

promoting access to White Rose research papers



Universities of Leeds, Sheffield and York
<http://eprints.whiterose.ac.uk/>

This is an author produced version of a paper published in **Language Processing and Knowledge in the Web, Lecture Notes in Computer Science.**

White Rose Research Online URL for this paper:

<http://eprints.whiterose.ac.uk/77161/>

Paper:

Alfaifi, A, Atwell, E and Abuhakema, G (2013) *Error Annotation of the Arabic Learner Corpus: A New Error Tagset*. In: *Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference, GSCL 2013, 25-27 September 2013, Darmstadt, Germany*. Springer , 14 - 22 (9).

http://dx.doi.org/10.1007/978-3-642-40722-2_2

Error Annotation of the Arabic Learner Corpus

A New Error Tagset

Abdullah Alfaifi¹, Eric Atwell¹, and Ghazi Abuhakema²

¹University of Leeds, Leeds, UK
{scayga, e.s.atwell}@leeds.ac.uk
²College of Charleston, SC, USA
AbuhakemaG@cofc.edu

Abstract. This paper introduces a new two-level error tagset, AALETA (Alfaifi Atwell Leeds Error Tagset for Arabic), to be used for annotating the Arabic Learner Corpora (ALC). The new tagset includes six broad classes, subdivided into 37 more specific error types or subcategories. It is easily understood by Arabic corpus error annotators. AALEETA is based on an existing error tagset for Arabic corpora, ARIDA, created by Abuhakema et al. [1], and a number of other error-analysis studies. It was used to annotate texts of the Arabic Learner Corpus [2]. The paper shows the tagset broad classes and types or subcategories and an example of annotation. The understandability of AALETA was measured against that of ARIDA, and the preliminary results showed that AALETA achieved a slightly higher score. Annotators reported that they preferred using AALETA over ARIDA.

Keywords: error, tagset, Arabic, corpus, learner.

1 Introduction

The benefits of learner error annotation are multi-faceted and extend to fields such as Contrastive Interlanguage Analysis (CIA), learner dictionary making, Second Language Acquisition, and designing pedagogical materials. CIA is still one of the most frequently used approaches for analyzing a learner corpus, as it enables researchers to observe a wide range of instances of underuse, overuse, and misuse of various aspects of the learner language at different levels: lexis, discourse and syntax [3]. Analyzing errors will also enable researchers and educators to understand the interlanguage errors caused by L1 transfer, learning strategies and overgeneralization of L1 rules. Secondly, learner corpora were – and still are – used to compile or improve learner dictionary contents, particularly by identifying the most common errors learners make, and then providing dictionary users with more details at the end of relevant entries. These errors are indicated in words, phrases, or language structures, along with the ways in which a word or an expression can be used correctly and incorrectly [3, 4]. Also, error-tagged learner corpora are useful resources to measure the extent to which learners can improve their performance in various aspects of the target lan-

guage [4, 5]. Compilers of longitudinal learner corpora usually include this goal in their aims. Examples of these include the LONGDALE project: LONGitudinal DATABASE of Learner English [6], Barcelona Age Factor [7], and the ASU corpus [8]. Finally, analyzing learners' errors may be beneficial for pedagogical purposes such as instructional teaching materials development. It can, for instance, help in developing materials that are more appropriate to learners' proficiency levels and in line with their linguistic strengths and weaknesses.

2 Rationale for developing a new tagset for Arabic learner corpora

The classification of errors in Arabic texts should take into account the nature of the different aspects of linguistic description (e.g., lexis, morphology, syntax, semantics, etc.), and the tagset used for this classification should be readily understandable. These two principles are applied in a number of error tagsets that are used and are publicly available, such as Dagneaux, Denness [9] – used in the International Corpus of Learner English, Granger [10] – used in the French Interlanguage Database (FRIDA) corpus, Nicholls [11] – used in the Cambridge Learner Corpus, Izumi, Uchimoto [12] – used in the NICT JLE Corpus, and ARIDA [1] – used in the Pilot Arabic Learner Corpus.

