Spiralling towards perfection: an incremental approach for mutual lexicon-tagger improvement

Karlheinz Moerth[1], Stephan Procházka[2], Omar Siam[1,2], Thierry Declerk[3]
[1]Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Vienna
[2]Institute for Near Eastern Studies, University of Vienna
Spitalgasse 2, Hof 4, A-1090 Wien
[3]German Research Centre for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, Campus D3 2, D-66123 Saarbruecken
E-mail: karlheinz.moerth@oeaw.ac.at, stephan.prochazka@univie.ac.at, simar@gmx.at, thierry.declerck@dfki.de

Abstract

Our paper describes an experiment in which four different digital language resources are used to incrementally create added value in one another. The resources are a digital dictionary, a morphological analyser, a tagger and a digital corpus. We will show how the dictionary is used to improve the tagger, how the tagger is used to annotate a collaboratively produced digital text collection, i.e. the Egyptian language Wikipedia, thus improving easily available open data and lastly how the results of the annotation process are – in turn – utilised to enhance and improve the dictionary. The paper touches on several issues related to the particular tasks involved in the process: we discuss problems of dealing with data retrieved from the internet, we give details on the lemmatisation, the creation of word-class information and the generation of frequency data from the corpus and we touch on issues of dictionary creation and aspects of the dictionary-corpus-interface. A final topic are standards for the representation of the statistical information in the digital dictionary.

Keywords: eLexicography, tools, tagger, corpus-dictionary interface, variational linguistics

1 Introduction

The research described in this paper has grown out of a master thesis (Siam 2013) aiming at determining the 200 most frequent words of the Wikipedia Masri, the Wikipedia version written in colloquial Egyptian Arabic. We will attempt to explain how we reached beyond the 200 most frequent items and how we achieved the word counts. In addition, we will give an outline of our approach to integrate such data into an existing digital dictionary. While the first attempt simply aimed at the creation of a word list of the 200 most frequent words, the follow-up project was supposed to create data from the Wikipedia corpus to enhance an existing lexicographic resource with statistical information.

1.1 Research questions

This research has also been motivated by the obvious lack of any reliable information concerning word frequencies in varieties of spoken Arabic. This kind of data would be helpful in raising the quality of many practical linguistic applications such as dictionaries, language courses etc. A first serious count of the words of Modern Standard Arabic (MSA) was put forward by Landau (1959). The most recent publication on this topic is Buckwalter & Parkinson's *A Frequency Dictionary of Arabic* (2011). Both projects investigated written MSA [1]. To our knowledge, no publicly available statistical analysis of any Arabic dialect has been undertaken so far.

Other objectives of the project were text technological methodologies for the build-up of digital corpora of colloquial Arabic, improving a particular lexicographical resource by increasing the number of entries and adding corpus-based statistical information to the entries.

1.2 Egyptian Arabic

Arabic is one of the six official languages of the United Nations and spoken by some 450 million people. In the Ethnologue summary listing the major languages of the world according to the number of first-language speakers Arabic holds rank number five[2].

The sociolinguistic situation in the Arabic countries is characterised by a phenomenon which by many linguists has been described as diglossia, the coexistence of two linguistic varieties within the same community. While the written standard is almost the same across the Arab World, the spoken varieties differ considerably both from the written form and from one another. Students of Arabic have to master two quite different linguistic systems in order to be able to cope with a sufficiently wide range of everyday situations.

What we are concerned with here is the variety of the Egyptian capital Cairo which – for the sake of convenience – is often called Egyptian Arabic. Actually, Egyptian Arabic is in itself quite differentiated and split up into several groups. Nevertheless, the use of the catch-all term is justified by the fact that the variety of the capital is virtually understood everywhere in the country. Beyond the national level, Egyptian Arabic is widely used in communication beyond the borders of Egypt. Even though active skills may be limited, most Arabs are

[1] Buckwalter & Parckinson (2011) also list vernacular items that appear in their MSA corpora.

