Corpus Linguistic Tools for Historical Semantics in Arabic

Omaima Ismail, Sane Yagi and Bassam Hammo
University of Jordan, Jordan

Abstract: *In this paper, we present a set of corpus linguistic tools for conducting historical semantic research in the Arabic language. We compiled a Historical Arabic Corpus (HAC) that spans more than 1500 years of continuous language use. With techniques from the field of Natural Language Processing (NLP), the tools we presented here have been used to create the HAC and to explore lexical semantic change. The development of these tools is aimed at offering a catalyst to the ambitions goal of compiling an Arabic dictionary on historical principles. HAC and the tools can also be used for conducting research in a variety of areas of linguistics.*

Keywords: Arabic historical corpus, diachronic semantics, etymology, computational lexicography, lingusitic tools, NLP resources.

1. Introduction

Corpus Linguistics is a sub-discipline of the scientific study of language that uses machine-aided tools for the compilation, retrieval, and analysis of a large body of classified, machine-searchable texts that are representative of authentic language use. It uses frequencies, collocations, and phrase structures to make generalizations about language use.

A corpus is a database of a large body of classified machine-searchable texts that are selected to be representative of language as spoken or written **in** a specific geographic region or period of time, **by** a specific group of users, and/or **for** a specific function. The texts are often meta-encoded with information about the author, date, place, and medium of publication, genre, language, degree of representativeness, etc. They are also annotated at word and/or sentence level with information about a word's part of speech (POS), grammatical function, morphological components, prosodic features, etc.

Corpora are at the basis of a variety of recent linguistic studies. They are considered a good resource for research in natural language processing, teaching languages, machine translation, language engineering, information retrieval, lexical analysis, lexicography, and many others. (Al-Sulaiti and Atwell, 2006)

Only few researches have been conducted on Arabic from a historical linguistic perspective. One reason for this is the fact that most NLP tools at the disposal of linguists have been geared to Modern Standard Arabic (MSA). The absence of an Arabic corpus on historical principles is at the root of this research deficiency.

Arabic has one of the longest traditions in lexicography, with the first full-fledged dictionary dating back to 786 C.E. It also has one of the richest works of lexicography with dictionaries that are onomasiological, semasiological, alphabetical, retrogradic, encyclopaedic, terminological, etc. What it does not have is a dictionary that traces the semantic development of words; ie, an etymological dictionary (a dictionary on historical principles).

Towards the ultimate end of initiating systematic work in the direction of historical linguistic research and lexicography, we have leveraged information technology. This paper aims at describing, previewing, and demonstrating a set of computational tools that would facilitate research in historical semantics and etymological lexicography. It provides a set of tools that were used to create and analyze the Historical Arabic Corpus (HAC) in order to extract historical semantic knowledge about Arabic.

The rest of the paper is organized as follows: Section 2 defines historical semantics and discusses semantic change. Section 3 provides a background on corpus linguistics and reviews some previous work. In section 4, we give a description of the research methodology we followed in the development of the corpus tools. Section 5 shows some experiments that were conducted for illustrating the utility of our work. Section 6 concludes and draws a roadmap for future research.

2. Historical semantics and meaning change

Historical semantics is a scientific enterprise that studies diachronic change of meaning. It identifies, describes, and explains this change and attempts to discover the conditions that motivate it, the factors that regulate it, and the mechanisms that propagate it. It also studies the consequences of this change for meaning relations within and across semantic fields. Furthermore, historical semantics studies the conceptual history of a language community by investigating the etymology of words.

Fritz (2012) asserts that historical semantics is also “a research area where fundamental problems of semantics tend to surface and which can be seen as a testing ground for theories of meaning and for methodologies of semantic description”, p. 2644.

On the benefit that historical semantics has gained from corpus linguistics, Fritz (2012) maintains that corpus linguistics has inspired historical semantics to reflect on the relationship between collocations of a word and its senses. Historical corpora facilitate the study of gradual changes in contexts of use and make it possible to discriminate between new senses that are transient and new senses that develop firmly.

Semantic change, meaning change, denotes the universal tendency of word meanings to become different over time by gaining new senses or losing old ones, replacing default senses, drifting in terms of word prototype, narrowing or widening category boundaries, pejoration or amelioration, and/or bleaching.

As in all languages, Arabic words change meaning over time. Below we present some examples of the major types of semantic shift that affected Arabic words. These examples have been culled from our Historical Arabic Corpus to illustrate both some types of semantic change and the potential of our compiled corpus. Table 1 shows examples of words which underwent semantic specialization. We found out that many of the cases of semantic change are due to specialization (or semantic narrowing).

Table 1. Examples of words that underwent specialization

Word	**Classical meaning**	**Meaning in modern contexts**
أمانة *?amana*	Trust	Safe deposit; honesty; integrity; fidelity; confidence
حَرَج *h̲araj*	Extreme distress	Shyness; modesty
نجم *najm*	A star in the sky	Acelebrity

Broadening the senses of a word is another type of semantic change. Classical Arabic words that have undergone semantic generalization in MSA are few (cf. Table 2).

Table 2. Examples of words that underwent semantic generalization

Word	**Classical meaning**	**Meaning in modern contexts**
رؤيا *ru:?ya*	Good dream	Vision, trance, dream
حيث *h̲aythu*	Adverb of place (where)	Adverb

Semantic pejoration (degradation) and semantic amelioration (elevation) also occur in modern Arabic contexts. Tables 3 and 4 illustrate both types.

