Workshop on Arabic Corpus Linguistics

11th and 12th April 2011

Lancaster University, UK

We gratefully acknowledge the generous sponsorship of this event by
Research in Motion (RIM: www.rim.com).

Tunisian Arabic Corpus: Creating a written corpus of an "unwritten" language

Karen McNeil[*] and Miled Faiza[†]

[*] Georgetown University

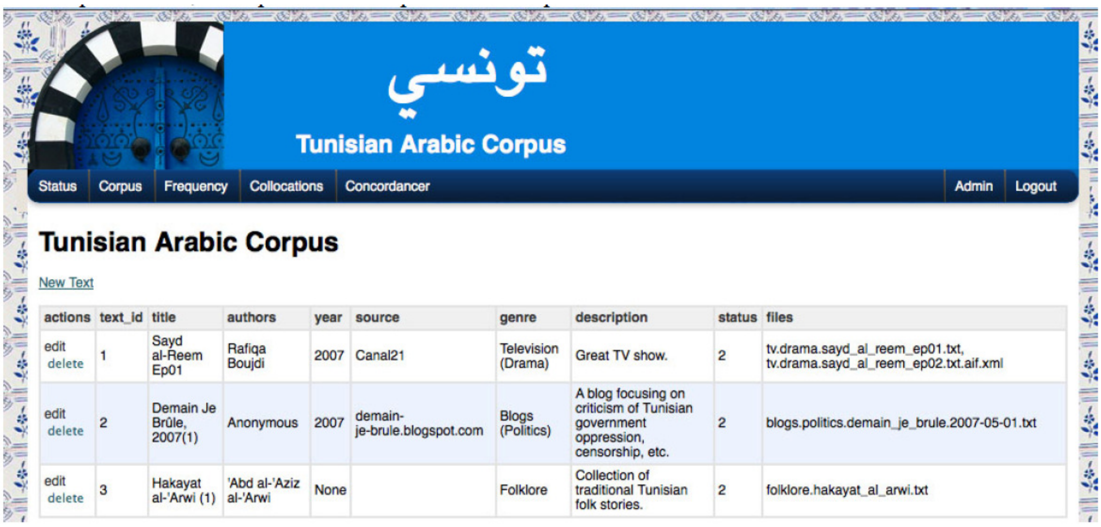
[†] University of Virginia

Arabic corpora have, to date, largely focused on Modern Standard or Classical Arabic, neglecting the spoken varieties which are the everyday form of communication throughout the Arab world. This is unfortunate, since many of the tasks for which corpora would be used (especially Natural Language Processing tasks like speechto- text systems) require corpora of the spoken language, not the written one. The traditional focus on the written variety, and stigma of the spoken "dialects," continues to be a barrier in Arabic linguistic research. Even when the importance of the spoken varieties is recognized, however, many challenges remain in creating a usable corpus of spoken Arabic. Chief among these is a difficulty in acquiring written sources: sources which provide a large part of the corpora in other languages (such as literature and news), are almost all written in Standard Arabic throughout the Arab world.

In this paper I will discuss my own efforts to overcome these challenges and create a corpus of Tunisian Spoken Arabic, with the goal of using the corpus to create a bilingual Tunisian-English Dictionary. This project is currently in progress, with the corpus containing approximately a quarter of a million words, and a final goal of one million words. The topics address specifically are:

- Utilization of traditional sources: Sources which are traditionally written in dialect and so are readily available for inclusion in the corpus, including plays, television/movie scripts, and folktales.
- Utilization of new media: New media has both made written materials easier to access, and expanded the domains in which it is considered acceptable to write in dialect. Some of these sources include blogs, emails, and Facebook postings.
- Transcription of spoken materials: Since Tunisian Arabic is, mainly, a spoken language, obtaining a complete corpus would be impossible without a significant inclusion of spoken materials. I will discuss the sources which I utilized, and some of the challenges presented by the transcription of these sources.
- Balance: Many of the genres traditionally included in corpora (like news) simply do not exist in Tunisian (even broadcast news is delivered in MSA), so creating a corpus which is "balanced" in the traditional sense is not possible. I will discuss the criteria by which I designed the corpus to be as balanced as possible for the language.
- Work-flow management: To organize the corpus materials and work-flow, I created a web application using the programming language Python and the web framework Turbogears. Although building this application required a significant amount of time, it was well worth it in that it allows me to manage

and organize the corpus files and metadata, and perform basic linguistic processing (such as frequency lists, collocations, and concordancing). In addition, the application acts as a central portal for all the people working on the project (including transcribers working from Tunisia), allowing them to download yet-to-be-completed files, and upload the completed transcripts.



Tunisian Arabic Corpus

New Text

actions	text_id	title	authors	year	source	genre	description	status	files
edit delete	1	Sayd al-Reem Ep01	Rafiqa Boujdi	2007	Canal21	Television (Drama)	Great TV show.	2	tv.drama.sayd_al_reem_ep01.txt, tv.drama.sayd_al_reem_ep02.txt.aif.xml
edit delete	2	Demain Je Brûle, 2007(1)	Anonymous	2007	demain-je-brule.blogspot.com	Blogs (Politics)	A blog focusing on criticism of Tunisian government oppression, censorship, etc.	2	blogs.politics.demain_je_brule.2007-05-01.txt
edit delete	3	Hakayat al-'Arwi (1)	'Abd al-'Aziz al-'Arwi	None		Folklore	Collection of traditional Tunisian folk stories.	2	folklore.hakayat_al_arwi.txt

Getting flexible: Developing a corpus of Iraqi Arabic to study multimodal communication

Kamala Russell[1], Atoor Lawandow[1], Amy Dix[1], Edward King[2], Frederica Lipmann[1], Danial Parvaz[3], Gina-Anne Levow[4], Dan Loehr[3]