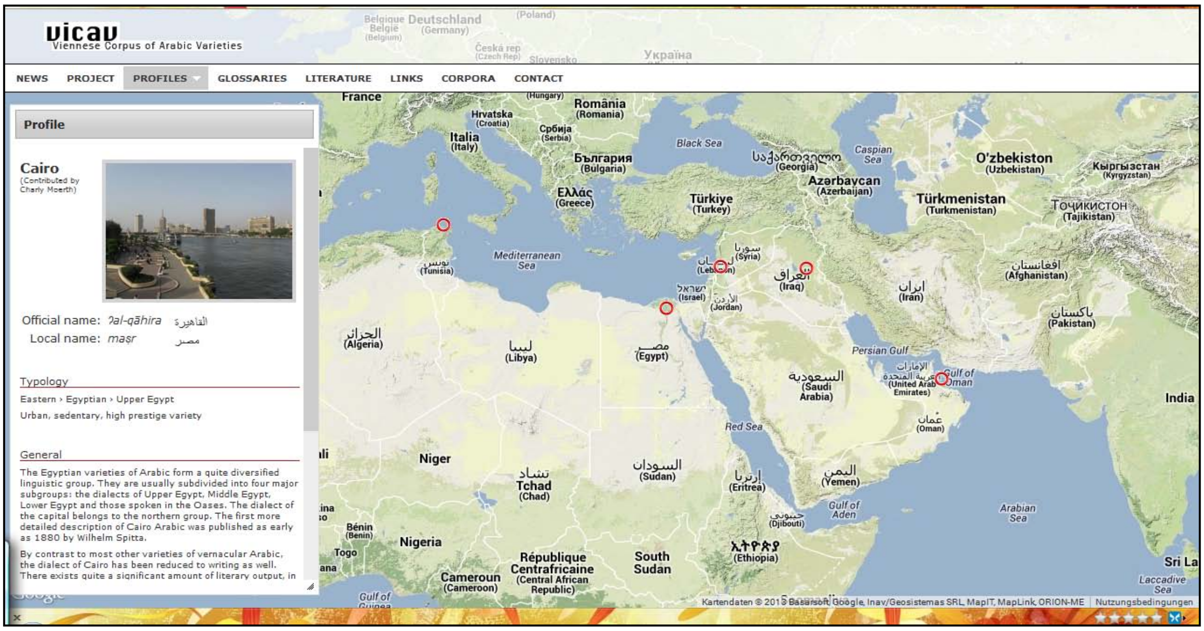
[2] http://www.ethnologue.com/statistics/size.

capable of understanding colloquial Egyptian Arabic throughout the Arabic World. Given Egypt's historical role in the region, the Egyptian brand of spoken Arabic is still regarded by many as the most prestigious form of colloquial Arabic.

Among the many forms of spoken Arabic, Egyptian Arabic is the one variety that in the recent past has also been widely reduced to written form. By contrast to other parts of the Arab World, colloquial language is not only used to write poetry, but also for prose and in particular drama. Numerous stage plays were written in the second half of the twentieth century. The advent of the internet has ushered in a new phase characterised by new types of communication. Meanwhile, many colloquial texts can be found in the Internet: on personal web sites, in public discussion forums, and in the social media. However, it is important to note that the use of colloquial language in writing has remained controversial and keeps being discussed with a lot of emotion the details of which need not be resumed here.[3]

2 The VICAV dictionaries

All of the described endeavours have been undertaken as part of the *Vienna Corpus of Arabic Varieties*, a joint project of the Austrian Academy of Sciences (Institute for Corpus Linguistics and Text Technology) and the University of Vienna (Centre for Near Eastern Studies). VICAV has been set up with two main purposes in mind: firstly, to serve as a virtual research environment targeting the particular needs of Arabic dialectology. The main concern will be textual and lexicographic data. Secondly, the platform will be used as a test bed for newly developed text technological methodologies and tools.



Profile

Cairo
(Contributed by Charly Moerth)

Official name: ʔal-qāhira القاهرة
Local name: maṣr مصر

Typology
Eastern › Egyptian › Upper Egypt
Urban, sedentary, high prestige variety

General
The Egyptian varieties of Arabic form a quite diversified linguistic group. They are usually subdivided into four major subgroups: the dialects of Upper Egypt, Middle Egypt, Lower Egypt and those spoken in the Oases. The dialect of the capital belongs to the northern group. The first more detailed description of Cairo Arabic was published as early as 1880 by Wilhelm Spitta.

By contrast to most other varieties of vernacular Arabic, the dialect of Cairo has been reduced to writing as well. There exists quite a significant amount of literary output, in

Figure 1: VICAV Interface

2.1 Arabic dialect lexicography

Arabic philology looks back on a tradition of several hundred years of fruitful lexicographic creativity. However, most of what has been achieved was directed towards the documentation of the classical language, in the recent past on the written standard. In the history of lexicography, dictionaries documenting Arabic dialects are a comparatively recent phenomenon. The first serious works were produced in the 19th century, the bulk of high-value productions stems from the second half of the 20th century. While there are a number of dictionaries for those interested in the Egyptian form of spoken Arabic, only a few (such as Hinds/Badawi 1986 or Jomier 1976) can be regarded as reliable lexicographic sources.

In recent years, Arabic lexicography has started to make use of digital technologies; and some recently published products also build on digital corpora (e.g. Hoogland et al. 2003). However, there are hardly any digital dictionaries available, let alone ones that come in a digitally reusable form, live up to modern ICT standards or cover varieties other than Modern Standard Arabic (MSA).