Table 3. Examples of words that underwent semantic pejoration

Quranic word	**Meaning in the Quran**	**Meaning in modern contexts**
حيوان *h̲ayawa:n*	True eternal life	Animal
إرهاب *?irha:b*	Deterrence of aggressors	Indiscriminate assault

Table 4. Examples of words that underwent semantic amelioration

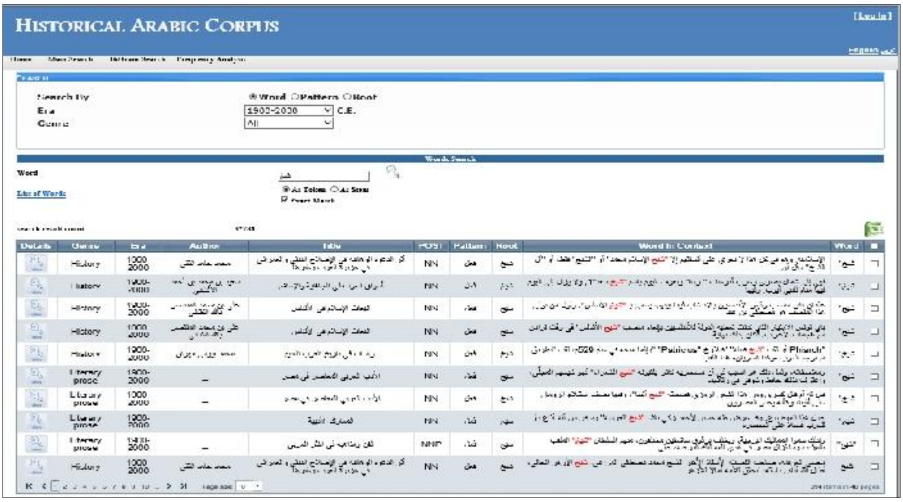
Word	Meaning in the Quran	Meaning in modern contexts
شيخ *shaikh*	Old man	Leader; prince; religious scholar
مَرَح *marah̲*	Falsehood	Delight and joy
زبانية *zaba:niya*	Guardian spirits of Hell	Accomplices

Now, let us focus on one word, شيخ‘*shaikh*’, and inspect how it acquired elevated senses in modern times. Figure 1 shows a snapshot from the concordance of *n-gram* textual phrases for the word ‘*shaikh*’ in current newspaper editorials from a variety of Arab countries. Whilst this word meant in Quranic Arabic 'old man', it means in Modern Arabic 'old man' as well as 'chief', 'religious leader', 'supreme religious leader', and 'scholar'.

In this type of semantic change, the original meaning of a word transforms to an entirely different sense. Table 5 shows few examples of this type of semantic change.

Table 5. Examples of words that underwent semantic transition

Quranic word	Original meaning in Quran	Meaning in modern context
جهاز *jiha:z*	Furniture	Machinery
ناصية *nas̲iya*	Exhausting	Erecting; fraudulent
يَشْرِى *yashri:*	Sells	Buys



HISTORICAL ARABIC CORPUS

Search By ● Word ○ Pattern ○ Root
Era 1900-2000 C.E.
Genre All

Word Search

Word شيخ
List of Words
● As Token ○ As Stem
☑ Exact Match

Details	Genre	Era	Author	Title	POS	Pattern	Root	Word In Context	Word	■
	History	1900-2000	[illegible]	[illegible]	NN	فعل	شيخ	[illegible]	شيخ	□
	History	1900-2000	[illegible]	[illegible]	NN	[illegible]	[illegible]	[illegible]	[illegible]	□
	History	1900-2000	[illegible]	[illegible]	NN	[illegible]	شيخ	[illegible]	شيخ	□
	History	1900-2000	[illegible]	[illegible]	NN	فعل	شيخ	[illegible]	شيخ	□
	History	1900-2000	[illegible]	[illegible]	NN	[illegible]	[illegible]	[illegible]	[illegible]	□
	Literary prose	1900-2000	–	الأدب العربي المعاصر في مصر	NN	[illegible]	شيخ	[illegible]	شيخ	□
	Literary prose	1900-2000	–	[illegible]	NN	[illegible]	[illegible]	[illegible]	[illegible]	□
	Literary prose	1900-2000		المعارك الأدبية	NN	[illegible]	[illegible]	[illegible]	[illegible]	□
	Literary prose	1400-2000	–	[illegible]	NNP	[illegible]	شيخ	[illegible]	[illegible]	□
	History	1900-2000	[illegible]	[illegible]	NN	فعل	شيخ	[illegible]	شيخ	□

Fig 1. A snapshot from the concordance, tracking the word شيخ‘*shaikh*’ in MSA

Table 6 illustrates how morphological change plays an important role in the semantic development of words. In modern contexts, some plural forms of Quranic words are now used as singular and new plural forms have been coined.

Table 6. Examples of Quranic words that underwent morphological change

Quranic word	**Meaning in the Quran**	**Meaning in modern contexts**
حَاجّ ‘ha:j’	Pilgrims	One pilgrim, the plural being حُجاج‘huja:j’
نَفَر ‘nafar’	A group of people	One person, the plural being أنفار ‘anfa:r’

3. Background and literature review

As corpora are large and structured collections of texts, electronically stored and processed, they are good resources for linguistic research, natural language processing, and data mining. In order to make corpora more useful for linguistic research, they are often annotated with different information depending on the purpose that the corpus is used for. For example, a corpus could be annotated with Part of Speech Tags (POST), and some morphological information such as a word’s lemma, stem, root, morphological pattern, etc. Corpora are considered fundamental to computational linguistic research and to the study of language use in the real word. They are used in lexicography, translation, language learning and teaching, data mining, etc.

Historical corpora consist of texts from periods that span the entire history of a language or part of it and they are used by historical linguists to explore the development of the language over time. Such corpora would reveal how words changed meaning and how grammatical structures changed. Historical corpora are indispensable for the compilation of historical dictionaries, not only for tracing the changes in a word’s meaning but also for providing quotations that illustrate the senses of a word. Arabic is in dire need for a dictionary on historical principles.

There are many Arabic corpora with a range of structures and annotations. In terms of annotation, Al-Sulaiti (2004) developed a corpus of contemporary Arabic, which included modern standard Arabic texts and samples of colloquial varieties. The purpose of this corpus was to enrich resources for teaching and researching Arabic. The corpus contained one million words, marked up with Extensible Markup Language (XML). They were collected mainly from magazines, newspapers, websites, and radio stations. XML is a language that utilizes a set of rules for encoding documents in a human-readable and machine-readable format. Al-Sulaiti’s XML encoding contains tags for general information about the texts but no tags for morphological or POS information.

Alansary, et al. (2007), built the International Corpus of Arabic (ICA) for the purpose of evaluating methods of information extraction from Arabic documents. They planned for it to contain 100 million words of Modern Standard Arabic, selected from a range of resources. They also built software for the ICA to query the corpus and to insert documents. It can detect the genre and source of inserted documents and it places them in an appropriate hierarchy. The structure of the corpus, however, is not encoded in XML. Alansary, et al. used annotation to add information about the inserted documents but they never used morphological annotation. They later appended their software with a morphological analysis module that utilized the Buckwalter analyzer (Alansary, et al., 2008). Some other XML structures were used for lexicography purposes, e.g., the iSPEDAL. In the light of their own research, the authors proposed an improved structured electronic dictionary in the form of a relational database or in the form of XML documents; most Arabic dictionaries are found in flat textual form. The system contains such data as affixes, morphological patterns, and derived words and all words are linked to their roots (Hajjar, et al., 2010).

