Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus

Wajdi Zaghouani[1], Nizar Habash[2], Houda Bouamor[1], Alla Rozovskaya[3], Behrang Mohit[4], Abeer Heider[5] and Kemal Oflazer[1]

[1]Carnegie Mellon University in Qatar
{wajdiz,hbouamor}@cmu.edu, ko@cs.cmu.edu
[2]New York University Abu Dhabi
nizar.habash@nyu.edu
[3]Center for Computational Learning Systems, Columbia University
alla@ccls.columbia.edu
[4]Ask.com
behrangm@ischool.berkeley.edu
[5]Qatar University
abeer.heider@qu.edu.qa

Abstract

We present our correction annotation guidelines to create a manually corrected non-native (L2) Arabic corpus. We develop our approach by extending an L1 large-scale Arabic corpus and its manual corrections, to include manually corrected non-native Arabic learner essays. Our overarching goal is to use the annotated corpus to develop components for automatic detection and correction of language errors that can be used to help Standard Arabic learners (native and non-native) improve the quality of the Arabic text they produce. The created corpus of L2 text manual corrections is the largest to date. We evaluate our guidelines using inter-annotator agreement and show a high degree of consistency.

1 Introduction

Learner corpora (or L2 corpora) are collections of texts written by non-native learners of the languages of the texts. They are generally marked by a high error rate, i.e., orthographic, lexical, and grammatical errors (Granger, 2003; Hammarberg and Grigonyté, 2014). Learners of Arabic as second language often struggle to produce fluent Arabic text. In addition to the significant structural and conceptual differences between Arabic and other languages (English, French, etc.), vocabulary learning is one of the biggest challenges. Apart from content selection and planning, the writer should find the appropriate words/expressions to express her ideas. Finding the best formulation to integrate within the stylistic context of a discourse, or using the terminology that is more adapted to the context might be more complicated. Learners of Arabic as a second language have to adapt to a different script and different grammatical rules. These factors contribute to the propagation of errors made by L2 speakers that are of different nature than those produced by native speakers (L1 speakers). Hence, in order to model learner language and produce highly efficient error detection and correction methods, it is extremely important to collect a large learner corpus, annotate it and analyze the errors contained in it.

Annotated L2 corpora can provide teachers, learners, second language acquisition researchers, lexicographers and language materials writers, with a valuable data resource. For instance, the annotated corpora can be used for Contrastive Interlanguage Analysis (CIA), since it enables researchers to observe a wide range of instances of under-use, overuse, and misuse of various aspects of the learner language at different levels. Moreover, L2 corpora can be used to compile or improve learner dictionary contents, particularly by identifying the most common errors learners make while providing immediate access to detailed error statistics. This can provide learners with a very useful feedback and help them improve their proficiency level.

Proceedings of LAW IX - The 9th Linguistic Annotation Workshop, pages 129–139,
Denver, Colorado, June 5, 2015.

These errors may take place in words, phrases, language structures, and the ways words or expressions are used (Granger, 2003). For Arabic, there are few projects that aim at developing Arabic learner corpora and annotating them but most of them are not freely available for users or researchers (Abuhakema et al., 2008; Hassan and Daud, 2011).

In this paper, we present our annotation method and our efforts for extending an L1 large scale Arabic language corpus and its manually edited corrections to include annotated non-native Arabic learner text (L2). This work is part of the Qatar Arabic Language Bank (QALB) project (Zaghouani et al., 2014b), a large-scale error annotation effort that aims to create a manually corrected corpus of errors for a variety of Arabic texts (the target size is 2 million words).[1] Our overarching goal is to use our annotated corpus to develop components for automatic detection and correction of language errors that can be used to help Standard Arabic learners (native and non-native) improve the quality of the Arabic text they produce. The previous version of our annotation guidelines focused on native speaker text. Our extended L2 guidelines are built on the existing L1 guidelines (Zaghouani et al., 2014a) with a focus on the types of errors usually found in the L2 writing style and how to deal with problematic ambiguous cases.[2] Annotated examples are provided in the guidelines to illustrate the various annotation rules and their exceptions. As with the L1 guidelines, the L2 texts should be corrected with a minimum number of edits that produce semantically coherent (accurate) and grammatically correct (fluent) Arabic. The guidelines also devise a priority order for corrections that prefer less intrusive edits starting with inflection, then cliticization, derivation, preposition correction, word choice correction, and finally word insertion. The corpus of L2 text manual corrections we create is the largest to date. We evaluate our guidelines using inter-annotator agreement and show a high degree of consistency.