2.2 The VICAV Egyptian-English Dictionary

The dictionary used for our experiments is a trilingual Egyptian Arabic – English – German dictionary. It is part of a series of digital dictionaries which are being compiled with comparative research questions in mind. So far there exist similar lexical resources for Moroccan (Rabat), Tunis and Damascus Arabic. Currently, a web interface is being developed that will allow to query these resources together. These dictionaries have not been intended as comprehensive dictionaries of the respective linguistic varieties which makes the selection of entries to be offered to the user an even more important question. The dictionaries are also used in language teaching.

[3] A good summary of the debate can be found in Davies 2006.

The University of Vienna offers courses in all four varieties on a regular basis as part of their Arabic Studies curriculum.

2.3 Dictionary editing

Technically, the dictionaries all build on TEI (Text Encoding Initiative) and ISO (TC37) related standards, and have been compiled exclusively in and for the digital medium. All editing tasks have been carried out by means of the Viennese Lexicographic Editor (VLE)[4]. This tool is built around an XML editor that provides a number of functionalities typically needed in editing linguistic data. VLE was developed at the ICLTT as part of a number of lexicographic projects. Among other things, the requirement specification prescribed support for varying XML formats. The application was supposed to process standard-based lexicographic and terminographic data such as LMF, TBX, and TEI. It is also provided with simple scripting capabilities, a configurable interface to access corpora and support for XML based validation mechanisms.

3 The corpus

In creating corpora for under-resourced languages, the concept of web-as-corpus obviously plays a particularly important role. With the steadily increasing prevalence of access to the Internet also in the Arab World, written manifestations of Egyptian Arabic have become available in large numbers. The probably largest single digital source written in colloquial Egyptian is the Wikipedia Masri (*miṣri* means Egyptian in Arabic, *maṣri* is the colloquial Egyptian form), the Wikipedia in Egyptian Arabic which has been growing continuously over the past few years. While Egyptian Arabic is only one of many varieties of colloquial Arabic, the Wikipedia Masri has so far remained the only Wikipedia written in an Arabic dialect.

3.1 Wikipedia and Wikipedia Masri

The use of Wikipedia for serious research may for some be a breach of taboo as grave as the use of colloquial language for scholarly purposes. Be that as it may, there exist only very few, for many languages virtually no other digital text collections of comparable size and quality.

Wikipedia is a quite new phenomenon and its growth over the past decade has had some really groundbreaking influences on the way many of us work in the digital medium. The idea to use so-called wiki software for creating a digital encyclopaedia on the internet was born sometime around the turn of the millennium. The first attempts at setting up a predecessor of what later should evolve into Wikipedia were undertaken by Jimmy Wales who started this collaborative project in 2001.

Wikipedia started in English, but the makers of the new project started to think about extending its scope to other languages early on. The first non-English Wikipedia was the German language version which was created as early as 16 March 2001, only two months after the launch of the project. Interestingly, the next one to follow was Catalan (a few hours after the German version). Others followed suit: Japanese, French, Chinese, Dutch, Esperanto, Hebrew, Italian, Portuguese, Russian, Spanish, and Swedish were the next to join. Arabic was introduced on 18 November 2001.[5]

The Wikipedia in Egyptian Arabic is a relatively recent offspring of the Wikipedia family. It came into existence in 2008. The number of articles has kept growing over the past few years. However, the community of contributors has remained rather limited (Siam 2012).

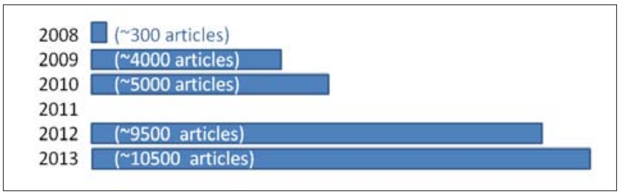


Figure 2: Growth of Wikipedia Masri

3.2 Cons and Pros

At the beginning of our project, the question arose repeatedly: why Wikipedia Masri. The question is not as straightforward as it looks on first sight. An argument that was brought against using this corpus was that it unarguably represents a very special type of language, a type of language particular to the medium. Actually, the language found in this digital resource does not reflect real spoken language, represents the written form of a language otherwise used primarily in oral communication.

As interesting as all the contained linguistic material may be, the kind of data we are dealing with poses some serious challenges. Scholars working on contemporary or historical non-canonical language are familiar with the whole range of issues such as for instance the high degree of graphematic variance. The main protagonists of Wikipedia Masri tried to create a guide for dealing with this kind of problems right at the outset of the project. They tried to achieve a true representation of many features of spoken Egyptian. However, the analysis of the data shows countless cases of orthographic inconsistencies. The word lists display many cases of spellings in the ‘new’ Wikipedia orthography next to traditional spellings.