Abuhakema et al's ARIDA tagset aforementioned is the sole error tagset specifically created for Arabic learner corpora, and it is based on the French Interlanguage Database FRIDA tagset. This adaptation from a French tagset, however, rendered some classification inconsistency with traditional Arabic linguistics. For example, in traditional Arabic, grammatical and syntactic errors are combined under one category called either grammar or syntax; in the ARIDA tagset, these are two different error categories. We recognize however that ARIDA's classification may prove appropriate to those trained in Romance languages where this distinction exists. Moreover, the ARIDA tagset is a three-layered tagset that include error domains, grammar categories and error categories. With a language as diverse as Arabic, we felt that two layers of tagging might be sufficient, and training annotators can be a less daunting task for the new tagset. While the ARIDA tagset uses three-character tags, the new tagset uses two-character tags. In addition, a number of the categories in the FRIDA-derived tagset have a literal translation into Arabic with no clarification of what they linguistically or practically mean, which renders them vague. Examples include *Adjective Complementation* "متممة الصفة", *Noun Complementation* "متممة الاسم", and *Verb Complementation* "متممة الفعل". Further, most of the morphological categories describe the error place and not the type. The sole exception is *Inflection confusion* "الخلط في التصريف", which describes an essential morphological error in Arabic learner production. In the *Form/spelling* category, Abuhakema lists important error types, like *hamza* "الهمزة" (ء) and *tanwin* "التنوين" (َ ِ ً), but neglects some others, like *tā' mutataarrifa* "الألف المتطرفة" (آ), *'alif mutataarrifa* "الألف المتطرفة" (أ), *'alif fāriqa* "الألف الفارقة" (إ), and *lām Šamsiya* "اللام الشمسية" (وا).

3 Basis of AALETA development

As a result of the above limitations, we developed another error taxonomy based on ARIDA and other error-analysis studies [13-16]. The reason for relying on the ARIDA tagset is that it includes two comprehensively well-described categories, Style and Punctuation. The other four studies investigate different real types of error in Arabic learner production using the bottom-up method where they analyzed their own samples then extracted the corresponding error-type lists. These studies do not aim to develop an error-type tagset to be used for further projects, such as learner corpora. Nonetheless, their error taxonomies are valid and adaptable since they include significant and comprehensive classes of learner error. Furthermore, we cannot overlook the authenticity of the texts from which these error types are derived; which adds to the validity of their taxonomies. The following is a brief overview:

- Alosaili [13] investigates errors of Arabic learners in their spoken production. His list of errors consists of three main classes: phonological, syntactic, and lexical errors, with sub-types under each domain. Some of these types are included in the tagset proposed in this study, specifically those related to orthography, as they were well-formed and cover clearly significant types.
- Alateeq [14] focuses on semantic errors and extracts a detailed list of them, which is adapted in the proposed tagset. Aside from these semantic errors, the study also lists several phono-orthographical, morphological, and syntactic types of error.
- Alhamad [15] focuses on the writing production of advanced level Arabic learners, and concludes with a list of error categories: phonological, orthographical, morphological, syntactic, and semantic errors. The most comprehensive errors are under orthography and syntax, which are added to the tagset we created.
- Alaqeli [16] examines learners' written errors in a particular type of sentence: a verbal sentence "الجملة الفعلية". This study, therefore, has a limited number of error types under two categories: morphological and syntactic. However, errors under the morphological category are deemed worthy of inclusion in the tagset suggested, due to their comprehensiveness.

