Annotating an Arabic Learner Corpus for Error

Ghazi Abuhakema, Reem Faraj, Anna Feldman, Eileen Fitzpatrick

Montclair State University

Montclair NJ 07043

Email: abuhakemag@mail.montclair.edu, farajr@mail.montclair.edu, feldmana@mail.montclair.edu, farajr@mail.montclair.edu, farajr

fitzpatricke@mail.montclair.edu

Abstract

This paper describes an ongoing project in which we are collecting a learner corpus of Arabic, developing a tagset for error annotation and performing Computer-aided Error Analysis (CEA) on the data. We adapted the French Interlanguage Database FRIDA tagset (Granger, 2003a) to the data. We chose FRIDA in order to follow a known standard and to see whether the changes needed to move from a French to an Arabic tagset would give us a measure of the distance between the two languages with respect to learner difficulty. The current collection of texts, which is constantly growing, contains intermediate and advanced-level student writings. We describe the need for such corpora, the learner data we have collected and the tagset we have developed. We also describe the error frequency distribution of both proficiency levels and the ongoing work.

1. Introduction

We describe a pilot study in which we developed a tagset for error-annotation of Arabic learner data. We compiled a small pilot corpus of Arabic learner written productions and adapted the French Interlanguage Database FRIDA tagset (Granger, 2003a) to the data. We chose FRIDA in order to follow a known standard and to see whether the changes needed to move from a French to an Arabic tagset would give us a measure of the distance between the two languages with respect to learner difficulty.

2. Language Learner Corpora

Computer Learner Corpus research is grounded in both corpus linguistics and Second Language Acquisition (SLA) studies. It uses the methods and tools of corpus linguistics to gain better insight into authentic learner language at different levels – lexis, grammar, and discourse. (Pravec, 2002; Granger, 2003b).

2.1 Contrastive Interlanguage Analysis (CIA)

Learner corpus research has concentrated on Contrastive Interlanguage Analysis (CIA), which involves two types of comparison -1) native productions (NS) vs. non-native productions (NNS) to highlight the features of non-nativeness in the learner language; 2) two or more varieties of NNS to determine whether non-native features are limited to one group of learners, in which case they are most probably transfer-related phenomena, or whether they are shared by several groups of learners, which would point to a developmental, or interlanguage, issue.

2.2 Computer-Aided Error Analysis

Computer-aided Error Analysis (CEA) has led to a much more limited number of publications than CIA due to the cost of manual error annotation. Apart from articles describing error tagging systems, there are a few articles covering certain specific error categories including lexical errors (Man-Lai et al., 1994; Källkvist, 1995; Lenko-Szymanska, 2003), tense errors (Granger, 1999; Fitzpatrick and Seegmiller, 2004) and a more recent article (Neff et al., 2007) covering the range of error types in the ICLE corpus from Spain. These analyses offer great promise for identifying the sources of error (L1 interference, features of novice writing, limited vocabulary and language structure, etc.) so the need to annotate for error and to reduce the cost of annotation by automating where possible is great.

3. Error Tagging

There are two ways to annotate learner data for error. One approach is to reconstruct the correct form (e.g. Fitzpatrick and Seegmiller, 2001). The other approach is to mark different types of errors with special tags (e.g. Granger, 2003a). The former is used for developing instructional materials that can provide (automatic) feedback to learners; the latter is used for SLA research to compare type of error and error frequency among different learners at different levels of language development.

3.1 Applications of Error Tagging

Error tagging is a highly time- and labor-consuming task. At the same time, a corpus annotated for error provides an invaluable resource for SLA research and practice. For researchers, errors can reveal much about the process by which L2 is acquired and the kinds of strategies the learners use in that process. For language instructors, errors can give hints about the extent to which learners have acquired the language system and what they still need to learn. Finally, for learners themselves, access to the data marked for error provides important feedback for improvement.

3.2 FRIDA (French Interlanguage Database)

Error tagging in FRIDA implements both reconstruction

and tagsets. To develop an error tagset for learner Arabic, we adapted the FRIDA tagset designed specifically for French. We chose FRIDA because of the explicit description of the tags in Granger (2003a). FRIDA is a three-level error annotation system, with 9 domains, 36 error categories and 54 word categories. The domain level is the most general: it specifies whether the error concerns typography and spelling, morphology, grammar, lexis, syntax, punctuation, register, or style. Each error domain is subdivided into a variable number of error categories. For example, the lexical domain L groups all lexical errors due to: 1) insufficient knowledge of the conceptual meaning of words; 2) violations of the co-occurrence patterns of words; 3) violations of the grammatical complementation patterns of words. The word categories (adjective, adverb, article, etc.) are subdivided into 54 subcategories, such as 'simple, comparative, superlative, complex for adjective errors. This particular tier makes it possible to sort errors by grammatical category and to draw up a list of relevant errors for each category.

4. A Pilot Arabic Learner Corpus

To the best of our knowledge, there are no learner Arabic corpora available for public use. Prior lack of interest in Arabic as a foreign language, the existence of more than thirty dialects and subdialects of the language, and previous technical difficulties with non-roman scripts have meant that resources for the systematic investigation of the acquisition of Arabic by non-native speakers are extremely scarce. Currently, not only is there a lack of learner corpora resources for critical languages, but there is no portable software that can be easily adapted to generate instructional materials automatically based on specified criteria, such as the level of linguistic complexity, different levels of competence, genre, target linguistic structure or discourse style. The current demand for the rapid generation of teaching materials for Arabic makes the creation and internet dissemination of a learner corpus such as this a critical need.

