ARIDA: An Arabic Inter-Language Database and Its Applications: A Pilot Study

Ghazi Abuhakema
The College of Charleston

Anna Feldman
Montclair State University

Eileen Fitzpatrick
Montclair State University

Abstract

This paper describes a pilot study in which we collected a small learner corpus of Arabic, developed a tagset for error annotation and performed simple Computer-aided Error Analysis (CEA) on the data. For this study, we adapted the French Interlanguage Database (FRIDA) (Granger, 2003a) tagset to the data. We chose FRIDA in order to keep our tagging in line with a known standard. The paper describes the need for learner corpora, the learner data we have collected, the tagset we have developed, its advantages and disadvantages, the preliminary CEA results, other potential applications of the error-annotated corpus of Arabic, and the error frequency distribution of both proficiency levels as well as our ongoing work.

Key words: learner corpus, inter-language, tagset, error annotation, computer-aided error analysis (CEA)

Language: Arabic

Language Learner Corpora and Applications

Learner corpora are electronic collections of authentic language data produced by learners of a foreign/second language. 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. It was not until the early 1990s that academics, English Foreign Language specialists and publishing houses alike began to recognize the theoretical and practical potential of computer learner corpora. Several projects were launched, including the International Corpus of Learner English (ICLE) (Granger 2003b), the Longman Learner Corpus (LLC), and the Hong Kong University of Science and Technology (HKUST) Learner Corpus (speakers of Cantonese learning English), (see Pravec 2002, Granger 2003a for more details).

Interlanguage and Contrastive Interlanguage Analysis (CIA)

Interlanguage is a term that was coined by Selinker (1972) to refer to an emerging linguistic system that is developed by a learner of a second language (L2) who has not become fully proficient yet but is only approximating the target language, preserving some features of the first language (L1) in spoken or written target-language productions as well as creating innovations. Interlanguage errors are caused by many phenomena, including L1 transfer, strategies of L2 learning and communication, and overgeneralization of the target language patterns.

Learner corpus research has concentrated on 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 difficulty). A wide range of topics has been covered, but vocabulary frequency (Altenberg 2002), modals (Neff et al. 2004), connectors, (Flowerdew 1998), collocations and prefabricated phrases (Nesselhauf 2003) have received more attention.

Computer-Aided Error Analysis (CEA)

Computer-aided Error Analysis (CEA) has led to a much more limited number of publications than CIA due to the cost of error annotation, which is manually done. Apart from articles describing error tagging systems, there are a few articles covering certain specific error categories (lexical errors (Chi et al., 1994; Kallkvist, 1995; Lenko-Szymanska 2003); tense errors (Granger, 1999; Fitzpatrick & 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 in the new culture, 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.

Error Tagging

There are two ways to annotate learner data for error. One approach is to reconstruct the correct form (Fitzpatrick & Seegmiller 2001). The other approach is to mark different types of errors with special tags (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. We have begun our study of learner Arabic using both reconstruction and special tag annotation, with the FRIDA tagset as our model for the latter approach. Here we report only on the annotation with special tags.

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 provided 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, syntax, agreement, lexis, 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 error categories. In addition, correct forms are also inserted in the text next to the erroneous forms.

Applications of Error Tagging

Learner corpora can serve as a teaching resource for Foreign/Second Language Teaching (FLT/SLT) and contribute empirical insights for Second Language Acquisition (SLA) research. Corpora annotated with linguistic information are particularly useful because they encode the distributional, morphological, and lexical aspects of inter-language and thus are essential for validating generalizations about language acquisition and supporting the development of new hypotheses and theories. Learner corpora play a crucial role in identifying areas of relevance for Foreign and Second Language practice. The use of corpora for obtaining examples does not depend on a specific method for evaluating the data --- both qualitative and quantitative analysis of data are possible (Díaz-Negrillo et al., 2009). According to Huston (2002), learner corpora make learners feel that their language is authentic and make it possible to prioritize the most frequently-used grammatical constructions and vocabulary items. Moreover, they help identify instances of overuse and underuse of spoken language and the extent to which NNS deviate from NS norms. Subsequently, the contrast helps in describing interlanguage, developing teaching materials, and training language instructors (Kennedy, 1998).

On the other hand, some researchers have argued that language corpora do not consider the inextricable link with the source language because of learner language phenomena such as overuse, underuse, and misuse. Such researchers thus regard such learner errors as inauthentic. (Tan, 2005) Tan, however, calls to revisit the notion of "authentic"., and contends that

much of the target language learners produce is "authentic" shaped by the cultural influences of the learners' local context.