Table 1. Error taxonomies in some Arabic studies

Alosaili	Alateeq	Alhamad	Alaqeli
أخطاء في الأصوات	أخطاء صوتية إملائية	أخطاء نحوية	أخطاء نحوية
Phonological errors	phono-orthographical errors	Syntactic errors	Syntactic errors
أخطاء في تراكييب	أخطاء صرفية	أخطاء صرفية	أخطاء صرفية
Syntactic errors	Morphological errors	Morphological errors	Morphological errors
أخطاء في المفردات	أخطاء نحوية	أخطاء إملائية	Morphological errors
Lexical errors	أخطاء نحوية	Orthographic errors	al errors
	Syntactic errors	أخطاء صوتية	
	أخطاء دلالية	Phonological errors	
	Semantic errors	أخطاء دلالية	
		Semantic errors	

4 AALETA tagset

As described, there was a need to develop an error tagset that can provide users (e.g., researchers of Arabic, teachers, etc.) with easily understood broad classes or categories and comprehensive error types. The suggested taxonomy, AALETA, includes 37 types of error, divided into 6 classes or categories: orthography, morphology, syntax, semantics, style, and punctuation. AALETA has two levels of annotation in order to simplify its use and evaluation at this early stage of development. A third layer can be added later when these two layers have achieved a high percentage of accuracy in their use. Each tag consists of two Arabic characters (with an equivalent tag in English). The first character in each tag indicates the error class or category (Table 2), while the second symbolizes the error type (see the example of morphological error in Table 3). For example, the tag *OH* indicates an *[o]rthographical* error in *[H]amza*.

Table 2. Representing error categories in the tagset

Error Category	Orthography الإملاء	Morphology الصرف	Syntax النحو	Semantics الدلالة	Style الأسلوب	Punctuation علامات الترقيم
First part in the Arabic tags	ا	ص	ن	د	ع	ن
First part in the English tags	O	M	X	S	T	P

Table 3. Examples of error types (under the morphological category)

Morphological error الأخطاء الصرفية	Word in- flection صيغة الكلمة	Verb tense زمن الفعل	Other morpho- logical errors أخطاء صرفية أخرى
Second part in the Arabic tags	ص	ز	خ
Second part in the English tags	I	T	O

This taxonomy is flexible and is to be modified based on studies, evaluation, or relevant results. In addition, at the end of each category, there is an item named “*Other [...] errors*”, which can handle any error(s) that do not yet have match(es).