The encyclopaedic aspiration of the Wikipedia project made it necessary to introduce a great number of neologisms. The texts contain a large amount of specialised vocabulary that has never before been used in Egyptian Arabic. Another issue is the large number of

[4] It is freely available at http://www.oeaw.ac.at/icltt/vle

[5] https://en.wikipedia.org/wiki/History_of_Wikipedia

named entities, especially those coming from outside of the Arabic language community.

In spite of the many problems, we decided to conduct our first experiments making use of Wikipedia Masri, mainly for a lack of real workable alternatives. Another option would have been to harvest the internet and to build a corpus of such data from scratch. However, building web-corpora for Arabic dialects is not as straightforward as with other languages. While there exist countless personal websites, discussion forums and similar material which contain text passages in colloquial Arabic, these texts are usually intermingled with MSA Arabic. The major task in dealing with this kind of data would be to identify relevant texts, text passages, on a more granular level, to find those sentences that reflect the linguistic registers we are interested in. The fact that no other coherent textual resource is available for the variety under investigation also made a strong case in favour of Wikipedia Masri.

Wikipedia Masri covers a wide range of topics, there are many of the basic categories found in other Wikipedias: famous persons, history, geography, mathematics, culture and arts, philosophy and religions, society and social sciences, health and medicine, natural sciences and technology are the main classified columns on the main page.[6]

Another argument in favour of this corpus was that all Wikipedias are available under a Creative Commons Attribution-Share Alike License which allows researchers to use the data for virtually any purpose except republishing the data under non-free terms.

3.3 Creating the corpus

To be able to extract statistical information from a corpus, a number of intermediate steps have to be taken which each require specialised tools. Steps involved in the build-up of corpora can differ. In the case of our particular corpus, we discern three obvious steps. Firstly, we needed tools to download the corpus. Secondly, the corpus had to be prepared for linguistic treatment and thirdly linguistic markup had to be created.

In all such projects, linguists are still confronted with a tangible lack in easily available and usable tools. In recent years, the situation has somewhat improved as far as tools for Modern Standard Arabic is concerned. Although there have been attempts to use some of these tools for colloquial forms of Arabic, the situation looks rather bleak for the colloquial varieties of Arabic. Specialised tools are required, resources for this kind of research are scarce. One of the goals of our project was the creation of a tool to annotate texts reflecting colloquial Arabic with word class and lemma information.

3.3.1 Acquisition

Getting hold of the texts of a Wikipedia appears to be very straightforward. All Wikipedias are exported in regular intervals from the central servers and put online as comprehensive ‘dumps’, large files containing the whole language version in one big document which can be downloaded from the Wikimedia Downloads website. Usually there exist a number of options and versions: versions with the complete edit history, versions containing all log events and so on. All of this data is offered in form of XML documents which sounds like good news. Actually, this is the point where the troubles start, as the data of the Wikipedia proper is encoded in another format. Unfortunately, the only genuinely XML part of these documents contains metadata of the respective articles, the text proper is written in so-called Wiki markup.

3.3.2 Wiki markup

Wikipedias are authored in what is generally called lightweight markup languages. Such languages have a simple syntax and were designed to be easy to learn and easy to use. Typically, this kind of markup language is used in collaborative web-based publishing. The particular Wikipedia brand of syntax and keywords used to edit the encyclopaedia is also called Wiki markup. The big drawback of this kind of markup system is the lack of consistency in its application which is mainly due to the fact that – by contrast to anything positioned in the XML world – the wiki markup language cannot be validated, i.e. automatically checked for structural integrity and logical consistency. To make things even more complicated, each language-specific Wikipedia has additional particular language-specific conventions which make the markup inaccessible to anybody not conversant in that language.

Many attempts have been undertaken at converting this kind of data towards more expressive, more reusable formats. The hitherto most important such undertaking is the DBpedia project, a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. The DBpedia website enables users to query relationships and properties associated with Wikipedia resources, including links to other related datasets. DBpedia builds on RDF (Resource Description Framework) and plays an important part in the Linked Data project.

3.3.3 Converting Wiki markup to TEI

At the ICLTT, researchers have experimented with three approaches: (a) writing specialised scripts to deal with each linguistic variety, (b) making use of existing software, and (c) working on the XHML instantiations of the Wikipedia articles.

The first option has proved to be time-consuming and very unrewarding. While the basic syntax of the Wiki languages is the same across the various Wikipedias, each of these contains many language-specific particularities

[6] http://arz.wikipedia.org/wiki/الصفحه_الرئيسيه

which makes tools developed for one Wikipedia unusable for others. For this reason this first approach was soon discarded.

Some of the existing software turned out to be usable for very specific purposes and languages only. Other tools did not work without considerable setup overhead. However, the evaluation of pieces of software such as Wikipedia2XML [7] (a collection of Python scripts) or Sweble [8] (a wikitext parser written in the Java programming language) for our particular purposes is an ongoing endeavour.