[1]University of Chicago
[2]Stanford University
[3]MITRE, Corporation
[4]University of Washington

Our corpus of Iraqi Arabic (IA) natural speech data comprises *Praat* TextGrid (Boersma 2001) transcriptions of 36 interactive storytelling elicitations, two to ten minutes in duration, each. Each TextGrid comprises six parallel tiers. The multimodal nature of our data and objects of analysis made it necessary to have tiers encoding data from the phrase level down to the phone level. The first tier is a transcription of the speech segmented at the phrase level, in an adapted Arabic orthographic script. In addition to the Arabic words, this tier includes filled pauses, breaths, and non-speech sounds. The second tier is a phrase-by-phrase transliterated version of the first, using the *Hans Wehr* transliteration system (see Cowan 1974). This transliteration is generated automatically using software, developed in house, that transcodes Arabic into Latin character text. The third tier consists of English translation equivalents of each phrase. The fourth tier is a word-by-word segmentation of the *Hans Wehr* transliterated version, generated automatically using an automated speech alignment software, *Sonic* (Pellom 2001), modified to work with Arabic audio and text. The fifth tier is a word-by-word English gloss of the Arabic. The sixth tier is a phone-level segmentation represented in Hans Wehr transliteration, generated automatically from the word-by-word parse, again using *Sonic*.

This corpus was developed by non-professional transcribers in support of a larger research project concerned with multimodal communication. They were native speakers of Iraqi Arabic and university students with no prior training in linguistics, use of the International Phonetic Alphabet, or transcription methods. Using the software *Arabic Editor*, they transcribed the natural language data in a fully vowellized Arabic script, augmented with two characters from Persian orthography (gaf and che) to represents sounds common to IA but non-existent in Modern Standard Arabic (MSA). Typically, MSA is written without representing most vowel sounds. Including vowelling was necessary to support further automated processing, particularly transliteration into Latin character text and text-speech alignment. Regarding choice of *Hans Wehr*, strengths of this transliteration system include that it is widely known, readable, represents a one-to-one mapping of characters to phonemes, and is easily adapted to *Praat* and *Sonic*.

The English translation equivalents and word glosses were also created by our native Iraqi Arabic speakers, each of whom is proficient in American English as a second language. The translation attempts to accommodate the requirements of the LDC (Linguistics Data Consortium), non-Arabic speaking gesture researchers, and non-professional transcribers. They are a combination of lexical and idiomatic equivalents. Native speakers of English checked the translations. In word glossing, the goal was consistency and clarity of reference, without following the morphological notation conventions typical in transcriptions created for linguistics research.

References

Arabic Editor, Text editor for Microsoft Windows. Basis Technology Corporation.

Boersma, P. (2001). *Praat, a system for doing phonetics by computer*. Glot International, 5(9-10):341-345.

Cowan, Milton J. [Editor] (1974). Hans Wehr, A Dictionary of Modern Written Arabic. London: MacDonald & Evans, Ltd..

Pellom, B.L. (2001). *Sonic: The University of Colorado Continuous Speech Recognizer*. Technical Report: TR-CSLR-01, Center for Speech & Language Research. University of Colorado at Boulder.

Building Arabic corpora to measure online Arabic content

Anas Tawileh and Mansour Al Ghamedi

This paper presents the development and analysis of two Arabic language corpora for the purpose of estimating the size of Arabic online content. So far, little effort has been invested to produce an objective and dependable estimation of the size of the Arabic indexed web. A recent project undertaken by King Abdulaziz City of Science and Technology for the development of an indicator of the Arabic online content was designed based on corpus linguistics. As part of this project, two Arabic language corpora were constructed to establish the foundation for the calculation of the indicator. The first corpus was built based on the articles in the Arabic Wikipedia (over 95,000 articles in total), and the second was constructed by crawling more than 75,000 pages from the Arabic web extracted from the Open Directory Project. The development of the corpora entailed extracting and removing markup tags and directives, converting the encoding of the collected text into Unicode and storing the text in the corpus.

The corpora were then analyzed to compile a list of words in each corpus, and for each word in these lists calculate the word and document frequency in the corpora. Based on these calculations a word list that contains 25 words was extracted based on Zipf's distribution to form the basis of the indicator's estimation. The indicator will be estimated by sending each of these words to relevant search engines, and calculate the projected size of the Arabic indexed web accordingly. An estimation of the overlap between the search engines involved was performed to enable reliable calculations of the indicator.

In order to facilitate further utilization and maximize their value, these corpora were released under an open license that promotes reuse and adaptation. This paper will elaborate on the corpora development process, discuss the theoretical foundation for the project, offer insights into the results and outline future progress.

Aspects of the lexical and grammatical behaviour of Arabic idioms

Ashraf Abdou
Cairo University

This paper reports a corpus-based study of some aspects of the lexical and grammatical behaviour of Arabic idioms. The term *idiom* is understood here as: a multiword unit that has a syntactic function within the clause and has a figurative meaning in terms of the whole or a unitary meaning that cannot be derived from the meanings of its individual components.

Six hundred and fifty four idioms in Modern Standard Arabic have been gathered from dictionaries and examples observed in everyday readings and interactions. A representative sample of 70 idioms has been randomly selected. The corpus data of these idioms have been obtained mainly from the *All Newspapers* section of *Arabicorpus*, a corpus of Arabic developed by Dilworth Parkinson. This section contains texts from five major Arabic newspapers, with a total word count of more than 83 millions.

The study addresses the following points in the lexical and grammatical behaviour of idioms (1) lexical variation, (2) perspective-adaptation (which refers to cases where an idiom may have two or more variants that differ in terms of e.g. transitivity and intransitivity, causativity, and reflexivity), (3) changes in the lexicogrammatical complexity of idioms, (4) inflectability, and (5) the use of active and passive voice.