Table 4. AALET: error taxonomy for Arabic learner corpora

Error Category مجال الخطأ	Error Type نوع الخطأ	A-tag الرمز العربي	E-tag الرمز الإنجليزي	
Orthography الإملاء 'al'imlā'	1. hamza (ء، أ، إ، ؤ، ئ، هـ)	<إه>	<OH>	
	2. التاء المتطرفة (ة، ت. tā' mutaṭarrifa)	<إة>	<OT>	
	3. الألف المتطرفة (ا، ي) 'alif mutaṭarrifa)	<إي>	<OA>	
	4. الألف الفارقة (كتبوا) 'alif fāriqa)	<إت>	<OW>	
	5. اللام الشمسية (الطالب) lām šamsiyya)	<إا>		
	6. التنوين (○○○) tanwin)	<إل>	<ON>	
	7. الفصل والوصل (Conjunction) faṣl wa waṣl)	<إو>	<OF>	
	8. تقصير الصوائت الطويلة Shortening the long vowels (○○○ → اوي)	<إف>	<OS>	
	9. تطويل الصوائت Lengthening the short vowels (اوي → ○○○ القصيرة)	<إق>	<OG>	
	10. الخطأ في ترتيب الحروف داخل الكلمة Wrong order of word characters	<إط>	<OC>	
	11. استبدال حرف أو أحرف من الكلمة Replacement in word character(s)	<إس>	<OR>	
	12. وجود حرف أو أحرف زائدة Character(s) redundant	<إز>	<OD>	
	13. وجود حرف أو أحرف ناقصة Character(s) missing	<إن>	<OM>	
	14. أخطاء إملائية أخرى Other orthographical errors	<إخ>	<OO>	
Morphology الصرف 'aṣṣarf	15. صيغة الكلمة Word inflection	<حصص>	<MI>	
	16. زمن الفعل Verb tense	<صزز>	<MT>	
	17. أخطاء صرفية أخرى Other morphological errors	<صصخ>	<MO>	
Syntax النحو 'annaḥw	18. الموقع الإعرابي أو علامة الإعراب Case/Mood Mark	<تنب>	<XC>	
	19. التعريف والتذكير Definiteness	<تبع>	<XF>	
	20. التذكير والتأنيث Gender	<تند>	<XG>	
	21. العدد (الأفراد) (Singular, Dual and plural) والتثنية والجمع	<تف>	<XN>	
	22. ترتيب المفردات داخل الجملة Word(s) order	<تنت>	<XR>	
	23. وجود كلمة أو كلمات زائدة Word(s) redundant	<تنز>	<XT>	
	24. وجود كلمة أو كلمات ناقصة Word(s) missing	<تنن>	<XM>	
	25. أخطاء نحوية أخرى Other syntactic errors	<تنخ>	<XO>	
	Semantics الدلالة 'addalāla	26. اختيار الكلمة المناسبة Word selection	<دب>	<SW>
		27. اختيار العبارة المناسبة Phrase selection	<دق>	<SP>
28. Failure of expression to indicate the intended meaning قصور التعبير عن أداء المعنى المقصود		<دد>	<SM>	
29. Wrong context of citation from Quran or Hadith الاستشهاد بالكتاب والسنة في سياق خاطئ		<دس>	<SC>	
30. أخطاء دلالية أخرى Other semantic errors		<دخ>	<SO>	
Style الأسلوب 'al'uslūb	31. أسلوب غامض Unclear style	<دسغ>	<TU>	
	32. أسلوب ركيك Prosaic style	<دسض>	<TP>	
	33. أخطاء أسلوبية أخرى Other stylistic errors	<دسخ>	<TO>	
Punctuation علامات الترقيم 'alāmāt 'at-tarqīm	34. الخلط في علامات الترقيم Punctuation confusion	<دط>	<PC>	
	35. علامة ترقيم زائدة Punctuation redundant	<دتز>	<PT>	
	36. علامة ترقيم مفقودة Punctuation missing	<دتن>	<PM>	
	37. أخطاء أخرى في علامات الترقيم Other errors in punctuation	<دتنخ>	<PO>	

5 Scope of error tags

The following example, from the Arabic Learner Corpus¹, includes two errors, orthographical OT: character redundant in اللتي “which” [’*allatī*] and stylistic TP: prosaic style in أعطيت أنا لك “I gave you” [’*a ‘aytu ‘anā ‘anta*]). It demonstrates how these errors can be annotated with the appropriate tags when the error is one morpheme (first error) or more (second error). Beside the error annotation, the example here shows lemmas, part-of-speech, and grammatical function tags, and a method of word segmentation in XML (Extensible Markup Language) format:

```
<err type="OD" errform="اللتي" crrform="التي">
  <w>اللتي
    <t token="اللتي" lemma="التي" pos="NR" fun="VA"></t>
  </w>
</err>
<w>كنت
  <t token="كن" lemma="كان" pos="VP"></t>
  <t token="ت" lemma="ت" pos="RR" fun="NK"></t>
</w>
<w>قد
  <t token="قد" lemma="قد" pos="PB"></t>
</w>
<err type="TP" errform="أعطي أنا لك" crrform="أعطيتك">
  <w>أعطي
    <t token="أعطي" lemma="أعطي" pos="VP"></t>
  </w>
  <w>أنا
    <t token="أنا" lemma="أنا" pos="NP" fun="NV"></t>
  </w>
  <w>لك
    <t token="ل" lemma="ل" pos="PP"></t>
    <t token="ك" lemma="ك" pos="RR" fun="GF"></t>
  </w>
</err>
```

6 Measuring understandability of AALETA

To measure the understandability of AALETA against the tagset developed by Abuhakema et al. [1], two annotators (indicated by T1 and T2) were asked to find errors in a sample of learner texts (the same sample for each annotator), and to mark these errors with tags using the proposed refined taxonomy. Both annotators have masters’ degrees and have taught Arabic as a Foreign Language for several years.