The remainder of this paper is organized as follows. First, we give an overview of related work in Section 2; then we describe the corpus and the annotation guidelines in Sections 3 and 4. Afterwards, we present our annotation tool and pipeline in Sections 5 and 6. Finally, we present an evaluation of the annotation quality and discuss the L2 annotation challenges in Section 7.

2 Related Work

Currently available manually corrected learner corpora are generally limited when it comes to the language, size and the genre of data. Several corpora of learners of English annotated for errors are publicly available (Rozovskaya and Roth, 2010; Yannakoudakis et al., 2011; Dahlmeier et al., 2013), ranging in size between 60K words and more than one million words. Dickinson and Ledbetter (2012) annotated errors in student essays written by learners of Hungarian at three proficiency levels at Indiana University. The annotation was performed using EXMARaLDA, a freely available tool that allows multiple and concurrent annotations (Schmidt, 2010). Student errors were marked according to various categories of phonological, spelling, agreement and derivation errors.

For Arabic, very few learner corpora annotation project have been built. Abuhakema et al. (2008) annotated a small corpus of 9K words of Arabic written materials produced by native speakers of English in the US who learned Arabic as a foreign language. Part of the learners' texts were written while the learners were studying Arabic in the US, while others were produced when they went to study abroad in Arab countries. A tagset of error annotation based on the FRIDA (French Interlanguage Database) tagset (Granger, 2003) was developed to mark-up the learners' errors.

The Corpus of Malaysian Arabic Learners is another project mainly designed to investigate the incorrect use of Arabic conjunctions among learners. It includes 240K words, produced by various Malaysian university students during their first and second year of Arabic major degree. The corpus includes descriptive and comparative essays produced using Microsoft Word without any help from native speakers (Hassan and Daud, 2011). This corpus is currently not publicly available.

More recently, Farwaneh and Tamimi (2012)

[1] http://nlp.qatar.cmu.edu/qalb/

[2] The L1 guidelines are available at http://reports-archive.adm.cs.cmu.edu/anon/qatar/CMU-CS-QTR-124.pdf

introduced The Arabic Learners Written Corpus (ALWC). This corpus includes around 51K words written by non-native Arabic speakers in the United States and were collected over a period of 15 years. ALWC covers three levels (beginner, intermediate and advanced), and three text styles (descriptive, narrative and instructional). Another notable work in progress has been initiated by Alfaifi and Atwell (2012) aiming at building a ∼282K word Arabic learner corpus. The corpus consists of written and spoken materials produced by native and non-native learners of Arabic from pre-university and university levels. Unfortunately, the authors plan to annotate and correct only 10k words of errors in the corpus according to a labeling system inspired by Abuhakema et al. (2008).

3 Corpus Description

Since it was costly to compile our own corpus, we use two freely available L2 Arabic Corpora representing a total of 189K words:

- 51K words from the Arabic Learners Written Corpus (ALWC) (Farwaneh and Tamimi, 2012);
- 139K words from the Arabic Learner Corpus (ALC) (Alfaifi and Atwell, 2012).

The original files of ALWC were in a PDF format. In order to get raw data, we first export the PDF to text, then we manually verified the extracted text to ensure that the data was preserved.

The version of ALC we use is a collection of texts (narrative and discussion) produced by 92 learners of Arabic as a second language in Saudi Arabia and captured in November and December 2012. The corpus is divided according to students' level (beginner, intermediate, advanced).[3]

4 L2 Annotation Guidelines

Essays produced by learners of a Arabic as second language differ from those of natives, not only quantitatively but also qualitatively. Their writings display very different frequencies of words, phrases, and structures, with some items overused and others significantly underused. They also contain varying degrees of grammatical, orthographic and lexical errors. Moreover, sentences written by Arabic L2 speaker have often a different structure and are not as fluent as sentences produced by a native speaker even when no clear mistakes can be found. Therefore, the correction task is complicated by the fact that the acceptability level of a given sentence differs widely within the native speaker annotators as stated by Tetreault and Chodorow (2008). These issues can be related to linguistic factors such as inter-language (L1 interference), the student's teaching and learning methodology, and to the translation effect (conscious interference). Thus, correcting the Arabic L2 essays can be a very challenging task that requires a lot of interpretation efforts by the annotators. This will likely lead to lower inter-annotator agreement as there is often many possible ways to correct the L2 errors.