Attia, et al. (2010) used Arabic corpora of MSA to build a lexicon encoded in Lexical Markup Framework (LMF). Their model was made such that it would automatically obtain lexical information from corpora for the purpose of constructing a large lexical resource. They also provided a complete description of inflectional and syntactic behavior of Arabic lexical entries. They developed a web system called AraComLex to help in lexicographic work which featured morphological patterns, sub-categorization frames, and Arabic lemma. (Attia, et al., 2011)

Another work is the SALAH project. Boella, et al. (2011) proposed a model for segmenting and linguistically analyzing classical Arabic texts of Prophet Muhammad's traditions (Hadieth) and the narratives on his life and deeds (Sunna). It divides text units into: transmitters' chain (isnad) and text content (matn). The final system outputs an XML format that contains relations among transmitters and a lemmatized text corpus that is used in the automatic generation of concordance texts. The authors suggest that the system be used for information retrieval from Hadieth texts, and for verifying relations between transmitters.

One of the most important Arabic corpora is the Quranic Arabic Corpus, an annotated corpus of Quranic text that uses dependency grammar to provide multiple layers of annotation, including morphological segmentation, part-of-speech tagging, and syntactic analysis. The Quranic corpus is automatically annotated by Buckwalter Arabic Morphological Analyzer and is then manually verified. The fully annotated corpus can be browsed online, and is encoded in both XML and plain text format (Dukes and Habash, 2010). Another work on the Quran presents a Quranic corpus tagged with personal pronouns and antecedents. They named this product QurAna. SimQur (Sharaf and Atwell, 2012 a; Sharaf and Atwell, 2012 b) is yet another Quranic corpus but it is concerned with semantically related verses. These corpora are resources for Quran and hadieth scholars and students, as well as computational linguists and computer scientists. The majority of the corpora above, however, either refrain from using simple structures and annotations, or are limited in scope and restricted to primary religious texts (i.e., the Quran and Hadieth). None is concerned with how the Arabic language developed over time.

On the other hand, there are numerous English historical corpora that contain samples of texts from earlier eras. These corpora are used to study language variation in earlier periods of English as well as language change and development. The Helsinki Corpus (2011), for instance, is a structured multi-genre corpus that includes periodically organized text samples from Old, Middle, and Early Modern English. It can be used for giving general information on the occurrence of forms, structures and lexemes in different periods of English. Another English historical corpus is ARCHER (2014) (A Representative Corpus of Historical English Registers), which is a multi-genre corpus of British and American English covering the period of 1600-1999.

Other languages have historical corpora as well. For example, Sánchez-Marco, et al. (2011) present a general method that adapts existing NLP tools to facilitate dealing with historical varieties of languages. They implemented these tools for Old Spanish. They used them to automatically add linguistic information to texts and to annotate them for their historical corpus.

Until the present time, there has been no published historical corpora for Arabic and neither have there been techniques for dealing with historical texts or annotating them. Most Arabic corpora are created manually or by simple tools that compile texts in an XML format and add annotation as meta-data. Khoja (2009) created a software application that can download RSS feeds to compile a corpus. It uses Arabic blogs of both modern standard and colloquial Arabic. The software converts the blogs into a corpus encoded in XML. It uses such tags for meta-data as author, gender, country, blog URL, etc. O'Donnell, (2008) also created the UAM Corpus Tool, which is an application for annotating texts using different linguistic layers, where the user can define the hierarchy of tags appropriate for the layer. The annotations can be at document layer level (e.g., text type, writer, register, etc.), at semantic-pragmatic level, and at syntactic level (e.g., clause, phrase). While the central task of the corpus tool is annotation, it also provides other functionalities, such as cross-layer searching, semi-automatic tagging, production of statistical reports, visualization of the tagged corpus, inter-coder reliability statistics, etc. It also enables users to add annotation manually, and it stores the annotation data using XML.

Other tools used commonly for dealing with corpora are concordances, tools used in analyzing corpora, and searching for and retrieving words in context. Concordances have been shown to be an effective aid in the acquisition of a second or foreign language, because they facilitate the learning of vocabulary, collocations, grammar, and writing styles. Linguists can use concordance output to understand language behavior at morphological, lexical, syntactic, and semantic levels. Lexicographers can also use concordance output to identify multiple senses of a word.

There are numerous concordancing tools for English and European languages, like MonoConc, WordSmith, WordPilot, etc. and most of them are commercial products. Little work is there to support Arabic and its unique morphological features. Anthony (2005) created AntConc, which is a multiplatform, multipurpose freeware corpus analysis toolkit, designed specifically for use in the classroom. It consists of a concordance, word and keyword frequency generator, tools for cluster and lexical bundle analysis, and a word distribution plotter. It also offers the choice of simple wildcard searches or regular expression searches. Although it can handle corpus text in UTF-8 encoding, it is not very efficient in handling Arabic texts.

Roberts, et al. (2006) built a concordancer called aConCorde. It supports Arabic and has both Arabic and English interfaces, and can be used on multiple platforms. The aConCorde provides an interactive stem-based searching facility and displays Arabic text correctly, but it still has some limitations. The full text of a selected item can't be seen by users. It is proposed as a data-driven language learning tool for Arabic and as a tool for lexicographers and linguists.

Abbès and Dichy (2008) developed AraConc as an interactive software specifically for Arabic. It integrates the Arabic word-form analyzer and generator (MorphArab) that is based on a lexicon generated from the DIINAR.1 knowledge database (DIctionnaire INformatisé del'ARabe, version 1). The AraConc software allows the building of new lexical resources, and widens the descriptive scope of the analyses used to construct DIINAR.1. AraConc inputs texts, extracts word-forms, updates their occurrences, and sends them one by one to the morphological analyzer (MorphArab). Analyses are then saved together with the position of the word in a document, and dispatched into specific files. Also the results of analyses are stored in a relational structure, which offers users a great number of choices in the grouping of output information and in statistics.