Tagset Development

Developing a tagset to annotate learner errors is not trivial. As an essential prerequisite, we have to determine which learner language properties are useful or important to analyze in order to provide feedback and model second language acquisition. An intensive, interdisciplinary dialogue between the fields of Second Language Acquisition, Foreign Language Teaching and Natural Language Processing is needed to address this question.

A corpus annotated for error provides an invaluable resource for SLA research and practice despite the fact that error tagging is a highly time- and labor-consuming task. While traditionally the focus of research on learner corpora has been the identification and classification of learner errors, prominent strands of SLA research are concerned with the stages of the acquisition process (e.g., Pienemann, 1998), ``often independent of the accuracy of the execution of the patterns which are indicative of the different levels" (Díaz-Negrillo et al., 2009). In sum, SLA research observes correlations of linguistic properties (whether erroneous or not), and therefore, error annotation is an essential ingredient in the success of such research. 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 Arabic learner corpora available for public use[1]. In general, there is little research done in the area of data-driven instructional materials development for Arabic. Prior lack of interest in Arabic as a foreign language, the existence of more than thirty dialects and sub-dialects 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. 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 even 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.

[1] We have posed a request on Corpora List and other relevant listservs, but unfortunately, we did not receive suggestions on where to obtain Arabic **learner** corpora. Even though our pilot corpus is small, we have received a number of inquires about the corpus from Indiana University at Bloomington and the Center for Advanced Study of Language (CASL) at the University of Maryland.

Error Annotation of Arabic Linguistic Properties Relevant to Error Tagging

We have analyzed eight different texts written by four 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 language is English. They studied Arabic in an intensive program and then went to study abroad in Arab countries. Study abroad was a program requirement. Two students went to Egypt and the other two went to Syria. All spent one semester there. Some of the texts were written during their study years in the United States and others represented their productions while studying abroad. The texts were obtained upon their return. 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 students took language classes with one of the researchers before any data was collected. They volunteered to participate and there was no formal relationship between researchers and students that may have led to coercion. At the time, it was not possible to collect data from beginning students as these students wrote their assignments by hand and did not have the skill to type and save electronic versions of their writings. For this pilot study, the tagset was developed by one of the authors and applied by this author and a second annotator on different data in order to test the coverage of the tags. Both annotators are native speakers of Arabic and teach in higher education institutions.

The most salient difference between French and Arabic is in their lexical and inflectional morphological processes, 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 and

tense. 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 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 one or more 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 Arabic, many phonological distinctions that learners may not hear are reflected in orthography, e.g. long/short vowels and consonant doubling. So we added the long/short vowel distinction, emphatic/non-emphatic consonants, nunation (a mark of indefiniteness), hamza (a glottal stop), 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, number agreement, (in)definite agreement, gender agreement (Arabic utilizes different types of agreement to a great extent), and negation (Arabic has several negation particles which depend 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. We also 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, 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.

The Tagset for Learner Arabic

Table 1 (Appendix C) shows the Arabic tagset we are currently using. The first column shows the error domains while the second shows 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.

The domain is the most general level of tagging and includes nine categories: form/spelling, morphology, grammar, lexis, syntax, diglossia, punctuation and typos. Each domain is subdivided into a number of error categories. There are thirty six categories in this second annotation layer. The third annotation layer is based on the part of speech. Appendix A lists authentic examples of error domains and their categories. Appendix B provides excerpts from the annotated corpus.

The first domain has to do with errors related to spelling and/or form. These errors may be due to lack of distinction between long and short vowels, emphatic and non-emphatic sounds and familiarity with some spelling conventions such as the glottal stop and nunation which is a grammatical marker. Morphological errors are either derivational or inflectional. Learners may misuse or confuse affixes and may not inflect when inflection is fully or partially needed. In Arabic, they may also use incorrect infixes. The

grammar domain concerns aspects such as part of speech, number and gender agreement, tense, negation, etc. The lexis domain concerns complementation and most importantly meaning. Arabic has a rich lexicon and words have numerous meanings where the differences can be subtle. We added the diglossia domain to capture errors made by native speakers of any given dialect and those non -native learners who do not yet distinguish standard from dialectal. Punctuation errors include errors of misuse, deletion, and reversal of punctuation marks. For instance, the comma in Arabic has a reversed shape of direction compared to the comma in English (Abboud, 1983). The last domain refers to typos.