¹ ALC is accessed from: <http://www.comp.leeds.ac.uk/scayga/alc>

However, they have not worked on corpus analysis or been involved in any similar task. This can be an advantage, as it could reveal the extent to which the tagset can be understood and useable by untrained users. The texts were taken from ALC which comprises a collection of texts written by learners of Arabic in Saudi Arabia. The corpus covers two types of students, non-native Arabic speakers (NNAS) learning Arabic as a second language (ASL) for academic purpose (AAP), and native Arabic speaking students (NAS) learning to improve their written Arabic. Both groups are males at pre-university level.

Each annotator had to tag the texts twice, using ARIDA tagset first, and AALETA second. Annotators were asked to add the same tag to each repeated error. The assumption was that both error tagsets were clear enough to both annotators, and that they understood which tag is most appropriate to use. Therefore, the error categories and types of both tagsets (ARIDA and AALETA) were not explained to the annotators. This measurement may be sufficient to check whether a tagset can be independently understood against another tagset, considering that the differences between annotators are sometimes due to the annotator’s view of the error type, and not to tagset clarity.

The results show that T1 detected 80 errors, while T2 found 91, and they shared 42 errors; the comparison was performed by calculating matched tags between T1 and T2 in each tagset. When the annotators used the ARIDA tagset, they added the same *error-category* tags to 15 errors (36%) out of 42, and the same *error-type* tags to 14 errors (33%). By using AALETA, the annotators shared the same *error-category* tags on 27 errors (64%), and the same *error-type* tags on 22 errors (52%). Although AALETA achieved a higher score, it is still not perfect, which means that it needs more refinement, and that more tests are still needed using other texts and more annotators.

Determining whether a word/phrase was right or wrong was completely based on the annotator’s view. It was very likely that some differences in their decisions, particularly in some categories such as semantics and style, relate to the degree of linguistic knowledge of the annotator. The disagreements might have been minimized if annotators were given texts with errors already identified and were asked to mark the appropriate tag on each error. This method can be used in future experiments to avoid such differences.

Table 5. Annotating comparison between Abuhakema and AALETA error tagsets

Using Abuhakema’s tagset		
	Error Category	Error Type
No. of same tags (out of 42)	15	14
Percentage	36%	33%

Using AALETA

	Error Category	Error Type
No. of same tags (out of 42)	27	22
Percentage	64%	52%

When the annotators were asked "Which taxonomy was more understandable? And why?", both selected AALETA because of the logical order of its items, and its comprehensiveness. For the question "Which of them was quick and easy for annotating? And why?", they both chose AALETA, as they believe that by using AALETA it is easier to select the proper tag, and that the tags are clearer with no ambiguity or overlap.

7 Conclusions and further work

This paper introduces a newly-refined tagset for error annotation developed specifically for tagging Arabic learner corpora, and draws on ARIDA and other error classification studies. While ARIDA has its own advantages, we believe that it can be improved in ways that make the annotators' task less daunting. The tagset was used for tagging texts taken from the ALC at two levels: board classes and error types. An example of the tagging process is presented. The understandability of AALETA was measured against the ARIDA tagset. Although AALETA scored higher, further work is still needed to compare the two tagsets in more detail. Also, to minimize differences in classifying errors, texts with errors already marked can be given, where the annotators' task is to identify the error category and type. This test will present more reliable data about the validity level of each tagset. Thus further work in collaboration with specialists in corpus linguistics and Arabists is still needed – to refine AALETA to increase its suitability for use in further Arabic learner corpora as a standard error tagset, and affirm its understandability over ARIDA. To make it comprehensible and offer more information about learners' errors, another layer may need to be developed and assessed in terms of comprehensibility, validity and applicability. Since the texts were written by male students in one country, diversifying those texts to include more learners from both genders and other countries may yield different results and types of errors.