The main goal was to extract the texts in the language we were investigating, and to separate this core data from secondary information found in each Wikipedia page. To obtain reliable data it was necessary to be able to isolate the contents of a page from for instance link lists or lists of keywords which would have considerably flawed the results of any statistical analysis. Without such a separation, items such as the Egyptian equivalents of Wikipedia, link, file, template etc. would necessarily have been on top of the word list.

Our solution for this problem was to simply avoid processing of the Wikitext data. The only way to circumvent this was to work with the HTML output delivered by the MediaWiki software. First experiments with this approach showed that the data delivered from the Wikipedia website is neat and clean XHTML, HTML which complies with the basic XML conventions. Creating XML from this data proved to be easily manageable.

The final corpus creation workflow was very simple. First, all documents had to be downloaded from the Wikipedia website. There exist many tools to achieve such a task. GNU Wget is one the most widely used tools to retrieve content from web servers. We made use of the downloadBrowser that has been developed for other purposes at the department (Moerth et al 2011). The complete list of article names which are needed to download the data, is also available at the Wikipedia download site. From this a list of URLs can be created which can be used to perform the bulk-download.

Once the data were downloaded, they could be converted to TEI conformant XML documents. This conversion was carried out by means of a very simple XSLT stylesheet.

3.4 Basic pre-processing

The baseline preprocessing of the corpus also requires tokenisation. The approach taken at this stage is very straightforward, the tokeniser does not do anything more sophisticated such as decliticisation (splitting off clitics) or orthographic normalisation. All tokens which are simply defined by whitespace and punctuation were furnished with TEI conformant *w* tags. Repetitive secondary data such as keywords were excluded from processing at this stage in order to be able to analyse relevant textual data only.

4 Method/Workflow

Merely counting word forms would lead to inconclusive data. To determine the number of words in the corpus, it was necessary to furnish the corpus data with basic linguistic information, i.e. it had to be lemmatised. To our knowledge, there exist no freely available tagged data that could be used to train existing taggers. We had therefore to start from scratch. Our tool basically proceeds in a very simplistic manner and makes heavily use of the data contained in the dictionary.

The tagger works on the basis of a list of inflected word forms. For each token in the corpus, the tagger checks whether the word form is in the list. If the word form is found, the tagger assigns a POS and lemma attribute to the token. At the end of this process, the non-assigned tokens are statistically evaluated. If a word form could not be identified this means that the word does not exist in the dictionary. The words correlating with the most frequent unassigned word forms were then added to the dictionary.

4.1 Word form creator

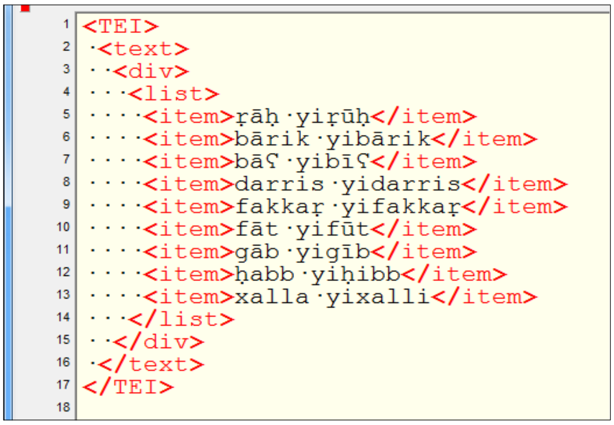
The word forms are produced by means of several tools, simple scripts that attempt to create virtually all possible word forms of the lemmas contained in the dictionary.

4.1.1 Verbs

The difficult part in creating all possible word forms from the finite set of items contained in our dictionary were the verbs. Arabic verb morphology is considerably more complex than nominal morphology. Still, many Arabic dialects have astonishingly regular paradigms. The Egyptian variety spoken in Cairo is one of those.

In order to obtain a list of all verb forms, the afore mentioned lexicographic editor is first used to export a TEI conformant list of all verbs contained in the dictionary. Each item is made up of two inflected forms: masculine singular third person perfect and imperfect. In Arabic philology the first one is usually used as the canonical form in dictionaries. Arabic has verbal nouns which are similar in some respects to infinitives. However, these are usually not used as headwords in dictionaries. Verbal nouns would not be very suitable as headwords as Arabic varieties show a high degree of overabundance in these forms. Many verbs can have several correlating verbal nouns which in some cases display semantic or functional differences, but often are simply competing forms.