Parts of Speech

In creating the tagset and classifying words into their respective parts of speech (POS), we follow the classical view. Arabic is different from Indo-European languages and thus should have its own tagset. Arabists normally base their studies on the traditional grammar of Arabic and not on that of Indo-European languages. Traditionally, Arab grammarians divided words into three main parts of speech: nouns, verbs and particles. These are subdivided into smaller categories.

Nouns. These typically describe a person, an object or an idea. Nouns can be derived from verbs, nouns and particles. These nouns can be classified further by number, case and gender. In traditional European grammatical theory, this class will include particles, pronouns, relatives, demonstratives and interrogatives.

Verbs. These are similar to English but with different tenses and aspects. Verbs can be classified further into perfect, imperfect, and imperative. But they can also be subcategorized according to number, gender, and transitivity.

Particles. This class includes prepositions, adverbs, and conjunctions.

Table 2 (Appendix C) shows the grammatical categories that will be used for the third annotation layer. This third tier of annotation will make it possible to sort errors by grammatical categories. Figure 1 shows the distribution of parts of speech, by type, in our current data set.

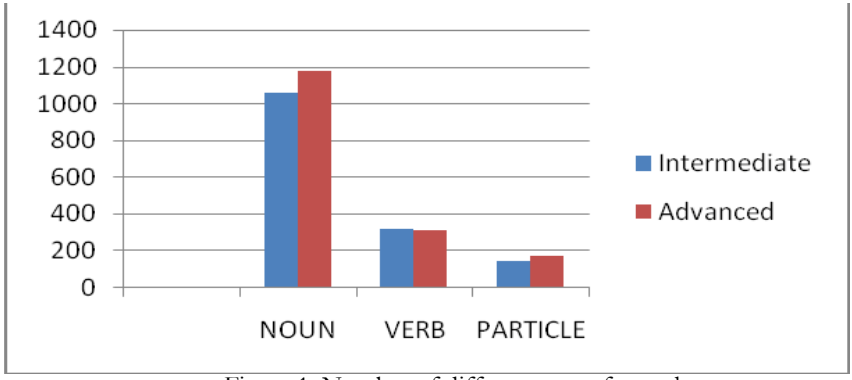

Figure 1. Number of different part of speech types.

The Tagging

While our corpus was not large enough to test inter-rater reliability, our test of the tagset's 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. In addition, the annotators found the tag phrases hard to remember so the phrases have been changed to keywords, as Table 2 shows. In addition, we looked at specific error categories covering a range of error types in our corpus. The next section describes our analysis in more detail.

Results

We have collected and annotated for error 3,818 tokens of intermediate learner texts and 4,741 tokens of advanced learner texts. The total number of errors in the former amounts to 259, whereas the total number of errors in the latter is 488.

Frequency of error types

The frequency of all error types taken together based on student level already provides useful data for pedagogical purposes. Figure 2 shows the most frequent errors by learner level. The difference in error types by learner level is statistically significant at the > .01 level.

One notable difference between the intermediate and advanced writers is that the former are still struggling with phonological/orthographic issues – such as the glottal stop known as hamza, which these students do not hear or confuse with other pharyngeal sounds resulting in misspelling. These students are also struggling with the many rules on where to write the hamza based on its location in the word and the surrounding vowels. The advanced group has left these errors behind and is 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. It is perhaps interesting to notice that the intermediate texts contain only an insignificant number of redundancy errors, i.e. the intermediate learners experiment with Arabic syntactic constructions much less than their advanced counterparts. We think that the redundancy errors are due to a misuse of a variety of syntactic constructions. We plan to parse the learner data to verify our hypothesis and

study which syntactic constructions cause most of the problems. Another noticeable error is the use of the conjunction و and in both intermediate and advanced texts as a separate orthographic word (instead of the usual attached form). While this conjunction stands by itself in English, it has to be connected to the following word in Arabic. Students at both levels are still struggling with this orthographic convention.