In order to annotate the L2 corpus, we use our annotation guidelines designed for L1 (Zaghouani et al., 2014b) and add specific L2 annotation rules. Annotation guidelines typically describe the core of the annotation policy. Our annotation guidelines describe the types of errors that are targeted and detail how to correct them, including how to deal with borderline cases. Many annotated examples are provided in the guidelines to illustrate the various annotation rules and exceptions.

As with the L1 guidelines, we adopt an iterative approach to write and improve the L2 guidelines by evaluating various rounds of annotation. The goal is to reach a clear and consistent set of directions for the annotators. For instance, several changes to the guidelines were needed to address the correction of dialectal words and whether or not to correct or ignore certain word categories.

In the following subsections we briefly review the main error types corrected and presented in the guidelines. Then, we detail the L2 specific errors and the L2 correction strategies adopted.

4.1 Guidelines for Error Correction

Errors in any natural language can be defined as a deviation from the standard language norms in word morphology, syntax, punctuation, etc. They can be classified according to basic types such as omis-

[3] A more detailed description of ALC is given at: http://www.arabiclearnercorpus.com/

sion, addition, or substitution errors; or in terms of word order and grammatical form errors. In order to help the annotators understand the types of errors to be corrected, we document them in the annotation guidelines. Furthermore, to reduce over-correction and improve annotation consistency, we instructed the annotators to avoid modifications of any informal or colloquial writing style, which is considered by some to be less acceptable than formal style.

We group the errors to be corrected into seven categories and define them in the guidelines as follows.

Spelling Errors: These occur when at least one of the characters in a word is deleted or substituted by another character, or when an extra character is inserted. Some of these errors result in non-words and some result in other correct words which can not be used in that context.

Word Choice Errors: These include the use of an incorrect word. It was made clear in the guidelines that only wrong word choices are considered for correction, while style changes should not be made since the goal is not to correct or improve the writing style of the text. Word choice errors are particularly frequent in the L2 Arabic student essays.

Morphology Errors: These are usually related to an incorrect derivation or inflection, or incorrect templatic or concatenative morphology. The annotator should be aware of the Arabic morphological inflection rules and their exceptions in order to be able to correct this type of errors.

Syntactic Errors: These include wrong agreement in gender, number, definiteness or case as well as wrong case assignment, wrong tense use, wrong word order, and missing word or redundant/extra words.

Proper Name Errors: These occur in the spelling of persons, organizations, and locations, especially those of foreign origin which could be incorrectly transliterated. If the text uses one of multiple widely acceptable transliterations, the annotators should not modify the word.

Punctuation Errors: Punctuation errors should be corrected according to the commonly accepted Arabic punctuation rules.

Dialectal Usage Errors: In comparison to Standard Arabic, where there are clear spelling standards and conventions, Arabic dialects do not have official orthographic standards partly since they were not commonly written until recently. Today, Arabic dialects are often seen in social media, but also in published novels (and there is even an Egyptian Arabic Wikipedia). Habash et al. (2012) proposed a Conventional Orthography for Dialectal Arabic (or CODA) targeting Egyptian Arabic for computational modeling purposes and demonstrated how to map to it in (Eskander et al., 2013) and (Pasha et al., 2014; Habash et al., 2013). CODAs for other dialects have also been proposed (Zribi et al., 2014; Jarrar et al., 2014). In our current annotation task we neither address dialectal Arabic spelling normalization (Eskander et al., 2013), nor do we systematically translate dialectal words into Standard Arabic (Salloum and Habash, 2013). We recognize that the Arabic language is in a diglossic situation and borrowing is frequent. Most of the texts provided for annotation are in Standard Arabic, but dialectal words are sometimes mistakenly used. We are interested in reducing various spelling inconsistencies that frequently occur. So, as was done in the L1 annotation effort (Zaghouani et al., 2014b), we asked annotators to flag the highly dialectal cases to be reviewed later by the annotation manager. The guidelines classify dialectal word issues into five categories inspired by Habash et al. (2008): dialectal lexical choice, pseudo-dialectal lexical choice, morphological choice, phonological choice and closed class dialectal words. Only the last three categories are considered for correction. For more details, see (Zaghouani et al., 2014a; Zaghouani et al., 2014b).