5. Error Annotation of Arabic

5.1 Linguistic Properties of Arabic Relevant to Error Tagging

The most salient difference between French and Arabic is in the basic word formation process, French being a stem and affix language and Arabic being a triliteral root language. However, like French, Arabic has inflectional affixes that mark gender, person, number, tense, etc. In addition, there are general errors that will be present for all L2s, e.g., errors involving word order, missing or confused elements, and spelling.

5.2 The Learner Data

We have analyzed eight different texts written by learners of Arabic as a Foreign Language. The level of the students was either intermediate (3818 words) or advanced (4741 words). The students are American native speakers of English who studied Arabic in an intensive program and then went to study abroad in Arab countries. Some of the texts were written during their study in the United States and others represent their writing while abroad.

For this pilot study, the tagset was developed by one author and applied by this author and a second author on different data in order to test the coverage of the tags. Once the tagset is complete, we will test for interrater reliability.

5.3 The FRIDA Tagset Applied to Arabic

We have adopted FRIDA's first level of tagging with only one addition: *diglossia*, a common error when students are exposed to the many dialects of Arabic. For the second level, we deleted some tags and added others. The tags that we dropped include upper/lower case, auxiliary and euphony (Arabic does not have these), diacritics, and homonymy, which will only occur in fully voweled texts and do not appear in learner writing. We do not anticipate using these tags on a larger scale set.

In terms of phonology, we added the long/short vowel distinction, emphatic/non-emphatic consonants, nunation (a mark of indefiniteness), hamza (a glottal stop that learners often do not hear), and shadda (consonant doubling). In terms of morphology and syntax, we added infixation, verb pattern confusion, negation (Arabic has several negation particles based on the form of the sentence and verb tense), and definite and indefinite structure (different from (in)definite agreement). The phenomenon of partial, or weak, agreement in Arabic caused us to modify the tagset to include full inflection, partial inflection, and zero inflection, which FRIDA does not need for French. We also made minor modifications to gender agreement, (in)definite agreement, and number agreement. In terms of style, we kept 'heavy', though we found no instances of turgid writing in our samples. We added 'pallid', for writing that is oversimplified.

We also anticipate that we will need more tags as we deal with texts of beginning and highly advanced learners. Additionally, as we apply FRIDA's third tagging level, we anticipate that we will need to adjust it to fulfill particular needs the corpus will dictate.

5.4 The Tagset for Learner Arabic

Table 1 shows the Arabic tagset we are currently using. The first column shows the error domains while the second demonstrates the error categories. For the tags

themselves, we either used the initial(s) and/or the root or part of the root of the word that represents each domain and category. The tags use the Arabic script and appear in brackets in the table.

