا	تعاطفا	NOUN	تعاطف
			،	،	PUNCT	،
،ويجب	ويجب	EDIT	ويجب	و+يجب	CCONJ+VERB	وجب
...	...	...	...	...	...	...

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.

		Arabic	English
(a)	**Text Count**	214	388
(b)	**Raw Word Count**	33,376	87,602
	Raw Word/Text	156	226
(c)	**Corrected Word Count**	31,661	87,621
	Corrected Word/Text	148	226
(d)	**Exact Match**	68.0%	80.3%
	Edit	25.7%	17.0%
	Delete	6.3%	2.7%
	Insert	1.2%	2.7%
(e)	**WS Token Count**	34,235	97,478
	WS Token/Text	160	251
(f)	**Morph Token Count**	42,927	98,452
	Morph Token/Text	201	254
(g)	**Al+Morph Token Count**	51,609	
	Al+Morph Token/Text	241	

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

ratio of exact match words (correct words) among raw text words is 68.0% in Arabic and 80.3% in English — Arabic has almost 1.6 times the number of errors in English. This is not surprising given the observation above about the difference in word count: since Arabic words are denser in content, there are multiple reasons for errors per word.

Correlation of CEFR and Text Correction The Pearson correlation between the number of exact matches (between raw and corrected texts) and the average CEFR levels as discussed above is about the same for Arabic (0.70) and English (0.71). The high correlation between these two independent measures of writing quality confirms our expectations, but leaves room for other writing aspects that are perhaps captured by the CEFR ratings but not the text correction.

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 ا أ إ آ ئ ؤ ء 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 clitic و *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 (ة) *a(t)* without its dots as (ه) – 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., *arawnd* 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 وكالقمر *wakalqamari* 'and like the moon' is tokenized as و + ك + القمر *wa+ ka+ al-qamari*.[5]

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 كتب *kataba* 'he wrote', سيكتبها *sa-yaktubuhā* 'he will write it', and ونكتب *wa-naktubu* 'and we write' are all lemmatized to كتب *katab*. 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 99.9%, 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 ال *al* 'the', which occurs in 20% of all words, the difference between English and Arabic in word count diminishes to less than 5%; see Table 6 (g).

Parts-of-Speech The number of unique POS tags in *untokenized* 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.,

[5] For more details on Arabic computational morphology, see Habash (2010).

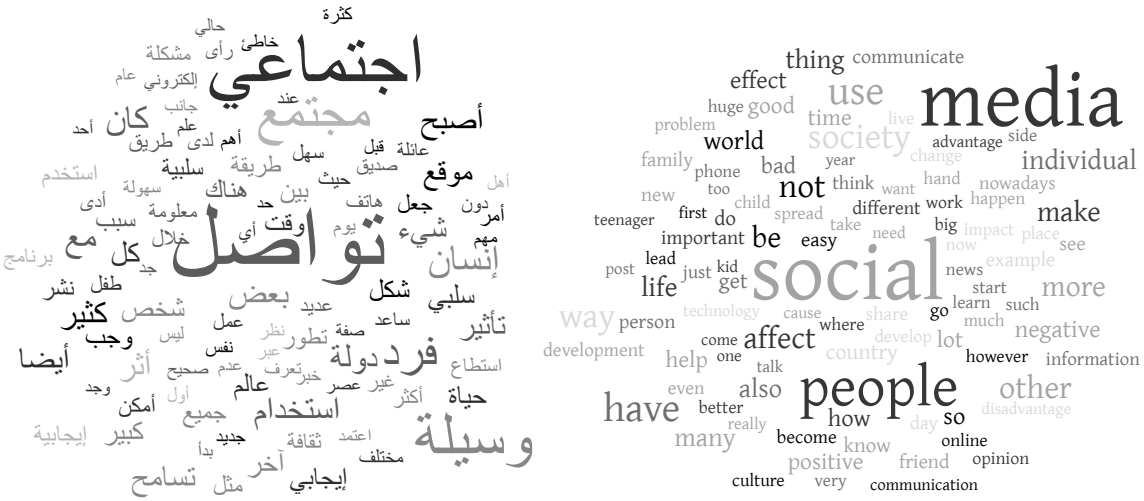

Figure 2: Arabic and English word clouds for the most frequent 100 (verb, noun, adjective, or adverb) lemmas in the Arabic and English texts. The top 10 in English are *social, media, people, use, have, society, way, not, other, affect*. The top 10 in Arabic are تواصل، اجتماعي، وسيلة، مجتمع، فرد، إنسان، كان، مع، استخدام، كثير *tawāṣul, ijtimā'ī, wasīla, mujtama', fard, insān, kān, ma'a, istikhdām, kathīr* 'connection, social, device, society, individual, human-being, be, with/together, use, very.'

CCONJ+NOUN+PRON, are the main reason why it has three times the number of untokenized POS tags compared to English. When comparing the distributions of specific tokenized POS tags, we notice a few important differences connected to the syntactic and morphological structures of the two languages. A two-tailed paired T-test on the shared texts submitted by the same students (N=205) shows that the differences are statistically significant for all POS ($p<.01$) except for ADJ, PART, PROPN, SCONJ and INTJ. We present some of our observations next.

Among the open class POS, we note that NOUNs are almost twice as common in Arabic as in English. This is connected to many determiners and prepositions in English being expressed as nouns in Arabic. VERBs are also more common in English than Arabic by 29%. This is not surprising given that Arabic heavily employs verbless copular sentences.

Among closed POS, we note the high usage of CCONJ (coordinating conjunction) in Arabic – almost twice as much as English. This may be explained by the high usage of the conjunction و *wa* 'and': 15 instances per Arabic text compared with 9 instances of English *and* per English text. We note the much higher usage of AUX in English compared to Arabic: 7.5 times. This is not unexpected since English makes heavy use of auxiliaries for different verb tense constructions, many of which Arabic marks inflectionally. Finally, we note the higher frequency of DET in English - almost 10 times that of Arabic. This is expected given that, among other things, Arabic has no indefinite article, and its very common definite article ال *al* 'the' is not tokenized in the PATB scheme we use.

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`.

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 Ramitechs.

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., Medveď, 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., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., and Vettori, C. (2014). The MERLIN corpus: Learner language 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). MulTINCo: 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`.

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., Cebiroğlu Eryiğit, G., Celano, G. G. A., Chalub, F., Choi, J., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, 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., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hà Mỹ, L., Haug, D., Hladká, B., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, 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., Mărănduc, 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., Muischnek, K., Mustafina, N., Müürisep, K., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Real, L., Rituma, L., Rosa, R., Saleh, S., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, 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., Tsarfaty, 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.