For more information on Arabic in the context of natural language processing, see (Habash, 2010).

4.2 Additional L2 Annotation Rules

Non-native essays often contain wrong lexical choices or unknown words due to misspelling and it is not easy for annotators to understand these words, interpret the errors and replace them with the correct form (the intended word chosen by the writer). In order to avoid any annotation inconsistency, we extend the general guidelines by adding new rules describing the error correction procedure in texts produced by L2 speakers.

4.2.1 General Correction Rules

First, non-native speaker texts should be corrected with a minimum number of edits.[4] However, correcting errors and making the text semantically coherent and grammatically correct is more important than minimizing the number of edits. Hence, annotators were asked to pay attention to the following three aspects :

- **Accuracy:** The accuracy of the text is very important and any missing meaning observed according to the sentence context should be added to ensure the coherence of the sentence.
- **Fluency:** Spelling, grammatical and agreement errors occur frequently in L2 texts and they should be always corrected. Word reordering is only permitted when it is needed to correct the meaning or the syntax.
- **Style:** L2 texts may be written in a style that is unfamiliar or unnatural to native speakers although the word order is acceptable, and the sentence conveys the meaning correctly. In such cases, the writing style should not be modified to match that of a native speaker.

4.2.2 Correction Priority Order

In order to abide by Arabic's grammatical and spelling rules, it is sometimes necessary to insert new words or do a major correction to unsuitable words selected by the non-native speakers. As the Arabic language is known to have a complex morphology, there is often many ways to correct errors, e.g., by changing the derivation or changing the inflection. To minimize the number of edits and correction, and to avoid any disagreement between annotators, we provide them with the following correction priority guidelines when a correction involving word edits is needed. By following the following predefined correction order, the annotators are more likely to produce a consistent annotation.

1. Correct inflection errors.
2. Correct cliticization errors.
3. Correct derivation errors; but keep root intact.
4. Correct preposition errors (by adding, deleting or substituting a preposition).
5. Correct lexical errors.

Inflection Correction: If the correction is not related to a preposition, the annotator should first try to correct the error by limiting the change to the inflection level (e.g., correction of gender and number). An example of inflection correction is shown in Table 1 in which the verb بدأنا *bdÂnA*[5] 'we started' was replaced by its correct form بدأت *bdÂt* 'I started'.

Cliticization Correction: Adding or changing the clitics.[6] In the example given in Table 1, two clitic corrections are made. In the first case, the annotator added the definite article ال+ *Al+* 'the' to مسجد *msjd* 'mosque'. In the second case, the annotator added the missing conjunction و+ *w+* 'and'.

Derivation Correction: Here we change the derivation while keeping the same root when possible. The example given in Table 1 shows the derivation of the correct form اغتسلنا *AγtslnA* 'to do a ritual wash' from the the same root of the contextually incorrect word غسلنا *γslnA* – the root is غ س ل γ s l 'washing-related'.

Preposition Correction or Insertion: Here we add missing prepositions or correct misused prepositions. An example is shown in Table 1. Prepositions are addressed specifically because they appear often as errors in the preliminary analyses we conducted. Errors in prepositions are not unique to Arabic L2; they are rather common for non-native speakers of other languages such as English (Leacock et al., 2010; Rozovskaya and Roth, 2010).

Lexical Correction: Finally, if it is impossible to fully correct the word using the previous four steps, there is a clear case of word choice errors and the annotator may have to replace the word used. This can be employed to especially correct inadequate lexical choices or unknown words. In the example given in

[4]The minimum edits approach in error correction have already been used in the Error-tagged Learner Corpus of Czech project (Hana et al., 2010)

[5]Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *Abtθ jHxdðrzsšSDTDˇ ς γfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ىء , ħ ة, ý ى.

[6]A clitic is a linguistic unit that is pronounced and written like an affix but is grammatically independent.