فات الإغطاء مجالات الأغطاء Form/spelling «التشريك حريف الملة الطريلة والقصيرة حلقة» Agglutination الحريف الملة الطريلة Vowel length confusion alba الطريلة (علقة) (حافي) Emphatic/non emphatic consonants Emphatic/non emphatic consonants الحريف المفخمة والمرققة <حخر) Consonant doubling (shaddat) الشدة حشد) Vowel length confusion المعرفة حفر) Consonant doubling (shaddat) المعرفة حفر) Consonant doubling (shaddat) المعرفة حفر) Other spelling errors (Glottal stop Cerivation-prefixation الإشتقاق – اللاحقة حشقے> Derivation-prefixation (Morphology Derivation-infixation (Yhitäbi – lubi حفر حشقے> Inflection – full المنصرف حصرف> Inflection – partial (لمنصر في المنصرف حصرفي Inflection – full (Adj. complement حمنهي (Gender agreement حمنه (Jhaba دنوع) Iberinte/Indefinite agreement (addi si bi الحرين حظقے> Iberlinite/Indefinite agreement (addi si bi الحرين حظقے> Iberlinite/Indefinite agreement (addi si bi الحرين حظقے> Iberlubai Eace	Error Domains	Error Categories		
Form/spelling Agglutination Itaid بين حروف العلة الطويلة elibad بين حروف العلة الطويلة elibad بين حروف العلة الطويلة emphatic/non emphatic consonants (consonants elucitis <حخر) Emphatic/non emphatic consonants Emphatic/non emphatic consonants (Consonant doubling (shaddat) exels Consonant doubling (shaddat) (Sources) Entraction (Sources) Glottal stop (Sources) Derivation-prefixation (Sources) Derivation-suffixation (Sources) Derivation-suffixation (Nuration) Derivation-suffixation (Sources) Derivation-full (Morphology) Derivation-infixation (Sources) Derivation-suffixation (Sources) Derivation-suffixation (Sources) Inflection - full (Adition) Adition - zero (Sources) Inflection confusion (Sources) Inflection confusion (Sources) Inflectinite agreement (Sources) Inflection confusion (Source) Inflection confusion (Source) Inflection confusion (Sourceseques) Infleat <ci>(Source</ci>	مجالات الأخطاء	فنات الاخطاء		
التسبيلة حسنة >> التسبيلة حريف العلة الطريلة والقصيرة حرعة، الجلرية الطريلة والقصيرة درعاة، Emphatic/non emphatic consonants () والمدينة، والمراققة حجر) Consonant doubling (shaddat) التعريف المغضة والمراققة حجر) Consonant doubling (shaddat) التعريف حيون Consonant doubling (shaddat) () Derivation () Derivation () Derivation-prefixation () Derivation-suffixation () Derivation-infixation () Inflection - partial () Inflection - partial () Inflection - zero () Inflection confusion () Inflection confusion () Class (POS) () Indexis & but set eds () Gender agreement () Q Hothofinite agreement <tr< th=""><th>Form/spelling</th><th>Agglutination</th></tr<>	Form/spelling	Agglutination		
Vowel length confusion الخلط بين حروف العلة الطويلة والقصيرة < علة> Emphatic/non emphatic consonants (consonant doubling (shaddat) (staddat) (consonant doubling (shaddat) (consonant doubling (shaddat) (consonant doubling (shaddat) (consonant doubling (shaddat) (consonant doubling errors (consonant fixed) (consonant fixed) (consonant fixed) (consonant doubling errors (conson-prefixation (conson-prefixation (conson-suffixation (conson-suffixation (conson-suffixation (conson-suffixation (conson-prefixation (conson-prefixation (conson-suffixation (conson-suffixation)<	الشكل <ش>	النسبيك <سبك >		
الخلط بين حروف العلة الطويلة والقصيرة < حفاتي Emphatic/non emphatic consonants () الحروف المغذمة والمرققة < حخار) الشدة < شدي Consonant doubling (shaddat) ()		Vowel length confusion		
والقصيرة < <d>والمصيرة </d> Emphatic/non emphatic consonants Emphatic/non emphatic consonants (consonant doubling (shaddat) Consonant doubling (shaddat) (consonant doubling estation) Consonant doubling (shaddat) (consonant doubling estation) Consonant doubling estation (consonant doubling errors) Consonant estation (conson estation) Conson estation (conson estation) Consonant estation (conson estation) Consonant estation (conson estation) Class (POS) (conson estation)		الخلط بين حروف العلة الطويلة		