In general, it has become evident that Arabic idioms show a degree of formal flexibility that is higher than what is suggested for them in some recent accounts of the phenomenon, e.g. Attia (2006). Both the transparency of the figurative images underlying idioms and the isomorphism of many of these expressions have been vital in accounting for this variability.

On the other hand, several explanations have been proposed for the restrictions on the formal variation of idioms. These include the discursive function of the idiom, some sociolinguistic factors related to the present state of diglossia in Arabic, the incompatibility between the grammatical meaning of the type of formal variation and the idiomatic meaning, and the possible interference of stylistic factors.

Despite a growing need for corpus-based research on Arabic phraseology to meet both practical and theoretical ends, and even with the availability of several suitable Arabic corpora, there is still a scarcity of this type of research. This work takes a step to fill in this gap with respect to the lexical and grammatical properties of Arabic idioms.

References

Attia, M. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In T. Salakowski, F. Ginter, S. Pyysalo, & T. Pahikkala (Eds.), *Advances in natural language processing. Lecture Notes in Computer Science*, 4139, pp. 87-98. Berlin: Springer.

Compiling a modern corpus-based collocation dictionary of Arabic

Sattar Izwaini
Department of Arabic & Translation Studies, American University of Sharjah, United Arab Emirates

Traditional Arabic collocation lexicons such as *fiqh al-lughah* فقه اللغة and *al-mukhassas* المخصص are about one thousand years old and are full of obsolete usages. This also applies to general classical Arabic dictionaries that also include such collocations within their entries. There is therefore a need to compile a modern dictionary that excludes such usages. In addition to the collocations that have fallen out of use, a large number of new collocations have emerged with words assuming new combinations, which creates the need for the newly created collocations to be documented and included in an updated collocation dictionary. For example, the collocation denoting a group of airplanes (*sirb taa'iraat* سرب طائرات) incorporates a word that is originally used for a group of birds (*sirb hamaam* سرب حمام). Another example is the noise of a tank (*hadeer al-dabbabah* هدير الدبابة) which has borrowed the sound of sea waves (*hadeer al-bahr* هدير البحر). Another kind of new collocations in Arabic is the calque translations of English expressions such as *green light*, *bottle neck* and *money laundering*. Naturally, such Arabic collocations cannot be found in collocation or general dictionaries whether old or modern.

It is therefore clear why users need an updated lexicon of collocations that can be referred to while writing, translating or carrying out research. This paper reports on a project that is still in the making, namely compiling a modern collocation dictionary of Arabic. It presents the broad lines of data and methodology used in this project. A corpus of old collocation and general dictionaries in electronic form has been created. This corpus comprises of seven dictionaries (three of collocation and four for general use). More dictionaries and texts will be added in due course when they become available in electronic form. The complied data also incorporates two corpora of modern Arabic newspapers with diverse topics and areas (politics, health, religion, sports, finance, education, science & technology, and art & music) of about 3 million words. Lexical combinations are identified and extracted in this corpus as well as in other corpora that are available online, for example ArabiCorpus (see http://arabicorpus.byu.edu/). Word combinations are recorded as candidate collocations while archaic combinations are removed. The entries are mainly nouns, verbs, and adjectives. For example, nouns are cited along with their noun, verb and adjective collocates. It is also envisioned that collocates are listed in a separate section where cross-references are made to their nodes to facilitate easy use.

Collocational patterns in a corpus of Modern Standard Arabic

Safwat Ali Saleh
Department of Linguistics and English Language, Lancaster University

Compared with English, relatively little corpus-based work has been done on Arabic in general, and on collocation in particular. Most previous studies of collocation in Modern Standard Arabic MSA have neither relied on corpus data, nor employed statistical measures to identify collocations. In fact, in most studies to date, *collocation* is not rigorously defined; nor has a precise classification of grammatical patterns of *colligation* been proposed, or even a *semantic* or *pragmatic* analysis of node-collocate relationships.

In this talk, I will take up a corpus-based approach to give a more detailed analysis of the lemma *HARB* 'WAR' as it co-occurs in Al-ahram newspaper corpus which consists of 91 million tagged words. The primary aim is to explore the linguistic structure and semantic properties of collocations in MSA through answering the following research questions: a) how can we identify collocates around the node *HARB* in the corpus? b) What are the grammatical patterns realised between the node and its collocates? c) What are the semantic preferences and discourse prosody associated with *HARB* in the corpus? The analysis is conducted within the framework of Sinclair's model of the *Extended Lexical Unit* (Sinclair 1998, 2004), according to which an extended lexical unit consists of *lexical*, *syntactic*, *semantic* and *pragmatic* components (Stubbs, 2007: 179; 2009: 23). Accordingly, to define a linguistic unit, we have to specify its possible constituents (to define its semantic content), and the possible relations between these constituents (to define its structure) (Stubbs, 2001: 87).

A central finding is that empirical quantitative evidence can be given interpretation of the phrasal units of meaning at lexical, syntactic, semantic and pragmatic levels. Hence, the meaning(s) of a given word could be defined based on its preferred sequences with which it associates within its phrasal co-text. Such meanings, in turn, are closely related to the structures in which the word occurs.

References

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive Lexical Semantics* (pp. 1–24). Amsterdam: Benjamins. (Reprinted in Sinclair 2004: 131-48).

Sinclair, J. (2004). *Trust the Text: Language, corpus and discourse*. London: Routledge.

Stubbs, M. (2001). *Words and phrases : corpus studies of lexical semantics*. Oxford ; Malden, MA: Blackwell Publishers.