It is evident from the above that the available tools are useful but there is need for work that can provide well-structured schemas and tools for Arabic corpora and that are capable of handling morphological annotation, and building an annotated corpus automatically. Historical corpora for Arabic has not received enough attention from scholars; hence, this research is very important for the construction, query, and analysis of an Arabic historical corpus.

NLP in the Arabic language is still in its initial stage compared to the work in the English language, which has already benefited from extensive research by scholars from all over the world. There are obstacles that slow down progress in Arabic NLP compared to the accomplishments in English and other European languages (Al-Daimi and Abdel-Amir, 1994). These hurdles include:

- Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task.
- The absence of diacritics (which represent short vowels) in the written text creates ambiguity and therefore, complex morphological rules are required to identify the tokens and parse text material.
- Capitalization is not used in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations.

In addition to these, there is also lack of free Arabic corpora, lexicons, and sophisticated machine-readable dictionaries, resources that are essential to advancing research in NLP. In the last decade, the volume of Arabic textual data on the web has started to grow and Arabic software for browsing the web is improving. Unfortunately, much of Classical Arabic texts available on the web was posted as images, which makes it unsuitable for searching and machine processing. There is, however, a gradual increase in the amount of Arabic news material available on the web.

4. Methodology and developed tools

Compiling a historical corpus is not our primary target but rather the development of tools for building such a corpus, tools that can automatically annotate texts and categorize them per historical era. Here, we will explain our framework, corpus data, the proposed schema, the tools we developed, and our system's architecture, in addition to the tool's features and functionalities.

Corpus data

Before collecting the corpus data, we considered the issue of textual representation. The most important factor that we focused on was that the data must cover all time periods of Arabic. That is why we classified the corpus data time-wise starting from pre-Islamic times until the current century, and divided this time span into periods of 100 years each, calling them eras.

We also classified the corpus data into primary and secondary sources on the basis of how representative a text is of its time of authorship. A primary text is poetry and literary prose, as well as non-fiction that does not comment on texts of older eras. But secondary texts include the language used in commentaries on older texts, commentaries that are expected to reflect how language was used at the time when the commentator lived, while the language of the text being explained shows how people of older times used the language. Quran exegesis and critical commentary on poetry of older times are two examples of secondary texts.

Another factor we considered is a text's genre. It is linguistically well established that genre affects the language used in a text; hence, we classified our corpus texts into Literary Prose; Poetry; History; Philosophy; Religion; Science; Thought; Dictionaries; Others.

In addition to era, genre, and primary/secondary categorization, we collected general information about the texts to be compiled into the corpus, information such as document title and author. Below is a table that illustrates the variety of texts that make up our corpus on historical principles and how they have been annotated in terms of author, era, genre, and category.

XML schema

XML is an acronym for Extensible Markup Language. This is an open and popular standard for marking up text in a way that is both machine and human readable. By 'marking up text' we mean that the data in the text files is formatted such that it would include meaningful symbols that represent what that data is for.

We designed a novel XML schema for the corpus, creating tags for each document's metadata, and a text token's morphological annotation, and stored each token in a single tag together with its annotation attributes. The annotations we stored with each token are: root, morphological pattern, POST, and light stem.

Corpus Builder

The corpus builder application was developed to compile the corpus and encode it in the proposed XML schema automatically. The corpus builder takes as input a text file encoded in utf-8 together with its document meta-data, and processes the document. Furthermore, we decided to integrate a stemmer and a POS tagger in our corpus builder in order to create the corpus as required. We adapted Khoja's Arabic Stemmer (Khoja, 1999) to extract word roots and morphological patterns and stems. For tagging, we used Stanford Part-Of-Speech Tagger (Toutanova et al., 2003).

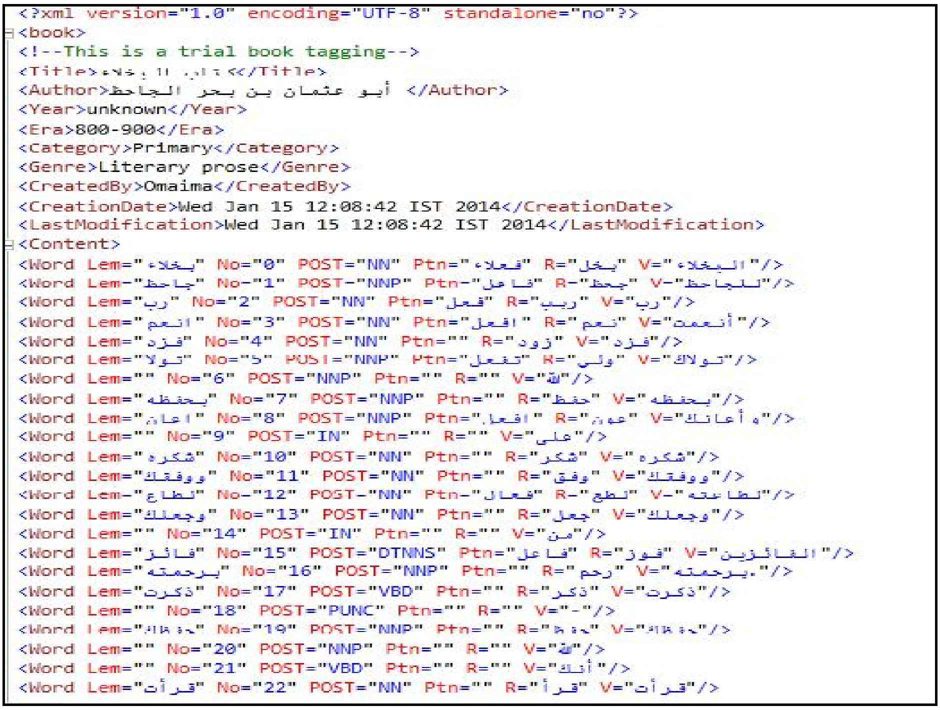
This Corpus Builder application:

1. Uploads a document with its meta-data
2. Processes and tokenizes the document's text
3. Extracts roots of words using Khoja's modified Arabic Stemmer
4. Retrieves the morphological pattern and light stem for each word in the document
5. Tags with Stanford Part-Of-Speech Tagger each word with its part of speech
6. Compiles all information into XML
7. Produces a morphologically and syntactically annotated XML file of the processed document.