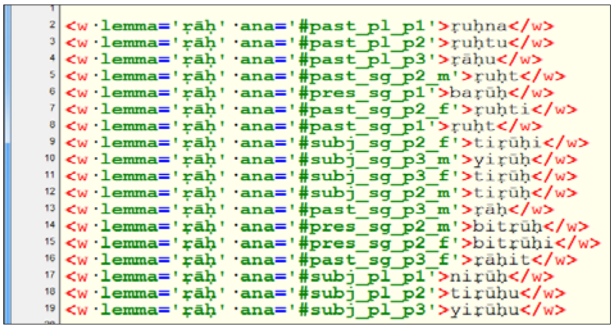
[7] wikipedia2xml.sourceforge.net
[8] http://sweble.org



```
<TEI>
·<text>
··<div>
···<list>
····<item>ṛāḥ·yiṛūḥ</item>
····<item>bārik·yibārik</item>
····<item>bāʕ·yibīʕ</item>
····<item>darris·yidarris</item>
····<item>fakkaṛ·yifakkaṛ</item>
····<item>fāt·yifūt</item>
····<item>gāb·yigīb</item>
····<item>ḥabb·yiḥibb</item>
····<item>xalla·yixalli</item>
···</list>
··</div>
·</text>
</TEI>
```

Figure 3: List of inflection bases

On the basis of these two verb forms almost all other verb forms can be created. This step is accomplished by a PYTHON script which reads the above list and creates a list of all possible word forms. There are only very few irregularities in the verbal inflection which can be handled in few lines of additional programming code.



```
<w·lemma='ṛāḥ'·ana='#past_pl_p1'>ṛuḥna</w>
<w·lemma='ṛāḥ'·ana='#past_pl_p2'>ṛuḥtu</w>
<w·lemma='ṛāḥ'·ana='#past_pl_p3'>ṛāḥu</w>
<w·lemma='ṛāḥ'·ana='#past_sg_p2_m'>ṛuḥt</w>
<w·lemma='ṛāḥ'·ana='#pres_sg_p1'>baṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#past_sg_p2_f'>ṛuḥti</w>
<w·lemma='ṛāḥ'·ana='#past_sg_p1'>ṛuḥt</w>
<w·lemma='ṛāḥ'·ana='#subj_sg_p2_f'>tiṛūḥi</w>
<w·lemma='ṛāḥ'·ana='#subj_sg_p3_m'>yiṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#subj_sg_p3_f'>tiṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#subj_sg_p2_m'>tiṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#past_sg_p3_m'>ṛāḥ</w>
<w·lemma='ṛāḥ'·ana='#pres_sg_p2_m'>bitṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#pres_sg_p2_f'>bitṛūḥi</w>
<w·lemma='ṛāḥ'·ana='#past_sg_p3_f'>ṛāḥit</w>
<w·lemma='ṛāḥ'·ana='#subj_pl_p1'>niṛūḥ</w>
<w·lemma='ṛāḥ'·ana='#subj_pl_p2'>tiṛūḥu</w>
<w·lemma='ṛāḥ'·ana='#subj_pl_p3'>yiṛūḥu</w>
```

Figure 4: List of inflected verb forms

This tool is also being used in an experimental web service that is being tested on the ICLTT Language Resources Portal which offers a graphical interface to the library.

4.1.2 The other word forms

With the other word forms, we proceed analogously. The only remaining forms that need special treatment are the plurals of nouns and adjectives. However, nominal morphology displays much less complexity and consequently yields less forms. Plural forms are not automatically created but taken from the dictionary as the patterns used to form plurals are highly unpredictable. Again, overabundance plays an important role here. Plural doublets are a quite frequent phenomenon. All the other word classes are not inflected.

4.1.3 Overall list of word forms

From these two lists the word form_creator produces a single list that is made of triplets consisting of a word form (in Arabic characters), a lemma (in transcription) and a morphosyntactic label.

…	…	…
بحور	baḥr	n_pl
بيبارك	bārik	v_subj_sg_3p_m
باحاول	ḥāwil	v_biPres_sg_1p
…	…	…

Figure 5: Excerpt from the complete list

4.2 Tagging

The list of all word forms is the basis for the tagging process in which the corpus is enriched with a basic layer of annotations. While rule-based tagging has not been very popular in computer linguistics, it is by many regarded as a fast and straightforward approach when dealing with under-resourced languages. It can at least help in creating first results without putting too much manual work in it. Our tool makes heavily use of the data contained in the dictionary, creating virtually all possible word-forms of the lemmas contained in the dictionary.

However, the tool does more than mere string matching. For each token, the tagger traverses up to four cycles.