	Inflection Error Correction
Original	كنت قد **بدأنا** في العام الماضي رحلة إلى مكة. *knt qd bdÂnA fy AlçAm AlmADy rHlħ Ălaý mkħ.*
Correction	كنت قد **بدأت** في العام الماضي رحلة إلى مكة. *knt qd bdÂt fy AlçAm AlmADy rHlħ Ălaý mkħ.*
English	'I had started a trip to Mecca last year.'
	Cliticization Error Correction
Original	وعندما وصلنا **مسجد** الحرام مع زملائي **عددهم** عشرة. *wçndmA wSlnA msjd AlHrAm mç zmlAŷy çddhm çšrħ.*
Correction	وعندما وصلنا **المسجد** الحرام مع زملائي **وعددهم** عشرة. *wçndmA wSlnA Almsjd AlHrAm mç zmlAŷy wçddhm çšrħ.*
English	'And when we got to the Holy Mosque with my ten colleagues.'
	Derivation Error Correction
Original	وقفت الحافلة عند الميقات ونزلنا منها و **غسلنا** ولبسنا ملابس الإحرام. *wqft AlHAflħ çnd AlmyqAt wnzlnA mnhA wγslnA wlbsnA mlAbs AlĂHrAm.*
Correction	وقفت الحافلة عند الميقات ونزلنا منها و **اغتسلنا** ولبسنا ملابس الإحرام. *wqft AlHAflħ çnd AlmyqAt wnzlnA mnhA wAγtslnA wlbsnA mlAbs AlĂHrAm.*
English	'The bus stopped at Miqat and we went down from it and we ritually bathed and we wore ritual clothing.'
	Preposition Correction
Original	لقد ذهبنا الحج هذا العام. *lqd ðhbnA AlHj hðA AlçAm.*
Correction	لقد ذهبنا **إلى** الحج هذا العام. *lqd ðhbnA Ălý AlHj hðA AlçAm.*
English	'We went to the Hajj this year.'
	Lexical Correction
Original	سأضع **المرآة** لكي أقرأ الكتاب. *sÂDç AlmrĀħ lky Âqrâ AlktAb.*
Correction	سأضع **النظارات** لكي أقرأ الكتاب. *sÂDç AlnĎArAt lky Âqrâ AlktAb.*
English	'I will put on the eyeglasses to read the book.'

Table 1: Examples of the different parts of the correction priority order

Table 1, the word المرآة*AlmrĀħ* 'mirror' was replaced by the word النظارات*AlnĎArAt* 'eyeglasses'.

5 The Annotation Tool

In order to ensure the speed and efficiency of the annotation process, as well as better management, we provide the annotators with a web-based annotation framework, originally developed to manually correct errors in L1 texts (Obeid et al., 2013). The annotation interface allows annotators to perform different actions corresponding to the following types of corrections: (a) *edit* misspelled words; (b) *move* words that are not in the right location; (c) *add* missing words; (d) *delete* extraneous words; (e) *merge* words that have been split erroneously; and (f) *split* words that have been merged erroneously.

In our final corpus output format, we record for each annotated file the list of actions taken by the annotator. These actions operate on one or two tokens depending on the action. We also supply token alignments starting from document tokenization to after human annotation.

6 The Annotation Pipeline

The annotation of a large scale corpus requires the involvement of multiple annotators. In our project, the annotation effort is led by an annotation manager, and the team consists of six annotators coming from three Arab countries (Egypt, Palestine and Tunisia) and a programmer. All annotators hold at least a university level degree and they have a strong Arabic language background.

The annotation manager is responsible for the whole annotation task including corpus compilation, the annotation of the gold-standard inter-annotator agreement (IAA) portion of the corpus, writing the annotation guidelines, hiring and training the annotators, evaluating the quality of the annotation, monitoring and reporting on the annotation progress, and designing the annotation tool specifications with the

programmer.

The annotation manager assigns tasks to annotators and controls the quality of produced annotations collected. Note that, we give the annotator the possibility to flag a word if he is not certain about its correction. This alerts the annotation manager to check it and correct it.