Table 7. Sample of corpus documents

Document Title	**Author**	**Era**	**Genre**	**Category**
ديوان امرئ القيس	امْرُؤُ القَيْس	Before 600 C.E.	Poetry	Primary
العين	الفراهيدي	700-800 C.E.	Dictionaries	Secondary
البخلاء	الجاحظ	800-900 C.E.	Literary prose	Primary
القانون فى الطب	ابن سينا	1000-1100 C.E.	Science	Primary
الإحاطة في أخبار غرناطة	لسان الدين ابن الخطيب	1300-1400 C.E.	History	Primary
تفسير الجلالين	جلال الدين المحلّي وجلال الدين السيوطي	1400-1500 C.E.	Religion	Secondary
آراء أهل المدينة الفاضلة	الفارابي	1500-1600 C.E.	Philosophy	Primary
الأعمال الشعرية الكاملة لابراهيم طوقان	ابراهيمطوقان	1900- 2000 C.E.	Poetry	Primary

Figure 2 shows a portion of an annotated XML file, and table 8 lists the symbols representing the annotations made in 2-5 above.



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<book>
<!--This is a trial book tagging-->
<Title>كتاب البخلاء</Title>
<Author>أبو عثمان بن بحر الجاحظ </Author>
<Year>unknown</Year>
<Era>800-900</Era>
<Category>Primary</Category>
<Genre>Literary prose</Genre>
<CreatedBy>Omaima</CreatedBy>
<CreationDate>Wed Jan 15 12:08:42 IST 2014</CreationDate>
<LastModification>Wed Jan 15 12:08:42 IST 2014</LastModification>
<Content>
<Word Lem="بخلاء" No="0" POST="NN" Ptn="فعلاء" R="بخل" V="البخلاء"/>
<Word Lem="جاحظ" No="1" POST="NNP" Ptn="فاعل" R="جحظ" V="للجاحظ"/>
<Word Lem="رب" No="2" POST="NN" Ptn="فعل" R="ربب" V="رب"/>
<Word Lem="انعم" No="3" POST="NN" Ptn="افعل" R="نعم" V="أنعمت"/>
<Word Lem="فزد" No="4" POST="NN" Ptn="" R="زود" V="فزد"/>
<Word Lem="تولا" No="5" POST="NNP" Ptn="تفعل" R="ولي" V="تولاك"/>
<Word Lem="" No="6" POST="NNP" Ptn="" R="" V="الله"/>
<Word Lem="بحفظه" No="7" POST="NNP" Ptn="" R="حفظ" V="بحفظه"/>
<Word Lem="اعانه" No="8" POST="NNP" Ptn="افعال" R="عون" V="وأعانك"/>
<Word Lem="" No="9" POST="IN" Ptn="" R="" V="على"/>
<Word Lem="شكره" No="10" POST="NN" Ptn="" R="شكر" V="شكره"/>
<Word Lem="ووفقك" No="11" POST="NN" Ptn="" R="وفق" V="ووفقك"/>
<Word Lem="لطاع" No="12" POST="NN" Ptn="فعال" R="لطع" V="لطاعته"/>
<Word Lem="وجعلك" No="13" POST="NN" Ptn="" R="جعل" V="وجعلك"/>
<Word Lem="" No="14" POST="IN" Ptn="" R="" V="من"/>
<Word Lem="فائز" No="15" POST="DTNNS" Ptn="فاعل" R="فوز" V="الفائزين"/>
<Word Lem="برحمته" No="16" POST="NNP" Ptn="" R="رحم" V="برحمته."/>
<Word Lem="ذكرت" No="17" POST="VBD" Ptn="" R="ذكر" V="ذكرت"/>
<Word Lem="" No="18" POST="PUNC" Ptn="" R="" V="-"/>
<Word Lem="سبحانه" No="19" POST="NNP" Ptn="" R="سبح" V="سبحانه"/>
<Word Lem="" No="20" POST="NNP" Ptn="" R="" V="الله"/>
<Word Lem="" No="21" POST="VBD" Ptn="" R="" V="أنك"/>
<Word Lem="قرأت" No="22" POST="NN" Ptn="" R="قرأ" V="قرأت"/>
```

Figure 2: Portion of an XML Corpus File

Table 8. Word tag attributes

Attribute	**Description**
No	The sequence number of the word in context
V	The word value as a token (the word itself)
R	Root of the word
Ptn	Morphological pattern of the word
POST	Part of speech tag for the word in context
Lem	Light stem of the word

Historical Arabic Concordancing and Searching System (HACSS)

The HACSS was created to facilitate tracing the development of linguistic aspects of Arabic across time. It does not use a database engine but rather XML corpus files and other XML files purposefully created for information retrieval. This system consists of four main modules: term indexer, term search engine, concordancer, dictionary editor.

Indexer

In Computer Science, an ‘index’ is “a list of keywords associated with a record or document, used especially as an aid in searching for information”[1]. It is used to assist in information retrieval systems in order to save search time especially if the number of documents to search through is huge.

In full text searching and when all results are equally needed, a simple inverted index or a Boolean index is enough for searching and retrieving all occurrences of a search term in all documents. The inverted index contains a list of references to documents associated with each term.

In HACSS, we created a set of index files for word stems and another set for roots. Each index file stores the terms that start with one of the letters of the alphabet. A corpus XML document is indexed for one time by identifying its unique terms (words and roots), then storing each of them in the appropriate index file in accordance with their word-initial letters. If the term already exists in the index file, we add only its new reference.