1. String matching
2. Declitisication + string matching
3. Pattern matching
4. Disambiguation of homographs

Phase 2 analysis is only performed when no results can be found in the list of word forms through string matching, when phase 1 yields no results. The tagger then applies an algorithm that attempts to deconstruct the word form. One of the major problems in analysing written Arabic is that the basic word forms contained in tokens may be 'surrounded' by a number of pre- and/or suffixes which in orthography are joined to the word. The process of peeling away these outer layers is called decliticisation. Consider the following token:

وبكتبها
wi-bi-kutub-ha
and-with-books-her
(=and with her books)

Figure 6: Clitics (a)

The program attempts to remove one clitic after the other and tries to find a match in the list of word forms. To make things more complicated, the above token could also be read as a verb form.

وبكتبها
wi-baktib-ha
and-I write-it/them

Figure 7: Clitics (b)

If 1 and 2 do not produce any results, a pattern matching algorithm is applied. The tagger tries to identify the most probable morphological pattern. In the last cycle, the tagger attempts to disambiguate homographs on the basis of statistical heuristics.

4.3 Evaluation of results

After the tagging is done, the tagger produces two output files: the corpus is rewritten in a one-token-per-line manner together with the POS and lemma information. In addition the program also writes a complete statistical analysis of the annotated tokens resp. lemmas. Both result sets can be saved as text or XML.

4.4 Closing the circle

The statistical analysis serves two main purposes. First of all, it is needed to find the next lexical items to be added to the dictionary and consequently to enhance and improve the dictionary. This in turn results into the improvement of the tagger which in the subsequent round will have more word forms to perform the analysis.

At the time of writing this report, the cycle has been run 12 consecutive times. The table below shows that the increase of 541 entries in the dictionary has resulted in an increase by 17% of recognised tokens.

Entries	Generated wordforms	Recognized tokens
1822	12991	48,50%
1829	12729	52,90%
1877	12800	56,60%
2012	13288	62,14%
2058	13378	62,66%
2149	13485	62,80%
2162	13505	63,59%
2356	13645	65,84%

Figure 8: First 12 refinement cycles

In many cases manual intervention is necessary. The many instances of homonyms cannot be assigned values automatically in most cases. Unsurprisingly, there are some very important items among the most frequent lemmas.

After the corpus was tagged, the most frequent lexical items not recognised by the tagger were added to the dictionary making use of a special interface built into the dictionary editor VLE, which allows to speed up the creation of new dictionary entries on the basis of the word forms proposed by the tagger for incorporation into the corpus.

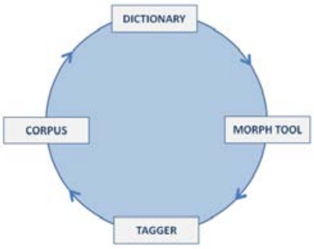


Figure 9: Workflow

5 Dictionary encoding

The statistical information the tagger creates can of course be used to different ends. The main goal of this project was the enhancement of the Egyptian dictionary with statistical information. One particular problem in this undertaking was the modelling of the encoding for this data.

5.1 Standards

In the preparatory phase of the dictionary project, many considerations concerning the encoding of the dictionaries were taken into account. Our lexicographers surveyed and tested various data formats in use. In spite of the comparatively small community of dictionary makers, there is little consensus on standards and many proprietary solutions. A great number of divergent formats have coexisted so far: Multilex and Genelex (GENEric LEXicon) are systems associated with the Expert Advisory Group on Language Engineering Standards (EAGLES). Other formats used in digital lexicography are OLIF (Open Lexicon Interchange Format), MILE (Multilingual ISLE Lexical Entry), LIFT (Lexicon Interchange Format), ISO 1951 (“Presentation/representation of entries in dictionaries – requirements, recommendations and information”) and OWL (Web Ontology Language). Without going into the details which have been discussed before (Budin et al. 2012), the final short list contained two candidates: LMF (Lexical Markup Framework) and TEI (Text Encoding Initiative).

5.1.1 LMF

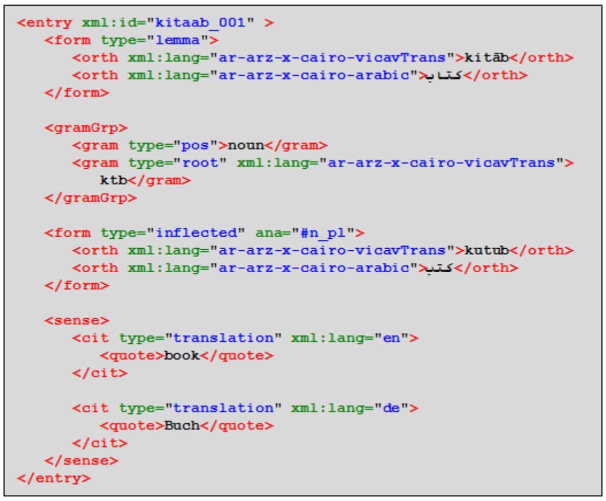
The ISO norm 24613:2008 has the full title *Language resource management – Lexical markup framework (LMF)* and is a standard for natural language processing (NLP) and machine-readable dictionaries (MRD). LMF was designed to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

5.1.2 TEI Dictionary Module

The second candidate on the short list was the dictionary module of the Text Encoding Initiative which is the de-facto encoding standard for dictionaries digitized from print sources. As such, “TEI for dictionaries” has a longstanding tradition. While LMF has been experimented with in some of the ICLTT’s dictionary projects, the current projects, in particular those working on the VICAV dictionaries, make use of a customisation of the TEI P5 dictionary module. The ICLTT’s dictionary schema is meant to be a multi-purpose tool targeting both human users and software applications. In the end, several arguments tipped the balance in favour of TEI.

