



UNIVERSITY OF LEEDS

This is a repository copy of *An evaluation of the Arabic error tagset v2*.

White Rose Research Online URL for this paper:

<http://eprints.whiterose.ac.uk/81592/>

---

**Proceedings Paper:**

Alfaifi, A and Atwell, ES (2014) An evaluation of the Arabic error tagset v2. In: Proceedings of the ACL 2014 - The American Association for Corpus Linguistics conference. ACL 2014 - The American Association for Corpus Linguistics conference, 26 - 28 Sep 2014, Flagstaff, USA. The American Association for Corpus Linguistics .

---

**Reuse**

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# An Evaluation of the Arabic Error Tagset v2

Abdullah Alfaifi

University of Leeds

[scayga@leeds.ac.uk](mailto:scayga@leeds.ac.uk)

Eric Atwell

University of Leeds

[e.s.atwell@leeds.ac.uk](mailto:e.s.atwell@leeds.ac.uk)

## Abstract

A survey of the literature shows that annotating errors of Arabic learners has not received much attention, and there is a need for a practical error tagset which can be used for Arabic learner corpora. This type of tagset is used in such corpora for several purposes, e.g., Contrastive Interlanguage Analysis (CIA), learner dictionary making, Second Language Acquisition, designing pedagogical materials, etc.

This paper evaluates the second version of a two-level error tagset developed for annotating the Arabic Learner Corpus (ALC). It includes six broad classes, subdivided into more specific error types. The paper shows the tagset, and an example of the annotation method used for tagging the ALC.

The inter-annotator agreement using the current revised version of the error tagset was higher compared to the first version (Alfaifi *et al.*, 2013). Four factors assisted in reaching this level of accuracy: (1) the tagset was reviewed by two experts in Arabic language, (2) the annotators were given texts with errors already identified, so their task was to classify and mark the appropriate tag on each error, (3) the annotators were trained during the experiment, (4) an error tagging manual was created which explains all error types in the tagset with rules and examples of how to tag learners' errors.

Two lists of varied sentences, 100 in each, were tagged for errors by three annotators; after tagging the first list they discussed their work to provide them with suitable training, and this allowed us to distinguish the value of the training among the other factors.

## Reference

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* (9): 14-22.