Emphatic/non emphatic consonants الحروف المغخمة و المرقفة <حخر) Consonant doubling (shaddat) Nunation التعرين <ثون> Glottal stop Image: Series Other spelling errors Consonsuffixation Ivation Verivation-prefixation Verivation-suffixation Verivation-suffixation Verivation-infixation Verivation-full Inflection – full Inflection – partial (inflection confusion <a 2"="" href="splittike-stage-s</th><th></th><th>والقصيرة <علة></th></tr><tr><th>consonants الحروف المغخمة و المرققة <حخر)</td> Consonant doubling (shaddat) الشدة <شده</td> Nunation Glottal stop Bayes Other spelling errors المعابة الحرى <خهج</td> Morphology Derivation-prefixation الاشتقاق البادئة <شقح> Derivation-prefixation الاشتقاق المتوسطة <شقح> Inflection – full المتوسطة <شقح</td> Inflection – full مالمنصرف <صرف> Inflection – partial (منع عن الصرف) Inflection confusion نوع الكلمة <نوع</td> Inflection confusion نوع الكلمة <نوع</td> Inflection confusion نوع الكلمة <iوع> Indutias is الجنس <di>Eas (POS) ig a lack <di>gender agreement نوع الكلمة <ig > Indutias is الجنس <di>gender agreement نوع الكلمة <ig > Indutias is الجنس <di>gender agreement is a liter Indutias is a liter <g > Indutias is a liter <g > Indutias is a liter <g > Indutis is a liter <g > In</th><th></th><th>Emphatic/non emphatic</th></tr><tr><th>الحروف المفضة والمرققة <حخر)</td> Consonant doubling (shaddat) التندة <شدد> Nunation التنوين <نون> Glottal stop Other spelling errors المحروف المغتير Other spelling errors Image: Construction Image: Construction Other spelling errors Image: Construction Image: Cons</th><th></th><th>consonants</th></tr><tr><th>Consonant doubling (shaddat) الشدة <شدد> Illing cross Image: Second Sec</th><th></th><th>الحروف المفخمة والمرققة ححخر)</th></tr><tr><th>الشدة <شدد> التنوين <نون> Image: Construct of the set of the s</th><th></th><th>Consonant doubling (shaddat)</th></tr><tr><th>التنوين <نون> Ititication العمرة <همز> Glottal stop العمرة <همز> Other spelling errors اخطاء هجائية أخرى <خهج> الصرف <ص> Morphology Derivation-prefixation الاشتقاق - اللاحقة <شقب> Inflection - suffixation الاشتقاق - اللاحقة <شقب</td> الاشتقاق المئوسطة <شقب</td> Derivation-suffixation الاشتقاق المئوسطة <شقب</td> Derivation-infixation الاشتقاق المئوسطة <شقب</td> Inflection – full المنصرف <صرف</td> المنصرف <صرف</td> المنصرف <صرف</td> المنصرف <صرف</td> المنصرف <صرف</td> المنصرف <صرف</td> المنصرف <صرف</td> المنصرف المنصرف المادو عان الصرف المعارف Inflection – partial المبنى (Lass (POS) نوع الكلمة <نوع</td> الموابقة في التمريف <ظفي > المطابقة في التعريف <طقع</td> Definite/Indefinite agreement نوع الكلمة <نوع</td> الموابقة في العرد المطابقة في العدر خلغ Number agreement نوع الكلمة <i/td> الموابقة في العدر المطابقة في العدر Number agreement نوع الكلمة الموابقة في العدر المطابقة في العدر Negation</td</th><th></th><th>الشدة <شدد></th></tr><tr><th>التنوين <نون> Glottal stop الهمزة <همز> Other spelling errors أخطاء هجائية أخرى <خهج> Morphology Derivation-prefixation الاشتقاق البادئة <شقب> Derivation-suffixation الاشتقاق البادئة <شقب> Derivation-suffixation الاشتقاق المتوسطة <شقب</td> Derivation-infixation الاشتقاق المتوسطة <شقب</td> Inflection – full المنصرف <صرف> المنصرف فير المنصرف (الممنوع من الصرف) Inflection – partial (منج > من الصرف) Inflection – zero المبني <بني> Inflection confusion نوع الكلمة <نوع> نوع الكلمة <iva >> Gender agreement نوع الكلمة <iva >> المطابقة في الجنس <طقح> Indulatis في الحد <طقع</td> Indulatis في الحد <di> المطابقة في الحد <di> المطابقة في الحد <طقع</td> الملغ <</td> الملغ <</td></th><th></th><th>Nunation</th></tr><tr><th>Glottal stop الهمزة <همز > Other spelling errors أخطاء هجائية أخرى <خهج</td> Morphology Derivation-prefixation الاشتقاق-البادنة <شقب</td> Derivation-suffixation الاشتقاق المتوسطة <شقب</td> Derivation-infixation الاشتقاق المتوسطة <شقب</td> Derivation-infixation الاشتقاق المتوسطة <شقب</td> Inflection – full المنصرف <صرف> فير المنصرف (الممنوع من الصرف) Inflection – partial (في التصريف (الممنوع من الصرف) Inflection confusion فير المنصرف (الممنوع من الصرف) Inflection confusion فو ع الكلمة <نوع</td> فو ع الكلمة <i وع</td> Grammar Class (POS) فو ع الكلمة <i وع<> فو ع الكلمة حنوع> Indelias في التصريف حظفي Indelias في التحريف <di> المطابقة في الحريف <di> Voice المطابقة في العدد <طقع<</td> المطابقة في العدد <طقع</td> المطبقة في العدد <طقع</td> المطبقة في العدد <طقع</td> المطبقة في العدد <طقع</td> المعاوم و المجهول <معج</td> المنو داني</td</th><th></th><th>التنوين <نون></th></tr><tr><th>الهمزة <همز> Other spelling errors أخطاء هجائية أخرى خفج الخطاء هجائية أخرى خفج Morphology Derivation-prefixation الاشتقاق-البادئة <شقب</td> Derivation الاشتقاق المئوسطة <شقح</td> Derivation الاشتقاق المئوسطة <شقح</td> Derivation الاشتقاق المئوسطة <شقح</td> Derivation الاشتقاق المئوسطة <شقح</td> Derivation المنصرف <صرف</td> الماصرف (الممنوع من الصرف) المنصرف <صرف من الصرف)</td> Inflection – full منح منح