The annotation manager selects and uploads the text files into the annotation system to create a new annotation project task. Once uploaded, the files are automatically tokenized and processed using MADAMIRA (Pasha et al., 2014), a morphological disambiguation tool that automatically corrects common spelling errors as a side effect of disambiguation. MADAMIRA uses a morphological analyzer to produce, for each input word, a list of analyses specifying every possible morphological interpretation of that word, covering all morphological features of the word. MADAMIRA then applies a set of models to produce a prediction, per word in-context, for different morphological features, such as POS, lemma, gender, number or person. The robust design of MADAMIRA allows it to consider different possible spellings of words, especially relating to Ya/Alif-Maqsura, Ha/Ta-Marbuta and Hamzated Alif forms, which are very common error sources. MADAMIRA selects the correct form in context, thus correcting for these errors which are often connected to lemma choice or morphology.

7 Evaluation

7.1 Inter-Annotator Agreement

Our annotation effort consists of a single annotation pass as commonly done in many annotation projects due to time and budget constraints (Rozovskaya and Roth, 2010; Gamon et al., 2008; Izumi et al., 2004; Nagata et al., 2006). In order to evaluate the quality of our correction annotations, we frequently measure the inter-annotator agreement (IAA) to ensure that the annotators are following the guidelines provided consistently. A high level of agreement between the annotators indicates that the annotation is reliable and the guidelines are useful in producing homogeneous and consistent data. We measure the IAA by averaging WER (Word Error Rate) over all pairs of annotations to compute the AWER (Average Word Error Rate).[7] For the purpose of this evaluation, the WER refers to an annotation error and it is measured against all words in the text. The higher the WER between two annotations, the lower is their agreement.

Table 2 compares the L1 and L2 portions of our corpus in two dimensions. First, we consider the amount of changes done over the whole corpus measured as WER between raw and corrected text. And secondly, we present the IAA numbers in terms of AWER. The IAA results are computed over 200 files (10,288 words) for the L1 corpus and 20 files (3,188 words) for the L2 corpus. Each of these files is corrected by at least three different annotators. We observe that the number of changes in L2 text is 50% more than that in L1, which is consistent with previous studies and our expectation of the complexity of the task. Furthermore, the IAA in L2 is over 10% absolute points worse than in L1. This is particularly disconcerting, but can be explained by the fact that the correction space for L2 text is larger as many different corrections are possible. In order to verify this hypothesis, we performed a second IAA round in which we provide the first IAA round text output to a second pool of three annotators and we measure how much they agree with the correction done by the first round annotator in term of IAA. The low average WER of 3.35 obtained show that there is a high agreement with the corrections done in the first round. We did not do the same second round for our L1 corpus annotations.

We perform a deeper analysis of the annotated corpus. Results are given in Table 3 and show again that there is a correlation between the number of changes and the level of annotators disagreement. It is clear that ALC is less challenging than ALWC as shown in the IAA of the first round and second rounds.

Overall, the high-level of agreement obtained in the second round shows that the annotators produced consistently similar results under the proposed guidelines; and their differences are all within acceptable variation. This of course makes the evaluation of automatic correction harder.[8]

[7] The annotation manager is excluded from this evaluation.

[8] This problem might be solved by considering multiple references in the evaluation process similarly to what is done in machine translation evaluation (Papineni et al., 2002). Unfortu-

Original	انوي ان ساتهي المقالة في عام الانسان قبل الثلاثاء. *Anwy An sAnthy AlmqAlħ fy ςAm AlAnsAn qbl AlθlAθA'.* 'I plan I will be-done the article in the year of humanity before Tuesday.'
Annotator 1	أنوي أن أنهي المقالة عن عام الإنسان قبل الثلاثاء. *Ânwy Ân Ânhy AlmqAlħ ςn ςAm AlĂnsAn qbl AlθlAθA'.* 'I plan to finish-off the article about the year of humanity before Tuesday.'
Annotator 2	أنوي أن أنهي المقالة عن عالم الإنسان قبل الثلاثاء. *Ânwy Ân Ânhy AlmqAlħ ςn ςAlm AlĂnsAn qbl AlθlAθA'.* 'I plan to finish-off the article about the human world before Tuesday.'
Annotator 3	أنوي أن أتهي من المقالة في عالم الإنسان قبل الثلاثاء. *Ânwy Ân Ânthy mn AlmqAlħ fy ςAlm AlĂnsAn qbl AlθlAθA'.* 'I plan to be-done with the article in The Human World before Tuesday.'