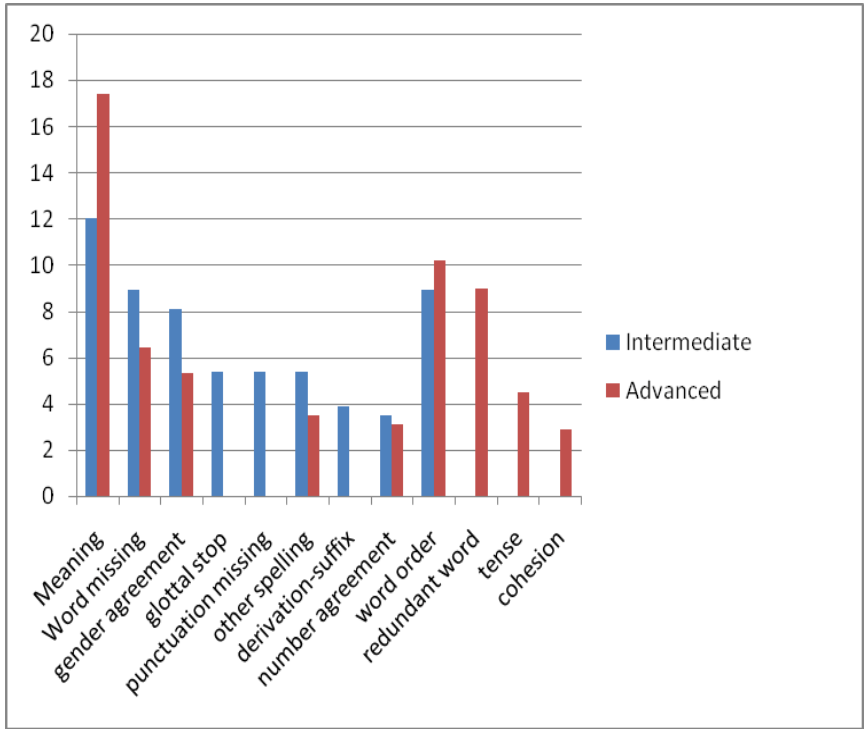

Figure 2: Frequency Distribution of Errors in Intermediate and Advanced Learner Texts.

Patterns of Overuse

Intermediate-level students tend to overuse the auxiliary be in Arabic texts, particularly in the past tense. This, again, seems to be a L1-transfer related phenomenon, since, unlike Arabic, English clauses always contain a verb. Another instance of overuse and L1 effect on the acquisition of Arabic comes from the use of prepositions in intermediate texts. Many English structures that require the use of a preposition have a preposition-less counterpart in Arabic. For example, the English *one of those* is translated into Arabic as أحد هؤلاء *one those* (taken from our learner corpus). Advanced students tend to overuse the particle قد which is used, among other things, for emphasizing the fact that the action has happened. The overuse of this particle leads to a heavy style. Advanced students also tend to use the relative pronoun after indefinite nouns, another overuse pattern that appears to stem from L1 interference. For instance, an incident which made translates into Arabic as جعل حادث an incident made without using the relative pronoun الذي *that* . Errors of redundancy often involve using two words that are semantically related. For instance, the English *the pace of expansion was faster* will translate into Arabic as الاتساع كان أسرع *the expansion was faster*. In Arabic, *pace* and *faster* are two related words that derive from the same root.

Part of Speech (POS) Usage

We also looked at the POS usage of the two types of students, intermediate and advanced. We did not notice striking differences in the distribution of POS in their writings and, to our surprise, our initial evaluation 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. According to our calculated value of

χ-square (9.7, d.f. = 7), the breakdowns of the POS types in advanced and intermediate Arabic learner texts are different only at the >0.2 level of significance. Advanced students do not seem to use a greater variety of verbs, nouns, or adjectives. In the case of verbs, the number of different verbs is even lower for the advanced students. We hypothesize that this is one of their error avoidance strategies. They thus tend to use what they have already learned well. Further studies on a larger learner corpus should test our hypothesis.

Patterns of Underuse

We also looked at particular instances where intermediate or advanced learners underused a specific lexical or grammatical category. We used the "missing" and "redundant" error categories to search for patterns of underuse and overuse respectively. We noticed that intermediate-level students tend to omit the particle أنّ that used to introduce a nominal clause in a complex sentence. For instance, in the clause ....تشير الأحصائيات أن *statistics point to* …, the particle was missing. This is clearly an L1-transfer phenomenon, since English (L1) does not have an analogous syntactic structure. Another instance of underuse in intermediate texts is missing "filling" words that are implied by the context but are still required to be present as part of the structure, e.g. هناكَ there is which is a predicate in a nominal clause. The absence of this particular form is odd since it is a direct equivalent of the English. Most of the missing prepositions are due to the direct translation of the English verbs into Arabic. English and Arabic verb translations do not follow the same sub-categorization patterns, but students tend to translate the English patterns into Arabic. Often the individual English verbs should be translated as phrasal verbs into Arabic, e.g. the English verb attack translates into Arabic as هجم على, where the indirect object follows the preposition

على on. A common error in the data has a direct object immediately following
على.