Stubbs, M. (2007). Quantitative data on multi-word sequences in English: The case of the word 'world'. In M. Hoey, M. Malhberg, M. Stubbs & W. Teubert (Eds.), *Text , Discourse and Corpora: Theory and Analysis* (pp. 163-189). London: Continuum.

Stubbs, M. (2009). Technology and phraseology with notes on the history of corpus linguistics. In U. Römer & R. Schulze (Eds.), *Exploring the Lexis–Grammar Interface* (pp. 15 - 31). Amsterdam; Philadelphia: John Benjamins Publishing Company.

Corpus analysis of conjunctions: Arabic learners' difficulties with collocations

Haslina Hassan and Nuraihan Mat Daud
Kulliyyah of Islamic Revealed Knowledge & Human Sciences,
International Islamic University Malaysia

This paper investigated Arabic majors use of conjunctions in an easy produced for Computer Applications for Language Studies course offered by the Department of Arabic Language and Literature of the International Islamic University Malaysia (IIUM). The essays were submitted through the university's Learning Management System (LMS). It serves as a corpus for this study. SketchEngine was used to track the frequency of conjunctions used by the students. However, SketchEngine was limited in its applications in that it is unable to detect the collocations for a number of conjunctions. Hence, Excel had to be used with Sketch Engine to get a more reliable data for this study. Findings revealed that there are a number of favourite conjunctions among the learners. The main problem with its usage lies in the use of collocations. The analysis revealed that out of more than 75 conjunctions available only five were commonly used by the students. Their usage, however, were not necessarily correct. There seemed to be a confusion in the application of these conjunctions, particularly those which carry similar meanings. The contexts of their applications are different. Mother tongue interference could be the reason for the confusion since the direct translation of the word can be used in the same context. This study highlighted the need to focus on these errors when teaching Arabic to second or foreign language learners.

The Leeds Arabic Discourse Treebank: Guidelines for annotating discourse connectives and relations

Amal Al-Saif and Katja Markert
School of Computing, University of Leeds

Discourse relations such as CAUSAL or CONTRAST relations between textual units play an important role in producing a coherent discourse. They are widely studied in theoretical linguistics (Halliday and Hasan, 1976; Hobbs, 1985), where also different relation taxonomies have been derived (Hobbs, 1985; Knott and Sanders, 1998; Mann and Thompson, 1988; Marcu,2000). Discourse relations can be signalled by explicit lexical indicators, so-called *discourse connectives* (Marcu, 2000; Webber et al., 1999; Prasad et al..2008a). In Example 1, the connective (لأن/because) indicates a CAUSAL relation which relates the arguments :(أحمد لم يذهب الى المدرسة/Ahmad did not go to the school) and (هو مريض/ he was ill). In addition, there is another connective in the example (لكن/ however) indicating CONSTAST relation and relates different arguments :(أحمد لم يذهب الى المدرسة/Ahmad did not go to the school) and (هو ذهب الى الطبيب/ he went to the doctor).

Ex1:

أحمد لم يذهب الى المدرسة لأنه مريض لكنه ذهب الى الطبيب

Ahmad did not go to the school because he was ill. However, he went to the doctor.

Discourse connectives are often used as an important feature in the automatic recognition of discourse relations, a task useful for many applications such as automatic summarization, question answering and text generation (Hovy, 1993; Marcu, 2000). Arabic NLP has a clear lack of such theoretical and corpus-based discourse processing studies.

We present the first effort towards producing an Arabic Discourse Treebank- the LADTB, a news corpus where all discourse connectives are identified and annotated with the discourse relations they convey as well as with the two arguments they relate. We discuss our collection of Arabic discourse connectives as well as principles for identifying and annotating them in context, taking into account properties specific to Arabic. In particular, we deal with the fact that Arabic has a rich morphology: we therefore include clitics, nouns and prepositions as connectives, as well as a wide range of nominalizations as potential arguments. We present also a dedicated discourse annotation tool for Arabic and a large-scale annotation study. We show that both the human identification of discourse connectives and the determination of the discourse relations they convey are reliable. LADTB corpus encompasses a final 6328 annotated discourse connectives in 535 news texts. LADTB v.1 will be released soon via LDC. It is used currently in Leeds for training and testing the first automated methods for discourse connective and relation recognition.

The dual tagging approach of the Modern Arabic Representative Corpus 2000 (MARC-2000)

Marc Van Mol
Katholieke Universiteit Leuven

At Leuven University the MARC-2000 Corpus has been developed. This corpus contains a representative sample of all kinds of Arabic Language material. The peculiarities of this corpus lies in the fact that all material dates from the beginning of the third millennium and that it has been collected at random in order to obtain a true representative sample of use of Arabic Language at that time. It contains both written and oral sources from different countries in the Arab world. This corpus was not copied from raw existing sources, such as the internet or CD's from newspapers. All data were entered into the computer in a specific way.

We distinguish, as far as the tagging is concerned between the primary or preparatory tagging of the corpus and the second or the definitive tagging of the corpus. The primary tagging is based on the partially use of the Arabic diacritical signs. These preparatory tags are very easy to implement by every Arab educated person and flows from the nature of Arab language itself. It demands a training of less than a week for a typist to master this way of Arabic typing. The preparatory tagging of the entire corpus of ca. 12,000,000 words lasted more than a decade. The definite tagging of the corpus is executed by confronting the primary tagged words in context with a lexicographical database (which served for the development of the in Belgium and Holland well known learners' dictionaries for Arabic). In this database the same primary tags were generated by programming all derived forms for all words in the database according to the same convention as the primary tags. The definite tags in this database are multiple. There are the elaborated tags based on Latin Parts of Speech and on the other hand the elaborated tags based on Arabic Traditional grammar.