فير المنصرف (الممنوع من الصرف) Inflection – full المنصرف (حمرف من الصرف) المواح من الصرف) المواح من الصرف من الصرف) Sender agreence المواح من الصرف (الممنوع من الصرف) Inflection confusion المواح من الصرف (حمرع) Sender agreement المطبقة في الجنس حظقج Sender agreement المطبقة في الجنس حظقج Definite/Indefinite agreement المطبقة في الحريف حطقج Sender agreement المطبقة في الحر (حميخ) Sender agreement المطبقة في الحري (حمج) Sender agreement المطبقة في الحريف حظفح Sender agreement المطبقة من الصيغة (صيغ) Sender agreement</th><th></th><th>Glottal stop</th></tr><tr><th>Other spelling errors أخطاء هجائية أخرى <خهج</td> Morphology Operivation-prefixation الإشتقاق-الباذئة <شقب> Derivation-suffixation Operivation-suffixation Operivation-suffixation Derivation-infixation Inflection - full</th><th></th><th>الهمزة <همز></th></tr><tr><th>أخطاء هجائية أخرى خفيج> أخطاء هجائية أخرى خفيج> Morphology Derivation-prefixation الإشتقاق البادئة حشقب> Derivation-suffixation الإشتقاق المتوسطة حشقم> Derivation - full الإشتقاق المتوسطة حشقم> Inflection – full المنصرف حصرف> Inflection – full المنصرف حصرف> المنصرف حصرف المنصرف حصرف> Inflection – full منع المنصرف حصرف> المنصرف حصرف Inflection – full منع المنصرف حصرف المنصرف حصرف Inflection – full منع المنح عن الصرف) المنصرف حضرف Inflection – partial (نوع الكلمة حنوع) Inflection confusion نوع الكلمة حنوع> القواعد حق Indetisit في الخس حطقج Gender agreement نوع الكلمة حنوع> القواعد حق Indetisit في الخس حطقج Number agreement المطابقة في الجنس حطقج المطابقة في الجنس حطقج المطابقة في الجنس حطقج Number agreement المطابقة في الجنس حطقج المطابقة في الحد حلقع المطابقة في الحد حلقع المطابقة في الحد حلقع المطابقة في الحد حلقع المطابق حليه <td</th><th></th><th>Other spelling errors</th></tr><tr><th>Morphology Derivation-prefixation Query and an experience of the second s</th><th></th><th>أخطاء هجائبة أخرى <خهج></th></tr><tr><th>المنافر المنافر المنافر البادئة حشق المرف (م) الاشتقاق – البادئة حشقب Derivation-suffixation الاشتقاق – اللاحقة حشقج Derivation-infixation الاشتقاق المتوسطة حشقم Inflection – full المنصر ف حصر ف المنصر ف (المعنو ع من الصر ف) المنصر ف (المعنو ع من الصر ف) المنصر ف (المعنو ع من الصر ف) المالغة في التصريف (حصر ف Inflection confusion الخلط في التصريف (Class (POS) نوع الكلمة حنوع Grammar Class (POS) نوع الكلمة حنوع المطابقة في الجنس حطقج المطابقة في الجنس حطقج المطابقة في العدد حطقے المطابقة في العدد حطقے المطابقة في العدد حطقع المطابقة في العدد حطقے المطابقة في العدد حطقے المطابقة في العدد حطقے المطبقة في العدد حطقے المطبقة في العدد حطقے المطبقة حيز المطبقة حصيغ المنه حنه المغور دامع حلي المغور</th><th>Morphology</th><th>Derivation-prefixation</th></tr><tr><th>Derivation-suffixation الاشتقاق – اللاحقة Inflection – infixation Inflection – full Inflection – full المنصرف <صرف> غير المنصرف (المعنوع من الصرف) Inflection – partial (في المنصرف (المعنوع من الصرف) فير المنصرف (المعنوع من الصرف) Inflection – partial (في المنصرف (المعنوع من الصرف) Inflection – zero Inflection confusion في التصريف (POS) ie 3 الكلمة <ie 3> Grammar Class (POS) ie 3 الكلمة <ie 3> Indelبقة في الجنس <di scie 4> Number agreement المطابقة في العدد <طقع> Indelبقة في العدد <طقع </td> Indelبقة خصيع Indelبقة Indel مو المجهول Indel مو المجهول Indita <t</th><th>الصدف <ص></th><th>الاشتقاق- البادئة <شقب></th></tr><tr><th>الاشتقاق – اللاحقة <شقح> Derivation infixation الاشتقاق المتوسطة <شقم> Inflection – full المنصرف <صرف> Inflection – partial Inflection – partial فير المنصرف (الممنوع من الصرف) Inflection – partial المبني <یني> Inflection – partial المبني <یني> Inflection – zero المبني <یني> Inflection confusion نوع الكلمة <ti>(Class (POS) نوع الكلمة <ti>(Class (POS) نوع الكلمة <ti>(Class (POS) نوع الكلمة <ti>(Definite/Indefinite agreement المطابقة في الجنس <طقح> Indelias في الجنس <طقح</td> Number agreement المطابقة في العدد <طقع</td> Indulation Number agreement المطابقة في العدد <طقع</td> Indulation Number agreement المطابقة في العدد <طقع</td> Induction Induction المطابقة في العدد <طقع</td> Induction Induction المطبقة خي العدد <طقع</td> Induction الموز دان خوع الموز دان خو</th><th></th><th>Derivation-suffixation</th></tr><tr><th>Derivation-infixation Inflection - full Inflection - full المنصرف <صرف> غير المنصرف (الممنوع من الصرف) inflection - partial (inflection - zero Inflection confusion ibit ibit Grammar Class (POS) ibit ibit Class (POS) ibit ibit Class (POS) ibit ibit Class (POS) ibit ibit Class (POS) ibit Indeltäs baj lit Indeltäs baj lit Number agreement ibit ibit Induit Induit Induit Induit Induit Induit Induit Induit</th><th></th><th>الاشتقاق – اللاحقة حشقح></th></tr><tr><th>الاشتقاق المتوسطة المنصولة المنصرف <صرف> Inflection – full المنصرف <صرف> المانصرف <صرف</td> غير المنصرف (الممنوع من الصرف) غير المنصرف (الممنوع من الصرف) غير المنصرف (الممنوع من الصرف) غير المنصرف (الممنوع من الصرف) المبني الموابق المالية Inflection confusion نوع الكلمة القواعد <ق> Grammar Class (POS) نوع الكلمة القواعد <ق> Gender agreement المطابقة في الجنس <طقج</td> المطابقة في الجنس <طقج</td> Inflectinite agreement المطابقة في الجنس <طقج</td> Number agreement المطابقة في الجنس <طقج</td> Number agreement المطابقة في الجنس <طقع</td> المطابقة في الجنس <طقج</td> المطابقة في الجنس <طقع</td> المطابقة في الجنس <طقع</td> المطابقة في الجنس <طقع</td> المطابقة في الجنس <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدم <طقع</td> المطابقة في العد <طقع</td> المطابقة في العدم <طقع</td> المعني المعنوع <طقع</td> المعني</th><th></th><th>Derivation-infixation</th></tr><tr><th>Inflection - full Inflection - full المنصرف <صرف> غير المنصرف (الممنوع من الصرف) عزر المنصرف (الممنوع من الصرف) Inflection - partial (فعه) المبني <بني> Inflection - zero المبني <بني> Inflection confusion في التصريف <خرف> Inflection confusion في التصريف <خرف> Inflection confusion في الخلط في التصريف <خرف</td> Indelبقة في الجنس <طقح> Gender agreement المطابقة في الجنس <طقح</td> Indelبقة في العدد <طقع</td> Inflexib Inflexib Inflexib Inflexib</th><th></th><th>Derivation-initiation</th></tr><tr><th>Inflection – Infl المنصرف <صرف> المنصرف <صرف</td> غير المنصرف (الممنوع من الصرف) المناحي المناحي المنصرف <صرف</td> المنصرف (الممنوع من الصرف) المنصرف <صرف</td> المناحي المناحي المناحي المنصرف <ضرف> Inflection – zero الملابق خياب Grammar Class (POS) نوع الكلمة <نوع> Indelبقة في الجنس <طقج</td> Indelبقة في الجنس <طقج</td> Indelبقة في الجنس <طقج</td> المطابقة في الجنس <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطابقة في العدد <طقع</td> المطبقة في العدد <طقع</td> المطبقة خصيغ المطبقة خي العد <لقع</td> الموز دان خي > الموز دان خي > الموز دان خي ></th><th></th><th>Inflaction full</th></tr><tr><th>Inflection – partial غير المنصرف (الممنوع من الصرف) حمنع> Inflection – zero Inflection confusion الخلط في التصريف <خرف> Inflection confusion نوع الكلمة <i وع</td> Grammar Class (POS) iega lلكلمة <i وع> Indelبقة في الجنس <dēs> Indelبقة في الجنس <dēs</td> Indelبقة في التعريف <طقع> Indelبقة في التعريف <طقع</td> Indelبقة في التعريف <طقع</td> Indelبقة في التعريف <طقع</td> Indelبقة في التعريف <طقع</td> Indelبقة في العدد <طقع</td> Indelبقة في العدد <طقع</td> Indelبقة في العدد <dēs</td> Indelبقة ذي العدد <dēs</td> Intai Number agreement Intais Negation Intais Negation Intais Intais Intais Adj. complementation منعي Adj. complementation</th><th></th><th>المنصر ف <صر ف></th></tr><tr><th>غير المنصرف (الممنوع من الصرف)
جمنع>
Inflection – zero
المبني <يني>
Inflection confusion
الخلط في التصريف <خرف>
Grammar
Class (POS)
iega lلكلمة <iega>
Gender agreement
iega lلكلمة <iega>
Gender agreement
المطابقة في الجنس <طفح>
Number agreement
Indelبقة في التعريف <طفع>
Number agreement
Indelبقة في العدد <طفع>
Number agreement
Indelبقة في العدد <طفع>
Number agreement
Indelبقة في العدد <طفع>
Indelبقة في العدد <طفع>
Lexis
Meaning
<i complementation
Indel, complementation
Adj. complementation</th><th></th><th>Inflection – partial</th></tr><tr><th>المبني جيني Inflection – zero المبني جيني Inflection confusion الخلط في التصريف <خرف> Grammar Class (POS) نوع الكلمة <ie ع> Ibe last (POS) indeltas bag Gender agreement Indeltas bag Gender agreement Indeltas bag Definite/Indefinite agreement Indeltas bag Number agreement Indeltas bag Tense Indeltas bag Voice Indeltas bag Indeltas bag Indeltas bag Ibe (Intersetas) Indeltas bag Intersetas) Indeltas bag Intersetas) Indeltas (Intersetas) Intersetas) Interse Intersetas) Interse Intersetas) Interse Interse Interse Intersetas) Interse Intersetas) Interse Interse Interse Interse Interse I</th><th></th><th>غير المنصر ف (الممنوع من الصرف)</th></tr><tr><th>Inflection – zero Inflection confusion ical confusion ical confusion Indeltation Indeltatin Indite </</th><th></th><th><منع></th></tr><tr><th>المبني <بني> Inflection confusion الخلط في التصريف <خرف> Grammar Class (POS) ie 3 نوع الكلمة <ie 3> Gender