Like the intermediate students, advanced students' instances of underuse seem to be due to L1 interference. Advanced students tend to omit the particle أن (obligatory in the past tense) after the two adverbs: قبل *before* and بعد *after* when followed by verbs. The English translation of the Arabic قبل أن أصبح is before he became, where no particle is required. Additionally, advanced students still have some difficulty with deriving adverbs from adjectives using prepositions. They prefer to use the accusative case ending to create adverbs from adjectives, which is another strategy in Arabic. This derivation method does not, however, apply to all adjectives. For example, the adjective عنيفا violent cannot function as an adverb in the accusative case.

Ongoing work

Our intention is to test this tagset on our most elementary writing students' work and modify it 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 tools to aid students in their acquisition of Arabic. Once we have a larger learner corpus, 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 (NNS) writings to highlight the features of nativeness and non-nativeness on the 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.

Conclusion

The study is a pilot study and provides a preliminary account of the types of errors intermediate and advanced students of Arabic may make. A larger corpus in terms of the number of tokens and other proficiency levels is needed to validate these findings. Researchers of other Less Commonly Taught Languages who may be interested in creating a tagset for their own languages to annotate learner corpora need to bear in mind the peculiarities of their languages. Adopting a well known tagset for this study was the guiding principle. However, when finalizing the tagset, the investigators needed to account for the differences between the Romance languages and Arabic, which belongs to a different language family and has a different writing system.

References

Abboud, P. & McCarus, E. (1983). Elementary Modern Standard Arabic. Cambridge, England: Cambridge University Press.

Altenberg, B. 2002. Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-

Tyson. T. (Eds.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (pp. 37-54). Amsterdam and Philadelphia: John Benjamins.

Chi, A., Wong, K., & Wong, M.. (1994). Collocational problems Amongst ESL l Learners: a corpus-based study. In L. Flowerdew & A. Tong (Eds.), Entering Text, (pp. 157-165), Hong Kong: Language Centre, The Hong Kong University of Science and Technology.

Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch. H. (2009). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. http://purl.org/dm/papers/diaznegrillo-et-al-09.html.

Fitzpatrick, E. & Seegmiller, M.S. (2001). The Montclair electronic language learner database. In G. Antoniou & D. Deremer (Eds.), Computing and Information Technologies: Exploring Emerging Technologies. New Jersey: World Scientific.

Fitzpatrick, E. & Seegmiller, M. S. (2004). The Montclair Electronic

Language Database Project. In U. Connor, & T. Upton (Eds.), Applied Corpus Linguistics: A Multidimensional Perspective. (pp. 223-237). Amsterdam: Rodopi.

Flowerdew, L. (1998). Application of learner corpus based findings and methods to pedagogy. In S. Granger & J. Hung (Eds.), Proceedings of the First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, (pp. 38–44), Hong Kong: The Chinese University of Hong Kong.

Granger, S. (2003a). Error-tagged learner corpora and CALL: A promising synergy.

CALICO Journal, 20, 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, 538–546.

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge, England: Cambridge University Press.

Kennedy, G. (1998). An Introduction to Corpus Linguistics. London, England: Longman Group UK Limited.

Lenko-Szymanska, A. (2003). Lexical problem areas in the advanced learner corpus of written data. In B. LewandowskaTomaszeczyk (Ed.), Practical Applications in Language Corpora (PALC 2001), Papers from the International Conference at the University of Lodz. (pp. 505-520}, Frankfurt am Main: Peter Lang.

Neff, J., Ballesteros, F., Dafouz, E., Martinez, F., Rica,J. P., Diez, M., and Prieto, P. (2004). Formulating writer stance: A contrastive study of EFL learner corpora. In U. Connor, & T. Upton, (Eds.), Applied Corpus Linguistics: A Multidimensional Perspective. (pp. 73-89). Amsterdam: Rodopi.

Neff, J., Ballesteros, F., Dafouz, E., Martinez, F., Rica, J. P., Diez, M., and Prieto, P. (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. (pp. 203-226). Amsterdam: Rodopi.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. Applied Linguistics, 24, 223–242.

