ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study

Anna Feldman and **Ghazi Abuhakema** and **Eileen Fitzpatrick**
Montclair State University
Montclair, New Jersey, 07043, USA
{feldmana,abuhakemag,fitzpatricke}@mail.montclair.edu

Abstract

This paper describes a pilot study in which we collected a small learner corpus of Arabic, developed a tagset for error-annotation of Arabic learner data, tagged the data for error[1], and performed simple Computer-aided Error Analysis (CEA).

Language Learner Corpora and Applications

Learner corpora research uses the methods and tools of Second Language Acquisition (SLA) studies and corpus linguistics to gain better insights into authentic learner language at different levels – lexis, grammar, and discourse.

One application of learner corpora is Contrastive Interlanguage Analysis (CIA), which involves two types of comparison – 1) native speech (NS) vs. non-native speech (NNS) to highlight the features of nativeness and non-nativeness of learner language; 2) two or more varieties of NNS to determine whether non-native features are limited to one group of non-native speakers (in which case it is most probably a transfer-related phenomenon), or whether they are shared by several groups of learners with different mother tongue backgrounds (which would point to a developmental issue).

Another application is Computer-aided Error Analysis (CEA) for identifying the sources of error (L1 interference, features of novice writing in the new culture, limited vocabulary and language structure, etc.). For this application, error annotation is essential.

Error Tagging

There are two ways to annotate learner data for error. One approach is to reconstruct the correct form (Fitzpatrick & Seegmiller 1999). The other approach is to mark different types of errors with special tags (Granger 2003). 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. We have begun our study of learner Arabic using both reconstruction and special tag annotation, with the FRIDA (Granger 2003) tagset for French as our model for the latter approach.

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 instruction. For SLA researchers, errors can reveal much about the process by which the second language (L2) is acquired and the kinds of strategies or methodology 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.

A Pilot Arabic Learner Corpus

To the best of our knowledge, there are no learner Arabic corpora available for public use. In general, there is little research done in the area of data-driven instructional materials development. 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 in dealing with non-roman scripts have meant that resources for the systematic investigation of the acquisition of Arabic by non-native speakers are extremely scarce.

Error Annotation of Arabic

Linguistic Properties of Arabic and 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 trilateral 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.

The Learner Data

We have analyzed eight different texts written by learners of Arabic as a Foreign Language. The level of the texts ranged between intermediate (3,818 tokens) and advanced (4,741 tokens). The students are Americans whose native

[1] **We thank Reem Faraj for helping us with data annotation.**

language is English. They studied Arabic in an intensive program and then went to study abroad in Arab countries. Some of the texts were written during their study years in the United States and others represented their productions while studying abroad. The classification into the intermediate and advanced levels was done based on the guidelines provided by the American Council on the Teaching of Foreign Languages (ACFTL) to rate written texts.

The FRIDA Tagset Applied to Arabic

We have adopted FRIDA's highest level of tagging, the domain, with only one addition: *diglossia*, a common error when students are exposed to the many dialects of Arabic. For the intermediate level, the error categories, we deleted some tags and added others. The tags that we dropped include *upper/lower case*, and *auxiliary* (Arabic does not have them), *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, the phenomenon of partial or weak agreement in Arabic caused us to modify the tagset to include *full inflection*, *partial inflection*, *zero inflection*, which FRIDA does not need for French, as well as *infixation*, *gender agreement*, *(in)definite agreement*, *number agreement* (Arabic utilizes different types of agreement), and *negation* (there exist a few negation particles based on the form of the sentence and verb tense). In terms of syntax, we added *definite* and *indefinite structure* (different from *(in)definite agreement*), *verb pattern confusion*, and *word confusion*.

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.

The reader is invited to visit `http://chss.montclair.edu/~feldmana/publications/flairs21-data/` to see the details of the tagset. We anticipate that we will need to add more tags as we deal with texts of beginning and highly advanced learners. Additionally as we apply FRIDA's third tagging level, covering word categories, we anticipate that we will need to adjust it to fulfill particular needs the corpus will dictate in terms of adding, expanding or deleting tags.

Computer-Aided Error Analysis (CEA)

The data discussed here is available at `http://chss.montclair.edu/~feldmana/publications/flairs21-data/`.

Frequency of error types

The frequency of error types based on student level already provides useful data for pedagogical purposes. We classified the most frequent errors by learner level. One notable difference between the intermediate and advanced writers is that the former are still partly struggling with phonological/orthographic issues (the glottal stops known as 'hamza', for which these students have difficulty mastering the spelling rules or even hearing) while the latter group have left these errors behind and are struggling, not surprisingly, with features of advanced writing such as word order and cohesion. Both groups still have difficulties with lexis and the morphologically marked agreement.

Part of Speech (POS) usage

Our initial evaluation of POS usage suggests that the advanced students' active vocabulary is not necessarily much richer: not only are the POS usages similar, but also the number of different members belonging to the same category (i.e. types) is comparable. Advanced students do not seem to use a greater variety of verbs, nouns, or adjectives. We hypothesize that this is one of their error-avoidance strategies.

Patterns of Underuse and Overuse

We used the "missing" and "redundant" error categories to search for patterns of overuse and underuse. We found that the patterns of underuse and overuse are largely an L1 transfer phenomenon. The differences between the beginners and advanced students' writing are in the type of errors. The former make more grammatical mistakes, whereas the latter have more stylistic and lexical issues.

Ongoing work

Our intention is to test this tagset on our most elementary writing students' work and modify it further if necessary. We plan to add an additional layer of annotation – reconstruction – where the mistakes will be corrected. This will allow us to run the standard tools, such as a POS-tagger and a parser to be able to analyze data further and start the work on automatic Arabic tutors. The POS-tagged data is important for implementing a more reliable error analysis as well as for further parsing. The parse trees will give us data about the syntactic development of Arabic learners, an area that has not been investigated enough, and will shed light on the redundant/missing errors mentioned in the paper.

In the future, we also plan to compare native (NS) and non-native writings (NNS) to highlight the features of nativeness and non-nativeness of learner language.

With respect to overuse/underuse patterns, since we noticed so many L1 transfer-related phenomena, we plan to compare this present data with Hebrew speakers learning Arabic. Modern Hebrew is another Semitic language with properties similar to Arabic and we expect that the patterns of overuse and underuse will be different for these students.

References

Fitzpatrick, E., and Seegmiller, M. S. 1999. The Montclair Electronic Language Database Project. In Connor, U., and Upton, T., eds., *Applied Corpus Linguistics: A Multidimensional Perspective*.

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