agreement المطابقة في الجنس <طقج</td> Definite/Indefinite agreement المطابقة في التعريف <طقج</td> Number agreement المطابقة في التعريف <طقع</td> المطابقة في العدد <طقع</td> المعاد < والمجهول <معج> المبني للمعلوم و المجهول <معج> المني <الفي <القي <القي </td> المن المن المن المن الفق المن المغر المغر المغر المغر المغر المغر المغر الفق الموابق</th><th></th><th>Inflection – zero</th></tr><tr><th>Inflection confusion الخلط في التصريف <خرف> Grammar Class (POS) ie 3 ibela <</td> ie 3 ibela <</td> ie 3 ibela <</td> indeliais 6 ibela indeliais 6</td</th><th></th><th>المبنى حبنى></th></tr><tr><th>الخلط في التصريف <خرف> Grammar Class (POS) نوع الكلمة <نوع> Ibela et ego Gender agreement المطابقة في الجنس <طقح> Definite/Indefinite agreement المطابقة في التعريف <طقع> Number agreement المطابقة في التعريف <طقع> Number agreement المطابقة في التعريف <طقع</td> Tense المطابقة في العدد <طقع</td> Voice المبني للمعلوم و المجهول <معج> Negation النفي <نفي> Meaning المعنى <عني> المفردات <ك</td> متممة الصفة Adj. complementation</th><th></th><th>Inflection confusion</th></tr><tr><th>Grammar Class (POS) iega IlZhak Class (POS) iega IlZhak Gender agreement Indelitäk is in the the the the the the the the the the</th><th></th><th>الخلط في التصريف <خر ف></th></tr><tr><th>نوع الكلمة <نوع> القواعد حق>
Gender agreement
المطابقة في الجنس <طقج>
Definite/Indefinite agreement
المطابقة في التعريف <طقع>
Number agreement
Tense
Tense
Voice
المبني للمعلوم و المجهول <معج>
Negation
Lexis
Lexis
Meaning
حني>
Adj. complementation
متممة الصفة <صنيه</th><th>Grammar</th><th>Class (POS)</th></tr><tr><th>Gender agreement المطابقة في الجنس <طقج</td> Definite/Indefinite agreement المطابقة في التعريف <طقع</td> Number agreement المطابقة في العدد <طقع</td> Tense المطابقة في العدد <طقع</td> المعادم والمجهول <معج> Negation النفي <نفي> المغاد حاليه Adj. complementation متممة الصفة <</th><th>القواعد <ق></th><th>نوع الكلمة حنوع></th></tr><tr><th>المطابقة في الجنس <طقج>
Definite/Indefinite agreement
المطابقة في التعريف <طقع>
Number agreement
المطابقة في العدد <طقع
Tense
Tense
Voice
الميني للمعلوم والمجهول <معج>
Negation
Lexis
Meaning
حنيه > Adj. complementation
متممة الصفة <صتم></th><th></th><th>Gender agreement</th></tr><tr><th>للمطابقة في التعريف حطقي
المطابقة في التعريف حطقي
المطابقة في التعريف حطقي
المطابقة في العدد حطقي
المطابقة في العدد حطقي
الميني للمعلوم والمجهول حمعج
الميني للمعلوم والمجهول حمعج
الميني للمعلوم والمجهول حمعج
الميني للمعلوم والمجهول حمعج
الميني حلقي
متممة الصفة حصتم</th><th></th><th>المطابقة في الحنس حطقج></th></tr><tr><th>المطابقة في التعريف حطقع>
المطابقة في التعريف حطقع>
المطابقة في العدد حطقع>
المطابقة في العدد حطقع>
الميني للمعلوم والمجهول حمعج>
الميني للمعلوم والمجهول حمعج>
النفي حنفي>
النفي حنفي>
المعنى حعني>
المعنى حعني>
المعنى حيني>
المعنى حيني></th><th></th><th>Definite/Indefinite agreement</th></tr><tr><th>المطابقة في العدد رحلقع Number agreement Indulties في العدد رحلقع Tense I colspan=">Indulties في العدد رحلقع Voice I colspan="2">Indulties to the second sec		المطابقة في التعريف حطقع>		
المطابقة في العدد حطقع> Tense الصيغة حصيغ > Voice Voice Voice Negation Lexis Lexis Meaning حني> Adj. complementation متممة الصفة حصتي>		Number agreement		
Tense الصيغة حصيغ > المبني للمعلوم والمجهول <معج> Negation Lexis المعنى <عني> Adj. complementation متممة الصفة حصتم>		المطابقة في العدد حطقع>		
الصيغة <صيغ > Voice المبني للمعلوم والمجهول <معج> Negation Lexis المعنى <عني> Adj. complementation متممة الصفة <صتم>		Tense		
للمبني للمعلوم و المجهول <معج> المبني للمعلوم و المجهول <معج> النفي <نفي> Lexis المعنى <عني> Adj. complementation متممة الصفة <صتم>		الصبغة حصيغ >		
المبني للمعلوم و المجهول <معج> Negation النفي <نفي> Lexis المعنى <عني> Adj. complementation متممة الصفة <صتم>		Voice		
النفي <نفي> المعنى حياي المغريات المعنى حياي المعنى حياي المعنى حياي المغريات حياي المغريات حياي المغريات حياي المغريات حياي المغريات حياي المغريات حياي المعنى حي مالمعنى حياي المعنى حياي المع المع المعنى حياي		المبنى للمعلوم والمجهول <معج>		
النفي <نفي> Lexis Meaning المفردات <ك> Adj. complementation متممة الصفة <صتم>		Negation		
لexis Meaning المعنى <عني> المفردات <ك Adj. complementation متممة الصفة <صتم>		النفي <نفي>		
المعنى <عني> المعنى حيني> المفردات حك> Adj. complementation متممة الصفة حصتم>	Lexis	Meaning		
Adj. complementation	المفر دات <ك>	المعنى <عنى>		
متممة الصفة حصتم>		Adi, complementation		
		متممة الصفة حصتم>		