Pienemann, M. (1998). Language Processing and Second Language Development: Processability Theory. Amsterdam and Philadelphia: John Benjamins.

Pravec, N. (2002). Survey of learner corpora. ICAME Journal 26, 81– 114.

Richard, J. C. (1992). Dictionary of Language Teaching and Applied Linguistics. Second edition. Essex, England: Longman Group UK limited.

Selinker, L. (1972). Inter-language. International Review of Applied Linguistics, 10, 209-231.

Tan, M. (2005). Authentic language or language errors? Lessons from a learner corpus. ELT Journal, 59(2): 126-134.

Appendix A: Error Domains and Categories: Authentic Examples

The following are sample authentic errors. Both tags and corrections are provided.

<
ش
>

>همز< أصبحت الأحوالَ في بيت مالكوم أسوئ (أسوأ) من قبل

>نون< هل هو إنسان أو جزء من الله أو الاثنان أو كان إنسان (إنسانا) أولا

>خهج< بالإضافة إلى الجيزيا (الجزية)

>علة< لأن وحدة (واحدة) من زوجات النبي محمد كانت مصرية

>ص<

>خرف< فرعوني يعني أجدادهم هم المصريون القديمون (القدماء) ولكن يقول معظم العرب الأمريكيين إنهم لهم (ان لهم) أصدقاء<شقح> تظاهر الناس بهم (بأنهم)<شقح>

<

ق
<
>نوع <كانت مصر و الشام متحدين ً(اتحاد)
>طقت< في الشرق الأوسط و الهند و إفريقيا الشمالي
(الشمالية) لذلك حرب (حارب) أذينة قائد
الانقلاب<نوع>
>طقج< دخلت (دخل)الجيش الروماني عن طريق الإسكندرية

<
ك
<
>عني< حصر الأقباط في ملابسهم وتصرفاتهم حصرا كثيرا

<
ن
<
لان قي الماضي الاحصاءات الأمريكية (لأن الاحصائيات الأمريكية
في <رتب> الماضي
كانت العاصمة المدينة الوحيدة (التي) مازالت في يد
الإمبراطورة<كفق> >ز<
>عمم< هذه كانت بداية الحروب ألصليبية>همز< بين المسيحيين و
المسلمين التي استمرت لميتين (لمائتي) سنةٍ

<
س
<

كان بداية فكرة "الوطن" فيه كل الناس في نفس طبقة أجنبية (بداية <>غمض<>ركك فكرة الوطن كانت أن كل الناس ينتمون الى طبقة أجنبية واحدة)

<
ق
>

>طفق< من اعمارهم)،(,

<
خ
>

صاحبا الفضل الأولان في اهتمامي هما جدي و جدتي الفلسطينان(الفلسطينيان)<طبع

Appendix B: A Sample

Error-tagged Text

The following text is annotated using the first two layers of annotation: error domains and error categories. The first tag represents the domain and the second the category. In the sample, ten errors have been annotated – three grammatical errors and one error in each of the other domains except for diglossia .

بدأت قصة مالكوم اكس منذ سنوات قبل ولادته بعد تأسيس منظمة جديدة في الولايات المتحدة باسم "امّة الاسلام" على يد رجل باسم "فارد محمّد". لا نعرف كثيراً عن حياة فارد، لكن مما نعرف عنه، انه كان رجلاً ممتعاً>ك<>عني<. في الحقيقة كان فارد من نيوزيلندا و ابيض البشرة، لكن عندما بدأت المنظمة، ادّعى انه كان من اصل سعوديّ، و انهّ كان من الاولاد >ق<>طقت< الملك. قبل>ن<>كفق< بدأت المنظمة، كان مملك>ق<>نوع< مطعماً، و يوماً ماُسُجِنَ من قبِلِ الشرطة لجريمة بيع المخدّرات في مطعمه. تخرّج >ص<>خلط< من السجن بعد ثلاث سناواتاً >ش<>خهج<، و هذه

المرة اصبح بائعاً. لكن في نفس الوقت، بدأ فارد يعلمّ "الاسلام" للسود في مدينة
"ديترويت" في ولاية "ميشيغن". في البداية كان>ق<>طقج< عنده طريقة خاصّة و سرّية لتِعليم دينه الجديد: كان يذهب من بيت الى بيت محاولاً أن يبيع ملابس، و كان يبدأ يتكلم عن الدين و افكاره، و عن الاحوال السيئة السود >ن<>رتب<في امريكة، و الحل الذي كان يقدّمه. ازدادت شهرته ببطء فبدأ يعقد اجتمعات >خ<>طبع< في بيوت خاصة، و بعد وقت قليل اسُتؤجِرت صالات عامة لخطباته >س<>ركك.<