Because of the possibility of combining these two tagsets the database tags are not simply twofold: viz. Latin and Arabic but multifold because a combination of European and Arabic tags is possible. So far, the secondary tags have been completely elaborated. The whole lexicographical database, originally in 4D format, has been transformed to a mysql database. The following steps will consist in the integration of the corpus in the database transforming the existing text format (so far encoded in Mac-ASCII) into utf-8.

Underneath the hood of arabiCorpus.byu.edu

Dilworth B. Parkinson
Brigham Young University

This paper will review what arabiCorpus.byu.edu does and does not do, and will give an inside look (given the time constraints of the presentation) into how it was programmed. arabiCorpus was designed not so much for Arabic language researchers, but for Arabic teachers and students to be able to quickly find numerous examples of specific Arabic words and constructions in a KWIC concordance format. This was done because much more technically sophisticated corpora have also been relatively inaccessible to 'normal' non-techie people because the interface was difficult to use and the results generated were not in an easily digestible format. This project therefore paid as much attention to user interface as it did to the manipulation of the texts. Because it is based on raw rather than lemmatized text, it is simply a given that it will generate false hits. The program provides some methods for reducing these, but also simply relies on the user to realize that false hits are a given with this (kind of) program.

First time users are often surprised to find out that the program seems to understand Arabic morphology, and can conjugate verbs, and understand prefixes and suffixes etc. This is, of course, simply a programming trick. The 'guts' of the search engine simply search for every single example of a particular input string, and then filter the results with some cleverly designed regular expressions that reflect Arabic morphology. This cuts out the false hits that can be predicted morphologically. Of course, it does nothing to cull out ambiguous forms which can only be distinguished by syntactic or collocational context.

The paper will give details both of how a well-trained user can take advantage of the tools available to get the best results possible, will show where the program succeeds and fails despite these tools, and will give a detailed example of how one of the part of speech filters was programmed. Plans for future development both of the corpus itself and the associated interface and engine will be presented at the conclusion of the talk.

Corpus linguistics resource and tools for Arabic lexicography

Majdi Sawalha and Eric Atwell
School of Computing, University of Leeds

Corpora have been used to construct dictionaries since the release of the Collins-Birmingham University International Database COBUILD (Ooi, 1998). Large and representative corpus provides detailed information about all aspects of written language that can be studied. Corpus analysis tools (such as Sketch Engine, www.sketchengine.co.uk) are used to build a detailed statistical profile of any word in the corpus, which enables lexicographers to understand the words or collocations, their behaviors, usages and indicating the connotations they may carry, etc. Oxford dictionaries (http://www.oxforddictionaries.com) represent an exemplar of the use of corpus in constructing dictionaries. Besides; citations which represent the objective evidence of language in use, are a prerequisite for a reliable dictionary but they have their limitations (Atkins and Rundell, 2008).

However, Arabic corpora have not been used to construct traditional monolingual Arabic dictionaries. The last Arabic dictionary المُعْجَمُ الوَسِيْطُ *mu‘jam al-wasīṭ* “Al-Waseet Lexicon” appeared in the 1960’s from the Arabic language academy in Cairo. The advances in corpora construction technologies, corpora analysis tools and the availability of large quantities of Arabic text of different domains, formats and genres on the web can allow us to build a large and representative lexicographic corpus of Arabic to be used in constructing new Arabic dictionaries (for instance the Arabic Internet corpus http://smlc09.leeds.ac.uk/query-ar.html which consists of 176 million words). A lemmatizing tool is needed to group words that share the same lemma to be studied. It also helps in finding the collocations of the word.

The second important resource of information needed to construct new Arabic dictionaries is the long established traditional Arabic lexicons. Over the past 1200 years, many different kinds of Arabic lexicons were constructed; these lexicons are different in ordering, size and aim or goal of construction. The traditional Arabic lexicons followed four main methodologies for ordering their lexical entries. These methodologies use the root as lexical entry. The main disadvantage of these methodologies is that the derived words of the root are not arranged within the lexical entry. Ordering of dictionary entries is the main challenge of constructing Arabic dictionaries.

Traditional Arabic lexicons represent a citation bank to be used in the construction of modern Arabic dictionaries. They include citations for each lexical entry from the Qur’an and the authentic poetry that represents the proper use of keywords. They provide information about the origin of the words. They also include the phrases, collocations, idioms, famous personal names and places derived from that root (lexical entry).

The corpus of traditional Arabic lexicons is a collection of 23 lexicons. It represents a different domain than existing Arabic corpora. It covers a period of more than 1200 years. And it consists of a large number of words about 14,369,570 and about 2,184,315 word types. The corpus of traditional Arabic lexicons has both types of Arabic text; vowelized and non-vowelized text.

References

Atkins, B. T. S. & Rundell, M. (2008) *The Oxford guide to practical lexicography* Oxford ; New York Oxford University Press.

Ooi, V. B. Y. (1998) *Computer corpus lexicography* Edinburgh, Edinburgh University Press.

Sawalha, M. & Atwell, E. (2010) Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *Language Resource and Evaluation Conference LREC 2010*. Valleta, Malta.

Semantic prosody as a tool for translating prepositions in the Holy Qur’an: A corpus-based analysis

Nagwa Younis
Ain Shams University, Egypt

One of the most challenging aspects of translating the Holy Quran is to reflect the shades of meaning conveyed by the use of certain prepositions in the Arabic text. Prepositions are used in the Holy Quran not only as a syntactic requirement but also as a semantic and rhetorical function. It is the hypothesis of this research that there is a ‘semantic prosody’ related to the use of one preposition or another in a certain linguistic context. The researcher hypothesises that there is a semantic prosody related to certain prepositions especially when they are preceded by the same verb. This semantic prosody makes it inaccurate for the translator to use the same English word as an equivalent for the translation of the same verb-preposition construction when the verb is followed by more than one preposition in various linguistic contexts. For example, when the passive verb ‘ʔolqeya’(was thrown) is followed by the preposition ‘ʕala’(lit. on/upon) it collocates with words that denote ‘heavy duty’ and ‘gross responsibility’, whereas when the same verb is followed by the preposition ‘ʔela’ (lit. to/towards) it has the collocates that denote ‘delivering/giving something’.