	N complementation			
	متممة الأسم <ستم>			
	V complementation			
	متممة الفعل <فتم>			
Syntax	Word order			
النحو <ن >	تيب الكلمات <رتب>			
	Word missing			
	كلمة مفقودة <كفق>			
	Word redundant			
	کلمة زائدة <کزد>			
	Cohesion			
	الترابط <ربط>			
Diglossia	Colloquial use			
ازدواجية اللغة <ز>	استخدام العامية <عمم>			
Style	Unclear			
الأسلوب <س>	غامض <عمض>			
	Simplistic			
	ركىك <>ركك			
	-1~~1			
Punctuation	Punctuation confusion			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل>			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل> Punctuation missing			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل> Punctuation missing علامة ترقيم مفقودة <طفق>			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل> Punctuation missing علامة ترقيم مفقودة <طفق> Punctuation redundant			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل> Punctuation missing علامة ترقيم مفقودة <طفق> Punctuation redundant علامة ترقيم زائدة <طزد>			
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل> Punctuation missing علامة ترقيم مفقودة <طفق> Punctuation redundant علامة ترقيم زائدة <طزد> حطبع>			

Table 1. The Error Tagset for Arabic

5.5 Evaluation

While our corpus was not large enough to test interrater reliability, our test of the tagset usability yielded results that will affect our work as we tag a larger corpus.

Each annotator covered only 500 words of text per hour due to the need to go up and down the levels of annotation to mark each error. A pull-down menu of tags at each level is planned to speed the annotation.

The frequency of error types based on student level already provides useful data for pedagogical purposes. Table 2 shows the most frequent errors by learner level.

Interr	med., wc= 3818	Advanc	ed, wc= 4741
31	<ك> < عني >	85	<كى >< عني >
	Meaning		Meaning
21	<ق>< طقج>	50	<ن ><رتب>
	Gender		Word order
	agreement		
15	< ن> < کفق>	44	<ن> <کزد>
	Word missing		Redundant
			word
14	<ش><همز>	31	<ن >< ک فق >
	Glottal stop		Word
	-		missing
14	حق> <طفق>	26	حق><طَقج >
	Punctuation		Gender
	missing		agreement

14	<ش ><خھج>	22	حق><صيغ>
	Other spelling		Tense
	mistakes		
10	<ص> <شقح >	17	<ش><خهج>
	Derivation-		Other
	suffixation		spelling
			mistakes
9	حق ><طقع>	15	حق><طقع>
	Number		Number
	agreement		agreement
9	حق> <رتب>	14	حق ><ربط>
	Word order		Cohesion
8	<ن><کر>	14	حق><نوع>
	Word missing		Class

Table 2. Most frequent errors by learner level.

5.6 Error Frequency Distributions

The tag frequency distributions are not surprising, but will be useful in terms of pedagogy. One notable difference between the intermediate and advanced writers is that the former are still struggling with phonological/orthographical issues (e.g., the glottal stops known as 'hamza', which are difficult to hear or involved in spelling rules) while the latter group have left these errors behind and are struggling, not surprisingly, with features of advanced writing like word order and cohesion. Both groups still have difficulties with lexis and the morphologically marked agreement.

6. Ongoing Work

Our intention is to test this tagset on our most elementary writing students' work and modify further if necessary. We will continue error tagging on the three levels of beginning, intermediate, and advanced, and make the tagged essays publicly available via the web for further second language acquisition analysis and design of pedagogical tools.

References

- Fitzpatrick, E. and Seegmiller, M.S. (2001). The Montclair Electronic Language Learner Database. In G. Antoniou and D. Deremer (ed.), *Proceedings of the International Conference on Computing and Information Technologies* (ICCIT).
- Fitzpatrick, E. and Seegmiller, M.S. (2004). The Montclair Electronic Language Database Project. In T. Upton and U. O'Connor, (ed.), Corpus Linguistics in North America 2002: Selections for the Fourth North American Symposium of the American Association for Applied Corpus Linguistics.
- Flowerdew, L. (1998). Application of Learner Corpus Based Findings and Methods to Pedagogy. In Proceedings of the First International symposium on

Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 38-44.

- Granger, S. (1999). Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus. In H. Hasselgerd and S. Oksefjell, (ed.), Out of Corpora: Studies in Honor of Stig Johansson. Amsterdam and Atlanta: Rodopi, pp. 19--202.
- Granger, S. (2003a). Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20(3), pp. 465–480.
- Granger, S. (2003b). The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, 37(3),538--546. Special issue on Corpus Linguistics.
- Källkvist, M. (1995). Lexical Errors Among Verbs: A Pilot Study of the Vocabulary of Advanced Swedish Learners of English. In U. Connor and T. Upton, editors, *Working Papers in English and Applied Linguistics*. University of Cambridge, pp. 103--115.
- Lenko-Szymanska, A. (2003). Lexical Problems in the Advanced Learner Corpus of Written Data. In Proceedings of Practical Applications of Language Corpora (PALC). Poland: Lodz.
- Man-Lai, A., et al. (1994). Collocational Problems Amongst ESL Learners. In L. Flowerdew et al., (ed.), Entering Text, pp. 157--165.
- Neff, J., et al. (2007). A Contrastive Functional Analysis of Errors in Spanish EFL University Writers Argumentative Texts: a Corpus-based Study. In E. Fitzpatrick, (ed.), Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse. New York: Rodopi, pp. 203--226.
- Pravec, N. (2002). Survey of Learner Corpora. ICAME Journal, 26:81--114.