Table 4: Example of multiple annotator corrections of an L2 erroneous sentence.

	Changes	**IAA**$_{Round1}$	**IAA**$_{Round2}$
L1 corpus	24.45%	3.80%	N/A
L2 corpus	37.64%	14.67%	3.35%

Table 2: Comparison between the L1 and the L2 corpus with the percentage of changes from the RAW source corpus and the inter-annotator agreement (IAA) on "all words" in terms of average WER (Punctuation is ignored). Round1 is basic IAA comparing two annotations starting from raw text. Round2 starts with the output of Round1.

	Changes	**IAA**$_{Round1}$	**IAA**$_{Round2}$
ALC corpus	32.65%	13.56%	3.13%
ALWC corpus	51.39%	19.12%	4.20%

Table 3: The percentage of changes from the RAW source corpus and the inter-annotator agreement on "all words" in terms of average WER in the two parts of our L2 corpus (Punctuation is ignored). Round1 is basic IAA comparing two annotations starting from raw text. Round2 starts with the output of Round1.

An analysis of the inter-annotator agreement errors shows that in some cases the annotators did not follow the correction priority order specified in the guidelines (Section 4.2.2) or disagreed on how to apply it. They also either did not pay attention or failed to correct spelling mistakes. In other cases, the disagreement is due to multiple possible interpretations of typos or wrong lexical choices.

In Table 4, we show some examples of disagreement among the annotators. The erroneous L2 sentence has multiple Alif-Hamza errors, an incorrect verb clitic and a confusing phrase في عام الانسان *fy ςAm AlAnsAn* 'in the year of humanity'. All the annotators corrected the Alif-Hamza errors and the verb clitic. However, they disagreed on how to correct the problematic phrase 'in the year of humanity' as (a) 'about the year of humanity', (b) 'about the human world', and (c) 'in the human world'. The different corrections interacted with the form of the main verb after clitic correction اتهي *Anthy* 'be-done' producing two corrections: أنهي *Ânhy* 'finish-off' (derivation change) or أتهي من *Ânthy mn* 'be-done with' (add a preposition). In conversations with the annotators about this case, they expressed strong opinions about what they considered to be the acceptable interpretation that justified their corrections.

7.2 L1 vs L2: Similarities and Differences

We selected a sample of 5K words from both the L1 and L2 corpora to compare their errors. Table 5 highlights the ten most frequent errors found in each corpus. Some errors are corpus-specific while other errors occur in both corpora. For example, the wrong word-order error, the redundant word error

nately, such a solution requires more annotations.

and the missing word error are mostly present in the L2 corpus. In contrast, errors such as punctuation errors, incorrect Hamza spelling, and nominal gender/number agreement are present in both corpora.

Err.	Native (L1)	Non-native (L2)
1	Punctuation	Punctuation
2	Hamza	Definiteness
3	Ha/Ta-Marbuta Confusion	Word Choice
4	Alif-Maqsura/Ya Confusion	Hamza
5	Case Endings	Conjunctions, Prepositions
6	Verbal Inflection	Missing Word
7	Agreement	Redundant Word
8	Definiteness	Agreement
9	Conjunctions, Prepositions	Case Endings
10	Word Choice	Word Order

Table 5: Most frequent errors observed in a sample of the L1 and L2 Corpus. The errors are sorted from the most frequent to the least frequent.

8 Conclusion and Future Directions

In this paper, we presented our Arabic L2 correction guidelines and a manually corrected L2 corpus that is the largest to date. We discussed the challenges inherent in learner corpus annotation and we presented our method for efficiently creating an Arabic L2 error corrected corpus. The results obtained in the evaluation suggest that the annotators produced consistently similar results under the proposed guidelines. We believe that publishing this corpus will give researchers a common development and test set for developing related natural language processing applications. A subset of our L2 corpus will be used as part of the Second QALB Shared Task on Automatic Arabic Error Correction in conjunction with the ACL-2015 Workshop on Arabic NLP.[9] This shared task follows the success of the First QALB Shared Task held in conjunction with EMNLP-2014 Workshop on Arabic NLP (Mohit et al., 2014). In the future, we will extend our annotation guidelines to address machine translation output correction (i.e., manual post-editing). We also plan to extend our systems for automatic correction of Arabic language errors (Jeblee et al., 2014; Rozovskaya et al., 2014) to handle L2 data, using the corpus discussed here for training and test purposes.