Search Engine

The HACSS’ search engine can search for a word or a phrase, a root, or a morphological pattern and is capable of retrieving words and their contexts from any specified era.

The Search Engine provides this set of functionalities:

1- Different types of search:

Searching by Word: The search engine looks for words in their light stem form. It also offers the chance to search for a word by exact or partial matching.

Searching by Root: This will retrieve all the words derived from that root that occurred in the corpus.

Searching by Morphological Pattern: This feature will retrieve all words that were coined in the morphological template of the search term.

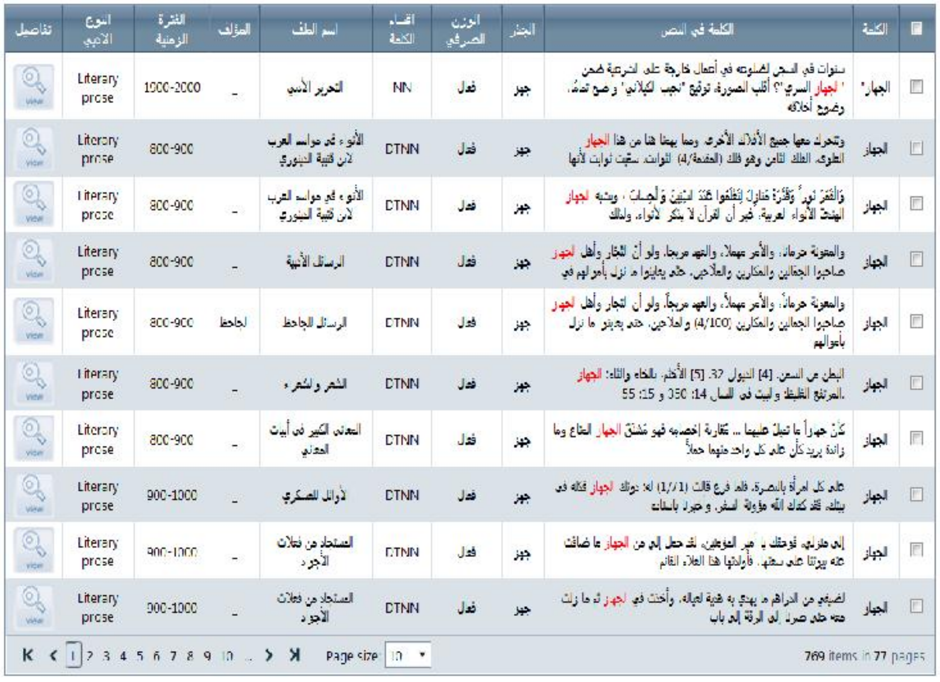
2- The system provides a list of eras and a list of genres to search within. The user can search for a term in a specific historical era, or in a specific genre.

Concordancer

The search engine’s results are displayed by the concordancer. It extracts the matched terms and their immediate contexts and compiles them in the form of a concordance list. The size of context is user determined; the user can specify for extraction and display the number of words preceding the search term and following it. The concordancer also displays morphological information stored in the corpus with the searched term (i.e. the root, pattern, and POST) in

[1] http://www.thefreedictionary.com/indexing

addition to the source document metadata (i.e. document title, author, genre, and historical era). Figure 3 shows a concordance page for the term ‘*Jiha:z*’.



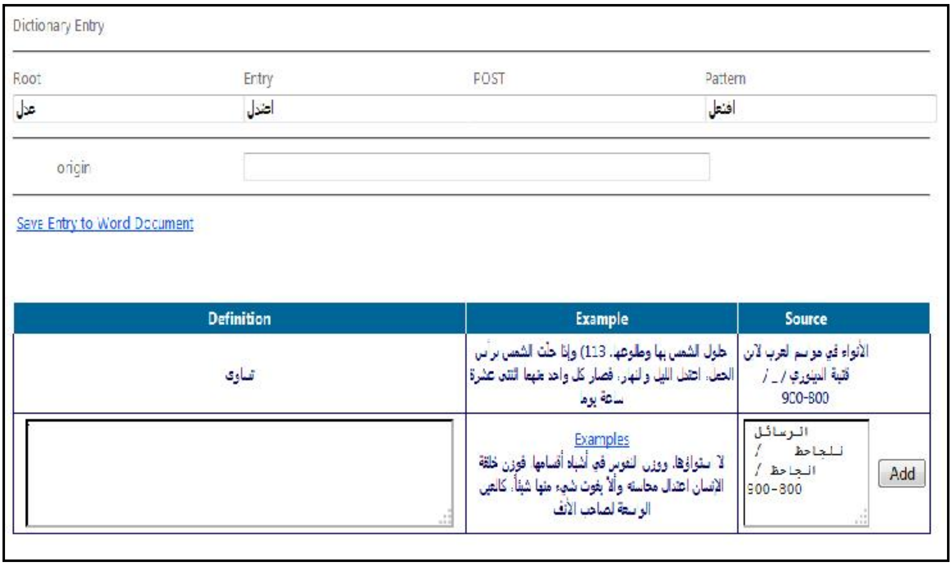
تفاصيل	النوع الأدبي	الفترة الزمنية	المؤلف	اسم الملف	أقسام الكلمة	الوزن الصرفي	الجذر	الكلمة في النص	الكلمة	☐
view	Literary prose	1900-2000	_	التحرير الأدبي	NN	فعال	جهز	سنوات في السجن لضلوعه في أعمال خارجة على الشرعية ضمن ' الجهاز السري'؟ أقلب الصورة، توقيع 'نجيب الكيلاني' وضح تماماً، وضوح أخلاقه	الجهاز'	☐
view	Literary prose	800-900		الأنواء في مواسم العرب لابن قتيبة الدينوري	DTNN	فعال	جهز	وتتحرك معها جميع الأفلاك الأخرى وما يهمنا هنا من هذا الجهاز الطوى، الفلك الثامن وهو فلك (المقدمة/4) الثوابت. سميت ثوابت لأنها	الجهاز	☐
view	Literary prose	800-900	_	الأنواء في مواسم العرب لابن قتيبة الدينوري	DTNN	فعال	جهز	وَالْقَمَرَ نُوراً وَقَدَّرَهُ مَنازِلَ لِتَعْلَمُوا عَدَدَ السِّنِينَ وَالْحِسابَ ، وشبه الجهاز الهند الأنواء العربية. غير أن القرآن لا ينكر الأنواء، ولذلك	الجهاز	☐
view	Literary prose	800-900	_	الرسائل الأدبية	DTNN	فعال	جهز	والمعونة حرماناً، والأمر مهملاً، والعهد مريجاً. ولو أن التجار وأهل الجهز صاحبوا الجمالين والمكارين والملاحين، حتى يعاينوا ما نزل بأموالهم في	الجهاز	☐
view	Literary prose	800-900	الجاحظ	الرسائل الجاحظ	DTNN	فعال	جهز	والمعونة حرماناً، والأمر مهملاً، والعهد مريجاً. ولو أن التجار وأهل الجهز صاحبوا الجمالين والمكارين (4/100) والملاحين، حتى يعاينوا ما نزل بأموالهم	الجهاز	☐
view	Literary prose	800-900	_	الشعر والشعراء	DTNN	فعال	جهز	البطن من السمن. [4] الديوان 32. [5] الأكثم، بالخاء والثاء: الجهاز المرتفع الغليظ. والبيت في اللسان 14: 350 و 15: 55.	الجهاز	☐
view	Literary prose	800-900	_	المعاني الكبير في أبيات المعاني	DTNN	فعال	جهز	كأنّ جهزاً ما تبيلُ عليهما ... عقارة إخصيه فهو مشتق الجهاز التاع وما زائدة يريد كأن على كل واحد منهما حملاً	الجهاز	☐
view	Literary prose	900-1000	_	الأوائل للعسكري	DTNN	فعال	جهز	على كل امرأة بالبصرة. فلم فرغ قالت (1/1) له: جوتك الجهاز فكله في بيتك. فقد كفاك الله مؤونة السفر، و [illegible]	الجهاز	☐
view	Literary prose	900-1000	_	المستجاد من فعلات الأجواد	DTNN	فعال	جهز	إلى منزليه فوحقك يا أمير المؤمنين لقد حمل إلي من الجهاز ما ضاقت عنه بيوتنا على سعتها. فأولدتها هذا العلاء القائم	الجهاز	☐
view	Literary prose	900-1000	_	المستجاد من فعلات الأجواد	DTNN	فعال	جهز	لضيفي من الدراهم ما يهدي به هدية لعياله، وأخذت في لجهز ثم ما زلت معه حتى صرنا إلى الرقة إلى باب	الجهاز	☐

K < 1 2 3 4 5 6 7 8 9 10 .. > K Page size 10 ▾ 769 items in 77 pages

Figure 3: A Sample Concordance Page for the Search Term ‘*Jiha:z*’

Dictionary Editor