References

1. Abuhakema, G., A. Feldman, and E. Fitzpatrick, *ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study*. Journal of the National Council of Less Commonly Taught Languages (JNCOLCTL), 2009. 7: p. 161-184.
2. Alfaihi, A. and E. Atwell. المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية (in Arabic) "Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors". in *8th International Computing Conference in Arabic (ICCA 2012) 26-28 December 2012*. 2012. Cairo, Egypt.

3. Granger, S., *The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research*. TESOL Quarterly, 2003. 37(3): p. 538-546.
4. Nesselhauf, N., *Learner Corpora and Their Potential in Language Teaching*, in *How to Use Corpora in Language Teaching*, J. Sinclair, Editor. 2004, Benjamins: Amsterdam & Philadelphia. p. 125-152.
5. Buttery, P. and A. Caines, *Normalising Frequency Counts to Account for 'opportunity of use' in Learner Corpora*, in *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*, Y. Tono, Y. Kawaguchi, and M. Minegishi, Editors. 2012, John Benjamins: Amsterdam. p. 187-204.
6. Meunier, F., et al. *The LONGDALE (Longitudinal Database of Learner English)*. 2010 [cited 2012 14 September]; Available from: <http://www.uclouvain.be/en-cecl-longdale.html>
7. Diez-Bedmar, M.B., *Written Learner Corpora by Spanish Students of English: an overview*, in *A Survey on Corpus-based Research. Proceedings of the AELINCO Conference*, P.C. Gómez and A.S. Pére, Editors. 2009, Asociación Española de Lingüística del Corpus: Murcia. p. 920-933.
8. Hammarberg, B., *Introduction to the ASU Corpus, a Longitudinal Oral and Written Text Corpus of Adult Learners' Swedish with a Corresponding Part from Native Swedes*. 2010, Stockholm University: Department of Linguistics.
9. Dagneaux, E., et al., *Error tagging manual*. 1996.
10. Granger, S., *Error-tagged Learner Corpora and CALL: A Promising Synergy*. CALICO Journal, 2003. 20(3): p. 465-480.
11. Nicholls, D. *The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT*. in *Corpus Linguistics 2003 Conference (CL 2003)*. 2003. Lancaster, UK.
12. Izumi, E., K. Uchimoto, and H. Isahara. *Error annotation for corpus of Japanese learner English*. in *Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005), 15 October 2005*. 2005. Jeju Island, Korea.
13. Alosaili, A.I., الأخطاء الشائعة في الكلام لدى طلاب اللغة العربية الناطقين بلغات أخرى: دراسة وصفية تحليلية (in Arabic) "*Common Errors in Speech Production of Non-Native Arabic Learners*". 1985, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
14. Alateeq, Z.M., تحليل الأخطاء الدلالية لدى دارسي اللغة العربية من غير الناطقين بها في مادة التعبير الكتابي (in Arabic) "*Semantic Errors Analysis of Non-Native Arabic Learners in Writing*". 1992, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
15. Alhamad, M.M., تحليل أخطاء التعبير الكتابي لدى المستوى المتقدم من دارسي العربية غير الناطقين بها في جامعة الملك سعود (in Arabic) "*Writing Errors Analysis of Advanced-Level Arabic Learners at King Saud University*". 1994, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
16. Alaqeeli, A.S., تحليل الأخطاء في بعض أنماط الجملة الفعلية للغة العربية في الأداء الكتابي لدى دارسي المستوى المتقدم (in Arabic) "*Error Analysis in Some Verbal Sentence Patterns of Arabic in Writing Production of Advanced-Level Learners*". 1995, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.