[9] http://www.arabic-nlp.net/wanlp

Acknowledgements

We thank anonymous reviewers for their valuable comments and suggestions. We also thank all our dedicated annotators: Noor Alzeer, Hoda Fathy, Hoda Ibrahim, Anissa Jrad, Samah Lakhal, Jihene Wafi. We thank Ossama Obeid for his continuous technical support during this project. This publication was made possible by grants NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation).

References

Ghazi Abuhakema, Reem Faraj, Anna Feldman, and Eileen Fitzpatrick. 2008. Annotating an Arabic Learner Corpus for Error. In *Proceedings of The sixth international conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Abdullah Alfaifi and Eric Atwell. 2012. Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors. In *The 8th International Computing Conference in Arabic (ICCA 2012)*, Cairo, Egypt.

Daniel Dahlmeier, Hwee Tou Ng, and Mei Wu Wu. 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.

Markus Dickinson and Scott Ledbetter. 2012. Annotating Errors in a Hungarian Learner Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LRECâĂŹ12)*, Istanbul, Turkey.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Samira Farwaneh and Mohammed Tamimi. 2012. Arabic Learners Written Corpus: A Resource for Research and Learning. *The Center for Educational Resources in Culture, Language and Literacy.*

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Third International Joint Conference on Natural Language Processing, IJCNLP*, pages 449–456, Hyderabad, India.

Sylviane Granger. 2003. Error-Tagged Learner Corpora and CALL: A Promising Synergy. *CALICO*, 20(3):465–480.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, Marrakech, Morocco.

Nizar Habash, Mona Diab, and Owen Rabmow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LRECâĂŹ12)*, Istanbul, Turkey.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Björn Hammarberg and Gintaré Grigonyté. 2014. Non-Native Writers' Errors a Challenge to a Spell-Checker. In *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014)*, Uppsala, Sweden.

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. Error-tagged Learner Corpus of Czech. In *Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV)*, Uppsala.

Haslina Hassan and Nuraihan Mat Daud. 2011. Corpus Analysis of Conjunctions: Arabic Learners Difficulties with Collocations. In *Proceedings of the Workshop on Arabic Corpus Linguistics (WACL)*, Lancaster, UK.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus Exploiting the Language Learners' Speech Database for Research and Education. *International Journal of The Computer, the Internet and Management*, 12(2):119–125, May.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a Corpus for Palestinian Arabic: a Preliminary Study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, Doha, Qatar.

Serena Jeblee, Houda Bouamor, Wajdi Zaghouani, and Kemal Oflazer. 2014. Cmuq@ qalb-2014: An smt-based system for automatic arabic error correction. *ANLP 2014*, page 137.

Claudia Leacock, martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated Grammatical Error Detection for Language Learners. *Synthesis Lectures on Human Language Technologies*, 3(1):1–134.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing*, Doha, Qatar, October.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 241–248, Sydney, Australia.

Ossama Obeid, Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Kemal Oflazer, and Nadi Tomeh. 2013. A Web-based Annotation Framework For Large-Scale Text Correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, Nagoya, Japan, October.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL Errors: Challenges and Rewards. In *NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, CA.

Alla Rozovskaya, Nizar Habash, Ramy Eskander, Noura Farra, and Wael Salloum. 2014. The columbia system in the qalb-2014 shared task on arabic error correction. In *Workshop on Arabic Natural Language Processing, EMNLP*, page 160.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Thomas Schmidt. 2010. Linguistic Tool Development between Community Practices and Technology Standards. In *Proceedings of the LREC Workshop Lan-*

guage Resource and Language Technology Standards State of the Art, Emerging Needs, and Future Developments.

Joel Tetreault and Martin Chodorow. 2008. Native Judgments of Non-Native Usage: Experiments in Preposition Error Detection. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189, Portland, Oregon, USA.

Wajdi Zaghouani, Nizar Habash, and Behrang Mohit. 2014a. The Qatar Arabic Language Bank Guidelines. Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014b. Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2355–2361, Reykjavik, Iceland.