The dictionary editor is an interface that the lexicographer interacts with. It fetches sentences from the concordance and plugs them into the example slot in the dictionary entry, extracts from the annotations and meta-tags such information as root, morphological pattern, part of speech, text author, and historical era, and plugs them into the appropriate slots in the Dictionary Entry Form. It also gives lexicographers the facility to add senses as deemed necessary, and to save the full dictionary entry in a Word document. The dictionary editor is designed to make the work of lexicographers easier and to save time. Figure 4 shows a snap shot of the dictionary editor's interface.



Dictionary Entry

Root	Entry	POST	Pattern
عدل	اعتدل		افتعل

origin

Save Entry to Word Document

Definition	Example	Source
تساوى	طول الشمس بها وطلوعها. 113) وإذا حلت الشمس برأس الحمل، اعتدل الليل والنهار، فصار كل واحد منهما اثنتى عشرة ساعة يوم	الأنواء في مواسم العرب لابن قتيبة الدينوري / _ / 900-800
	Examples لا تساوؤها. ووزن النفوس في أشباه أقسامها فوزن خلقة الإنسان اعتدال محاسنه وألا يفوت شيء منها شيئاً: كالعين الواسعة لصاحب الأنف	الرسائل للجاحظ / الجاحظ / 900-800

Figure 4: Dictionary Editor’s interface

5. Experiments and results

Here is a demonstration that uses our corpus tools to show word usage and the different senses of some terms which changed in time.

Change over time

Words undergo semantic change over time, and the word ‘*Jiha:z*’ *(ײַ̀Ųγ)* is no exception. Table 9 below shows sample sentences of this word extracted from our corpus.

Table 9: Senses of *“Jiha:z”* over time extracted from HAC

Era	Genre	Source	Word in Context	English Meaning
1100-1200 C.E.	Dictionaries	**مفردات القرآن** للاصفهاني	**الجهاز** ما يعد من متاع وغيره، **(قل تعالى: فلما** جهزهم بجهازهم)	Belongings
1300-1400 C.E.	History	**الإحاطة في أخبار غرناطة**	**وآخذ من** بنات الدهر حقي **جهاز** البيت استلب استلابا	Furniture
1800-1900 C.E.	History	**عجائب الآثار في التراجم والأخبار للجبرتي**	وفرشوها بأنواع الفرش الفاخرة ونقلوا إليها **جهاز** العروس والصناديق وما قدم إليها من الهدايا	Clothes
1900-2000 C.E.	Thought	**موسوعة** اليهود واليهودية و الصهيونية	حكومة الانتداب البريطاني هناك قررت تجنيدهم داخل **الجهاز** الحكومي كموظفين حتى يمكنها أن تبقيهم بمعزل	Staff
After 2000 C.E.	Science	تطبيقات البكتريا في **المجال الطبي**	وتتجاوز أنواعها الموجوده **في الجهاز** الهضمي أكثر من (3000) نوع من البكتريا	(Digestive) System
After 2000 C.E.	Science	تطبيقات البكتريا في **المجال الطبي**	يقومون بتسجيل وجود أي **أثر للمتفجرات من خلال جهاز** حديث تم تصنيعه في معامل أبحاث نيومكسيكو	Apparatus

HACSS is capable of calculating and graphing the frequency of a term’s usage over time. It shows the trend of popularity of a term across all eras. Figures 5 and 6 are charts that plot the chronology of some words in the corpus.

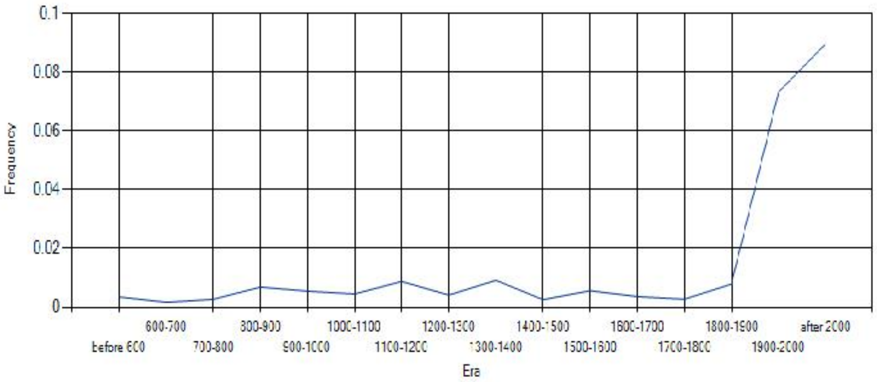


Figure 5: Timeline chart of the word *siyasa* (سياسة) ‘politics’

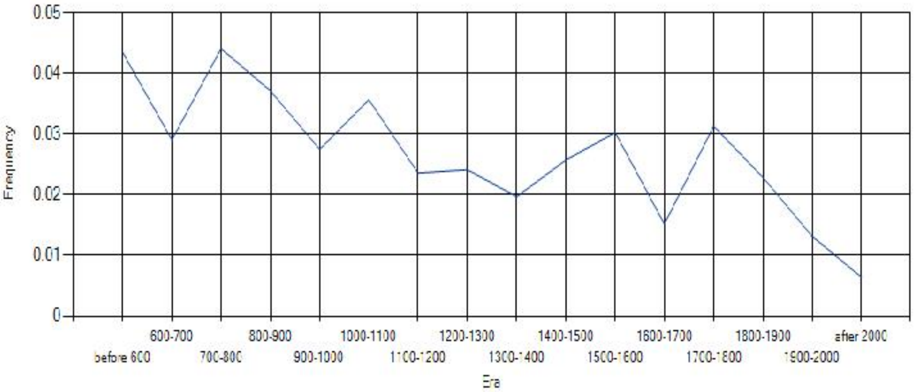


Figure 6: Timeline chart of the root ^{c}*tq* (عتق) ‘emancipation’

6. Conclusion and future work

The Historical Arabic Corpus together with the developed tools provide a good resource for extracting historical semantic knowledge. It has been made clear that HAC will enable linguists to study the Arabic language from a historical perspective, and will facilitate the work of lexicographers.

We have developed a set of tools for the creation and manipulation of a historical Arabic corpus. The corpus builder integrates a stemmer with a tagger to process and annotate documents, and then compile them into an XML corpus that uses a novel schema. We also created an indexer, a search engine, a concordancer, and a dictionary editor that together facilitate searching and extraction of linguistic knowledge from HAC, and facilitate the compilation of dictionary entries in a hypothetical dictionary on historical principles.

We aim to enhance HAC in the future by rendering more accurate annotation, expanding the corpus so that it would become more representative of Arabic through time, optimizing and adding functionalities to the search engine and concordancer, and reacting to linguists’ needs and offering them more flexibility.

Omaima Ismail
Computer Information Systems Department
University of Jordan, Amman, Jordan
omaima.i.ismail@gmail.com

Sane Yagi