Appendix C: Error Domains, Error and Grammatical Categories

Table 1. The Error Tagset for Arabic

Error Domains مجالات الأخطاء	Error Categories فئات الأخطاء
Form/spelling الشكل>ش<	Agglutination >شبك< التشبيك
	Vowel length confusion الخلط بين حروف العلة الطويلة والقصيرة >علة<
	Emphatic/non emphatic consonants الحروف المفخمة والمرققة> حخر<
	Consonant doubling (shaddat) >شدد< الشدة
	Nunation >نون< التنوين
	Glottal stop >همز< الهمزة
	Other spelling errors >خهج< أخطاء هجائية أخرى

Morphology < الصرف >ص<	Derivation-prefixation < الاشتقاق- البادئة >شقب<
	Derivation-suffixation

	الاشتقاق – اللاحقة >شقح<
	Derivation-infixation < الاشتقاق المتوسطة >شقم<
	Inflection – full <صرف> المنصرف
	Inflection – partial غير المنصرف (الممنوع من الصرف) >منع<
	Inflection – zero <بني> المبني
	Inflection confusion < الخلط في التصريف >خرف<
Grammar > ق< القواعد	Class (POS) <نوع> نوع الكلمة
	Gender agreement المطابقة في الجنس >طقج <
	Definite/Indefinite agreement < المطابقة في التعريف >طقت
	Number agreement <طقع > المطابقة في العدد
	Tense <صيغ> الصيغة
	Voice <معج> المبني للمعلوم والمجهول
	Negation <نفي> النفي
Lexis < ك< المفردات	Meaning > عني< المعنى
	Adjective Complementation < متممة الصفة >صتم<
	Noun complementation <ستم> متممة الاسم

	Verb complementation <فتم> الفعل متممة
Syntax <ن> النحو,	Word order >رتب< ترتيب الكلمات
	Word missing <كفق> كلمة مفقودة
	Word redundant <كزد> كلمة زائدة
	Cohesion <ربط> الترابط
Diglossia <ز> ازدواجية اللغة	Colloquial use <عمم> استخدام العامية
Style <س> الأسلوب	Unclear <عمض> غامض
	Simplistic <ركك> ركيك
Punctuation <ق> علامات الترقيم	Punctuation confusion <طخل> الخلط في الترقيم
	Punctuation missing <طفق> علامة ترقيم مفقودة
	Punctuation redundant <طزد> علامة ترقيم زائدة
Typos <خ> أخطاء مطبعية	>طبع<

Table 2: Grammatical Categories

Grammatical Categories الفئات القواعدية	Tag/Code الرمز
Noun الاسم > س<	Proper <علم>

	With alif maqsoura<قصر>
	Ending with a yaa <نقص>
	Diptot <منع>
	Nominative <سرف>
	Accusative <سنص>
	Genitive <سجر>
	Plural <جمع>
	Sg.-Pl. confusion <خجم>
	Simple adjective <بسط>
	Adjectival clause <صجم>
	Expanded adjective <صفض>
	Prepositional phrase in a adj. Position <صشب>
	Comparative <قرن>
	Comparative with tamyeez <مي>
	Free Pronoun– Demonstrative <ضشر>
	Free pronoun– personal<ضشر>
	Free pronoun– Relative <ضفص>
	Bound pronoun– Possessive <ضوص>
	Bound pronoun– Object<ضمل>
	Preposition confusion <جلط>
	Preposition missing <جفق>
	Preposition redundant <جزد>

	Adverbs <ظرف>
	Conjunctions <ربط>
Verb <ف الف عل>	Defective/Non defective <صع>
	Transitive/Intransitive <زعد>
	Nominative <فرف>
	Subjunctive <فنص>
	Jussive – Imperfect <فجز>
	Jussive – Imperative <جزء>
Punctuation <ق علامات الترقيم>	Stops <نقط>
	Commas <فصل>
	Colon <نقن>
	Semicolon <فصن>
	Exclamation marks <عجب>
	Question marks <سأل>
	Quotation marks <نصص>