The aim of study is to examine how the change of the semantic prosody concomitant with the change of preposition is reflected in translation. This is done through scrutinising a parallel corpus of six translations of the Holy Quran provided in the Quranic Corpus (Dukes, 2010).The study is only confined to examining the translation of prepositions in verb-preposition constructions where the preposition plays a role in changing the meaning of the verb. Special emphasis is given to the prepositions ‘ʕala’, ‘ʔela’ and ‘li-’. The results of the study shed light on some linguistic aspects in the translation of prepositions in the Holy Quran. These insights are of importance both in the field of Linguistics in general and Translation Studies in particular.

Using the Web to model Modern and Quranic Arabic

Eric Atwell
School of Computing, Leeds University

An initial survey (Atwell et al 2004) found few publicly-available Arabic language computing resources; but we found that Machine Learning could be used to adapt generic NLP techniques to Arabic (Abu Shawar and Atwell 2004, 2005). This required an Arabic text training set, so we developed the first freely-available Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006), and Arabic concordance visualisation toolkit (Roberts et al 2006).

We also developed tools for Modern Arabic text analytics: morphological analysis, stemming, and tagging (Sawaha and Atwell 2008, 2009, 2010), and Arabic discourse analysis (Al-Saif and Markert 2010). We have also extended our analytics techniques to Classical Arabic in the Quran, including question-answering (Abu Shawar and Atwell 2004), knowledge representation (Sharaf and Atwell 2009) and syntactic annotation (Dukes et al 2010).

The Corpus of Contemporary Arabic has been widely re-used in Arabic NLP research, for training and evaluation of systems. Our Quranic Arabic Corpus website http://corpus.quran.com/ has become a widely-used resource, not just by Arabic and Quranic researchers, but by general public wanting online tools to explore and understand the Quran. This has led us to propose "Understanding the Quran" as a new Grand Challenge for Computer Science and Artificial Intelligence for 2010 and beyond (Atwell et al 2010).

References

Atwell, Eric; Al-Sulaiti, Latifa; Al-Osaimi, Saleh; Abu Shawar, Bayan. 2004. A review of Arabic corpus analysis tools in: Bel, B & Marlien, I (editors) *Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*

Abu Shawar, Bayan; Atwell, Eric. 2004. An Arabic chatbot giving answers from the Qur'an in: Bel, B & Marlien, I (editors) *Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*

Abu Shawar, Bayan; Atwell, Eric. 2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, vol. 10, pp. 489-516.

Al-Sulaiti, Latifa; Atwell, Eric. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, vol. 11, pp. 135-171.

Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. 2006 aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora* journal, vol. 1, pp. 39-57

Sawalha, Majdi; Atwell, Eric. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers, in: *Proc COLING.2008 22nd International Conference on Computational Linguistics*.

Atwell, Eric; Al-Sulaiti, Latifa; Sharoff, Serge. 2009. Arabic and Arab English in the Arab World, in: *Proc CL2009 International Conference on Corpus Linguistics*.

Sawalha, Majdi; Atwell, Eric. 2009. Linguistically Informed and Corpus Informed Morphological Analysis of Arabic, in: *Proc CL2009 International Conference on Corpus Linguistics*.

Sharaf, Abdul-Baquee; Atwell, Eric. 2009. A Corpus-based Computational Model for Knowledge Representation of the Quran, in: *Proc CL2009 International Conference on Corpus Linguistics*.

Sawalha, Majdi; Atwell, Eric. 2010. Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text, in *Proc LREC.2010: Language Resources and Evaluation Conference*.

Sawalha, Majdi; Atwell, Eric. 2010. Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic, in *Proc LREC.2010: Language Resources and Evaluation Conference*.

Al-Saif, Amal; Markert, Katja. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. in *Proc LREC.2010: Language Resources and Evaluation Conference*.

Dukes, Kais; Atwell, Eric; Sharaf, Abdul-Baquee. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank, in *Proc LREC.2010: Language Resources and Evaluation Conference*.

Atwell, Eric et al. 2010. Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence, in *Proc GCCR.10 Grand Challenges in Computing Research for 2010 and beyond*

Arabic plurals in context: a corpus study

Petr Zemánek and Jiří Milička
Charles University, Prague

The paper focuses on analysis of the behaviour of plurals in Arabic. The presented analysis is based on the assumption that there is a strong link between the singular and plural of a noun which is reflected also in the usage of both forms, i.e. there is a strong match between the contexts in which the two forms appear.

The hypothesis outlined above was tested on corpora of both Classical and Modern Standard Arabic. The extent of the two corpora is 380 million word tokens for the diachronic corpus and 50 million word tokens for the MSA one.

The procedure consisted of several steps. A list of paired forms (singular and plural) was checked against the corpora, their contexts were mapped and the extent of concord of the mapped contexts was measured. The analysis also took into account possible polysemy of the forms in the list as well as possible differences between plurals belonging to one singular (such as *mawǧāt* vs. *ˊamwāǧ*, *buyūt* vs. *ˊabyāt* etc.) and considered the role of such differences in the disambiguation of singular meanings.

In order to check our results and provide a basis for comparison, we tested our algorithms on a corpus consisting of randomly transposed word tokens. These trials assured us that the methods we developed are free of gross errors.

All the steps taken with paired nouns were also carried out with a list consisting of pairs of forms derived from the same root, but not exhibiting the singular / plural relation. As an outcome we found out to what extent the similarities between contexts represent a general feature of all morphologically related words. Our analysis is a contribution to the discussion on whether we can rely on the context comparison when trying to determine the relation between words.

We conclude our paper by considering consequences of the results and possible practical applications within both corpus linguistics and natural language processing.

For a relational approach to modern literary Arabic conditional clauses

Manuel Sartori
Institut français du Proche-Orient

Based on novels written in Modern Standard Arabic published between 1963 and 2005 and from the entire Arab world, this article suggests how the hypothetical systems of this variety of language no longer correspond to the established "classical" model. Specifically, it demonstrates after having analysed them that the so called MSA grammar books are, facing the reality of the texts, descriptively inadequate. It then shows how the modern Arabic conditional clause, in its literary level, has created a kind of sequence of tenses, certainly influenced by European languages such as French and English. Therefore this is no longer the operator of the hypothetical system (iḏā, in and law) that enables us to understand the meaning of a conditional clause, but the relationship existing between the operator of the hypothetical system's protasis and the verbal form of the apodosis of that system.

Multifactorial methods for exploring contextual factors in the usage of Modern Standard Arabic come verbs

Dana Abdulrahim, John Newman and Sally Rice
University of Alberta

Within a usage-based, constructionist framework the behaviour of a lexical item is best understood in its context of use and not in isolation. It follows then that the syntactic structures in which it appears, the morphological inflections associated with it, the other lexical elements that co-occur with it in a phrase, etc., all contribute to the (conventionalized) meaning or function expressed by a linguistic item. This approach, therefore, calls for moving beyond single semantic, morphological, or syntactic properties of a lexical item and scrutinizing the entire lexico-syntactic frame in which it appears. The availability of corpora caters to such an analytical approach since they provide a large amount of naturally-occurring, contextualized uses (as opposed to introspective and elicited utterances that may not reflect actual language usage), as well as providing voluminous amounts of linguistic data that permit a quantitative treatment of the phenomena under investigation.

In this paper we will attempt to demonstrate the analytical potential of a corpus-based multivariant data frame in which a large number of the lexico-syntactic properties of utterances hosting certain lexical items are specified. The lexical items investigated via this method are four MSA verbs of COMING: ˀ**ata**, **ǧaˀa**, **qadima** and **ḥaḍara**. For each of these four verbs we constructed a data frame that is typically composed of a large number of corpus concordance lines where each verb appears in its natural context of use. In this data frame, every concordance line is examined and marked up for a large spectrum of morphosyntactic and semantic features. This includes the syntactic structure that hosts the verb, the patterns of verbal inflections for every instance of verb use (e.g. subject number, person, and gender, as well as aspect for the Arabic verb), the semantic properties of the other elements of the construction (e.g. subject animacy and semantic category), as well as the inclusion/exclusion of, for example, phrases denoting a starting point of the event (SOURCE), a terminal point of the event (GOAL), as well as specification of the PATH of motion.

The potential of such heavily annotated data frame can be explored in a number of ways and via different statistical tests that are designed to handle both mono- and multi-factorial datasets, and can therefore provide a reliable account of each verb's lexico-syntactic profile. In this paper we specifically report on multi-factorial statistical tests including cluster analysis and Hierarchical Configural Frequency Analysis (von Eye, 1990; Gries, 2008). The cluster analysis (Behavioral Profiles) test we conducted on this data frame was developed by Stefan Gries (2009) as a script to be run in R statistical software. The BP test provides a good measure of the distance between the four verbs as they are used in MSA, based on the variables each verb usage was coded for. This should determine whether, for instance, two of the four verbs appear in similar constructions and are therefore closer in usage. On the other hand, HCFA, also an R script developed by Gries (2009), is more concerned with highlighting the interaction between the different levels of variables and, therefore, determines what variables co-occur more frequently than would be expected by chance. Such a statistical test provides insights into what could constitute a prototypical usage of a certain lexical item.

References

Gries, Stefan Th. (2009). Statistics for linguistics with R: A practical introduction. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th. (2009). HCFA 3.2 – A Program for hierarchical configural frequency analysis for R for windows.

Gries, Stefan Th. (2009). BehavioralProfiles 1.01. A program for R 2.7.1 and higher.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, URL <http://www.R-project.org>.

von Eye, A. (1990). Introduction to configural frequency analysis: The search for types and antitypes in crossclassification. Cambridge: Cambridge University Press.

Using an Arabic corpus for recognition and translation of Arabic named entities with NooJ

Héla Fehri[*], Kais Haddar[†] and Abdelmajid Ben Hamadou[†]

[*] MIRACL-University of Franche-Comte and university of Sfax, Tunisia

[†] MIRACL-University of Sfax, Tunisia

To develop linguistic resources allowing the elaboration of named entity recognition and translation tool in any domain, we need to use a rich corpus. This permits to construct dictionaries with large coverage and rule systems that can treat morphological and syntactical phenomena. Moreover, the corpus is necessary to evaluate tools given to process named entities (NEs). Besides, the study corpus can help in the refinement stage applied to NE hierarchy of the chosen domain.

It is in this context that is situated the present paper. In fact our mainly objective is to construct a tool allowing the recognition of Arabic NEs and their translation into French language. Let's note that the domain used for this work is the sport domain.