Department of Linguistics
University of Jordan, Amman, Jordan
saneyagi@yahoo.com

Bassam Hammo
Computer Information Systems Department
University of Jordan, Amman, Jordan
b.hammo@ju.edu.jo3

References

Abbès, Ramzi, and Joseph Dichy. (2008). '*AraConc*, an Arabic concordance software based on the DIINAR. 1 language resource'. *The 6th International Conference on Informatics and Systems,*. Giza, Egypt. 127-134.

Alansary, Sameh, Magdy Nagi, and Noha Adly. (2007). 'Building an international corpus of Arabic (ICA): progress of compilation stage'. *The 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007*. 1-30.

Alansary, Sameh, Magdy Nagi, and Noha Adly. (2008). 'Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage'. *8th International Conference on Language Engineering, Egypt*. 1-23.

Al-Daimi, K., and Abdel-Amir, M. (1994). "The Syntactic Analysis of Arabic by Machine". Computers and Humanities, Vol. 28, No. 1, pp. 29-37.

Al-Sulaiti, Latifa, and Eric Atwell. (2006). 'The design of a corpus of contemporary Arabic'. *International Journal of Corpus Linguistics,* 11(2), 135-171.

Al-Sulaiti, Latifa. (2004). Designing and developing a corpus of contemporary Arabic. Doctoral dissertation, University of Leeds (School of Computing), UK.

Anthony, Laurence. (2005). 'AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom.' *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International. IEEE.*

ARCHER (A Representative Corpus of Historical English Registers). (2014). Available at: http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/. (Accessed on 20.12.2014).

Attia, Mohammed, et al. (2011). 'Lexical Profiling for Arabic'. *Proceedings of eLex*: 23-33.

Attia, Mohammed, Lamia Tounsi, and Josef van Genabith. (2010). 'Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic'. *Technical report, The NCLT Seminar Series, DCU, Dublin, Ireland.*

Boella, Marco, et al. (2011). 'The SALAH Project: segmentation and linguistic analysis of Ḥadīṯ Arabic texts'. *Information Retrieval Technology*. Springer Berlin Heidelberg. 538-549.

Dukes, Kais, and Nizar Habash. (2010). 'Morphological Annotation of Quranic Arabic'. *LREC.*

Friz, Gerd. (2012). 'Theories of Meaning Change: An Overview'. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning, Vol. 3.* Walter de Gruyter. 2625-2651.

Hajjar, Mohammad, et al. (2010). 'An Improved Structured and Progressive Electronic Dictionary for the Arabic Language: iSPEDAL'. *Internet and*

Web Applications and Services (ICIW), Fifth International Conference on. IEEE.

Helsinki Corpus of English Texts. (2011). Department of Modern Languages, University of Helsinki. Available at: http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/HC_XML.html. (Accessed 15 January 2015).

Khoja, Shereen, and Roger Garside. (1999). 'Stemming Arabic text'. Technical report, Computing Department, Lancaster University, Lancaster, UK.

Khoja, Shereen. (2009). 'An RSS feed analysis application and corpus builder'. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.. 115–118.

Khoja Stemmer. Available at: http://zeus.cs.pacificu.edu/shereen/research.htm. Retrieved in March. 2014.

O'Donnell, Mick. (2008). 'The UAM CorpusTool: Software for corpus annotation and exploration'. *Proceedings of the XXVI Congreso de AESLA*.

Roberts, Andrew, Latifa Al-Sulaiti, and Eric Atwell. (2006). 'aConCorde: Towards an open-source, extendable concordancer for Arabic'. *Corpora* 1.1.

Sánchez-Marco, Cristina, Gemma Boleda, and Lluís Padró. (2011). 'Extending the tool, or how to annotate historical language varieties'. *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*. Association for Computational Linguistics.

Sharaf, Abdul-Baquee M., and Eric Atwell. (2012a). 'QurAna: Corpus of the Quran annotated with Pronominal Anaphora'. *LREC*.

Sharaf, Abdul-Baquee M., and Eric Atwell. (2012b). 'QurSim: A corpus for evaluation of relatedness in short texts'. *LREC*.

Stanford Tagger. Available at: http://nlp.stanford.edu/downloads/tagger.shtml. Retrieved in March. 2014.

Toutanova, Kristina, et al. (2003). 'Feature-rich part-of-speech tagging with a cyclic dependency network'. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics.