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, 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.

Error Domains **مجالات الأخطاء**	**Error Categories** **فئات الأخطاء**
Form/spelling الشكل <ش>	Agglutination التشبيك <شبك >
	Vowel length confusion الخلط بين حروف العلة الطويلة والقصيرة <علة>
	Emphatic/non emphatic consonants الحروف المفخمة والمرققة <حخر)
	Consonant doubling (shaddat) الشدة <شدد>
	Nunation التنوين <نون>
	Glottal stop الهمزة <همز>
	Other spelling errors أخطاء هجائية أخرى <خهج>
Morphology الصرف <ص>	Derivation-prefixation الاشتقاق- البادئة <شقب>
	Derivation-suffixation الاشتقاق – اللاحقة <شقح>
	Derivation-infixation الاشتقاق المتوسطة <شقم>
	Inflection – full المنصرف <صرف>
	Inflection – partial غير المنصرف (الممنوع من الصرف) <منع>
	Inflection – zero المبني <بني>
	Inflection confusion الخلط في التصريف <خرف>
Grammar القواعد <ق>	Class (POS) نوع الكلمة <نوع>
	Gender agreement المطابقة في الجنس <طقج>
	Definite/Indefinite agreement المطابقة في التعريف <طقع>
	Number agreement المطابقة في العدد <طقع>
	Tense ا لصيغة <صيغ >
	Voice المبني للمعلوم والمجهول <معج>
	Negation النفي <نفي>
Lexis المفردات <ك>	Meaning المعنى <عني>
	Adj. complementation متممة الصفة <صتم>
	N complementation متممة الاسم <ستم>
	V complementation متممة الفعل <فتم>
Syntax النحو <ن >	Word order تيب الكلمات <رتب>
	Word missing كلمة مفقودة <كفق>
	Word redundant كلمة زائدة <كزد>
	Cohesion الترابط <ربط>
Diglossia ازدواجية اللغة <ز>	Colloquial use استخدام العامية <عمم>
Style الأسلوب <س>	Unclear غامض <عمض>
	Simplistic ركيك <>ركك
Punctuation علامات الترقيم <ط>	Punctuation confusion الخلط في الترقيم <طخل>
	Punctuation missing علامة ترقيم مفقودة <طفق>
	Punctuation redundant علامة ترقيم زائدة <طزد>
Typos أخطاء مطبعية <خ>	<طبع>

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.

Intermed., wc= 3818	
31	< عني > <ك> Meaning
21	<طقج> <ق > Gender agreement
15	<ن> < كفق> Word missing
14	<ش><همز> Glottal stop
14	<ق> <طفق> Punctuation missing

Advanced, wc= 4741	
85	<ك >< عني > Meaning
50	<ن ><رتب> Word order
44	<ن> <كزد> Redundant word
31	<ن ><كفق > Word missing
26	<ق><طقج > Gender agreement

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