From the Arabic collected corpus considered as a study corpus, we have firstly refined the inspired MUC NE hierarchy. Secondly, we have identified rules representing all possibilities of NE constituents and resolving problems related to the Arabic language like agglutination, vowelation, etc... These rules are described by grammars and dictionaries (Team names, player names, etc.) written in the linguistic platform NooJ. Problems related to the Arabic language are resolved using morphological grammars. However, the rules allowing recognition are represented by syntactical local grammars. Arabic dictionary entries should be voweled except for the later character for recognizing Arabic NEs whatever the corpus (voweled, not voweled or semi voweled). Thirdly, we have translated the extracted NEs. To experiment and evaluate the developed tool for recognition and translation in NooJ, we have used a corpus formed by 4000 texts of sport domain (different of the study corpus). This corpus contains texts of different newspapers like el sabeh, el Anwar, el chorouk, el ahram, etc. The performance measures of the obtained results gives 98% of precision, 90% of recall and 94% of F-measure.

As application, the developed tool can be used on the one hand to annotate corpora and on the other hand to identify sport corpora. In fact, if the corpus contains a representative number of NEs related to the sport domain and belonging to different categories of this domain, then we can deduce that this corpus is a sport corpus.

Moreover, the translation module allows to the no Arabic speaker to understand the main idea of sport corpora. Let's note that we have integrated a transliteration module that can improve the translation phase and be used in e-learning application.

Automated speech act classification in Arabic

Lubna A. Shala[*], Vasile Rus[*] and Arthur C. Graesser

[*] Department of Computer Science, University of Memphis†

† Department of Psychology, University of Memphis

Arabic Natural Language Processing (A-NLP) research has gained an increasing interest in the last few years for many reasons including underdeveloped computational methods to process it. Here, we present a fully-automated method for the task of speech act classification for Arabic discourse. The task of speech act classification involves assigning a category from a set of predefined speech act categories to a sentence to indicate speaker's intention. In particular, we worked with the following set of predefined categories: assertion, declaration, denial, expressive evaluation, greeting, indirect request, question, promise/denial, response to question, and short response.

Our approach to speech act classification is based on the hypothesis that the initial words in a sentence and/or their parts-of-speech are very diagnostic of the particular speech act expressed in the sentence. We have tested this hypothesis on more than 1000 Arabic sentences collected from several Arabic news sources including newspaper articles and television shows.

We experimented with two machine learning algorithms, naïve Bayes and Decision Trees, to induce speech act classifiers for Arabic texts. To model the task of speech act classification, we used as features the first 3, 4, or 5 words in a sentence (the so-called sentence-initial context), the parts of speech tags of these words, and both the words and tags, i.e. the word-tag pairs. The parts of speech of the words were automatically obtained using an Arabic tagger, AMIRA 2.0. To handle short sentences (less than 5 words, e.g. greetings), we used a NULL default part of speech category for the non-existing words.

We have also experimented with several other models in which we used bigrams and trigrams of parts-of-speech as features. The basic idea is to capture positional/sequential information about the parts of speech, which could be important when identifying speech acts. To obtain bigrams of parts-of-speech, we simply concatenated two consecutive parts of speech into one feature. As before, we only considered parts of speech for the first 3, 4, and 5 words in a sentence. We also introduced before the first word a fake part of speech, START, such that we could generate a bigram for the first word. Then, we paired the first part-of-speech with the second, the second with the third, and so on, generating five features for the first five words. These features were used in conjunction with the same algorithms to induce speech act classifiers.

A gold standard approach was used for evaluation in that the collected sentences were manually annotated by an Arabic scholar with correct speech acts. The evaluation was conducted based on a 10-fold cross-validation method in which the available data set is divided in 10 folds and for each fold a classifier is induced. The classifier is derived from 9 folds and tested on the remaining fold. The overall performance is the average over the 10 folds.

Combining corpus-based and linguistic models for Arabic speech systems

Hanady Ahmed[*] and Allan Ramsay[†]

[*] Qatar University

[†] Manchester University

Automatic generation (text-to-speech synthesis (TTS)) and recognition of spoken Arabic speech (automatic speech recognition (ASR)) is a challenging task. Automatic generation and recognition of any language is hard enough, but Arabic has a number of properties that make it even harder. In particular, the non-concatenative nature of Arabic morphology and the range of permitted word orders mean that is very hard to provide language models of the kind that are required for training speech recognizers, and the lack of diacritics in written Modern Standard Arabic (MSA) make it difficult to determine the underlying phonetic forms required for speech synthesis.

The proposed research aims to improve the performance of an existing computational linguistic treatment of Arabic in order to make it suitable for use in these areas. The existing engine was originally developed for use within a TTS system, and the planned research will allow this system to be used with a much wider lexicon and with fewer restrictions on the form of the input text than was the case with the prototype. The main aim of the proposed research, however, is to extend the natural language processing engine (NLP) so that it can also be used as the basis for a language model for speech recognition.

Speech recognition engines require a ‘language model’ to help constrain the search for words that match the acoustic properties of the speech signal. Such language models are typically supplied as context-free grammars.

The existing linguistic engine can be used to produce analyses of input text which can in turn be used to generate a context-free grammar of the kind that is required for speech recognition. The analyses produced by the linguistic engine are fine-grained dependency trees, annotated with a variety of syntactic and semantic features.

In order to use the current engine for this task, we need to add corpus-based information, e.g. statistical part-of-speech tagging, probabilities relating to various non-canonical word orders, converting phoneme-to allophone rules, and to extend its lexicon. The existing engine provides very fine-grained analyses, but it is easily swamped when faced with unrestricted text. The main aim of the current project is to improve the performance of the existing engine in the face of long sentences and a wide vocabulary, by adding statistical evidence to the existing rule-based approach and by extending the lexicon using resources such as Arabic Treebank , Buckwalter Arabic morphological analyzer, and SAMPA Analyzer. The outcome will be that the system can be used to generate language models for speech recognition, and that its existing deployment for speech synthesis will also become more widely applicable.