Most of the ICLTT’s encoding in other fields of research is being done in TEI. The scholars and lexicographers are well conversant with the basic set of rules of the TEI system. There are many examples of TEI dictionaries and successful projects making use of the TEI. TEI encoding is easily customizable and there exists an active and enthusiastic community that assists in discussions whenever problems arise.

The ICLTT's scholars have used a narrowly defined schema in their lexicographic projects that imposes a number of specific constraints which was meant as a mechanism to enhance interoperability. A typical, slightly simplified entry taken from the Egyptian dictionary is shown below:



```
<entry xml:id="kitaab_001" >
   <form type="lemma">
      <orth xml:lang="ar-arz-x-cairo-vicavTrans">kitāb</orth>
      <orth xml:lang="ar-arz-x-cairo-arabic">كتاب</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="root" xml:lang="ar-arz-x-cairo-vicavTrans">
         ktb</gram>
   </gramGrp>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-arz-x-cairo-vicavTrans">kutub</orth>
      <orth xml:lang="ar-arz-x-cairo-arabic">كتب</orth>
   </form>

   <sense>
      <cit type="translation" xml:lang="en">
         <quote>book</quote>
      </cit>

      <cit type="translation" xml:lang="de">
         <quote>Buch</quote>
      </cit>
   </sense>
</entry>
```

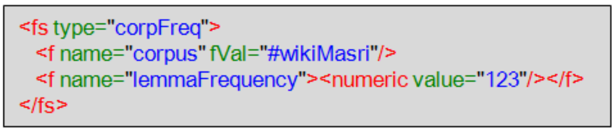
Figure 10: TEI P5 encoding of a dict. entry

5.1.3 LMF + TEI

Basically, the ICLTT strives to pursue both lines. While TEI is the prime tool for editing and production, they try to keep an eye on LMF in all lexicographic developments. For some dictionaries, they have developed XSLT styles to convert TEI output into LMF, a feature which is built into the VLE editor's export module.

5.2 Modelling the corpus queries

On looking for ways to encode statistical information inside dictionary entries we did not find examples on which we could build our decision making. TEI is very flexible, consequently there are many possible ways of solving the issue. In the particular case, feature structures were used which are a very versatile possibility to model linguistic data of any kind.



```
<fs type="corpFreq">
  <f name="corpus" fVal="#wikiMasri"/>
  <f name="lemmaFrequency"><numeric value="123"/></f>
</fs>
```

Figure 11: Feature structure for frequency data

This construct would allow to add data from other corpora and could be used in web based services to exchange such data. A number of additional items might be useful here. In addition to the corpus name, one might also think of parameters such as corpus size, mode of access, the query through which the data was retrieved or the date when the query was performed.

6 Plans for the future

In the described project the main focus was put on the integration of statistical information on word forms and lemmata with lexicographic data derived from one particular corpus. In doing so, we had a closer look at methods for creating statistical information and for integrating this data into one particular digital dictionary. While the current project has not yet been finished and results are not yet completely satisfactory, we have started to think about next steps. In the mid-term, we are planning to work along several lines.

First of all, we want to combine data drawn from several corpora into one dictionary. In addition to the Wikipedia Masri, we have started to harvest other data from the internet. While lots and lots of data is available, one of the major hurdles is to distinguish the different registers. Most of the data retrieved in this manner stems from personal websites, discussion forums and the like material which typically mix MSA Arabic and dialect. Different approaches are conceivable. One might filter on different levels, accepting only documents that display a modicum of tokens that can only appear in the dialect. Another way to get to the dialect data would be to filter on the level of paragraphs or sentences.

There is definitely a need to work on a consistent markup scheme that might be reused by other projects aiming at similar ends. The current solution based on feature structures should be regarded as a makeshift solution. There are plans to work on a comprehensive set of descriptive elements and attributes to describe frequencies of lexicographic items.

We are planning to modularise the system by setting up RESTful web-services that are capable of delivering the statistical information.

Lastly, we are planning to extend our experiments to other linguistic varieties. We assume that this approach of data-driven dictionary-based tagging can also be applied successfully to other varieties of Arabic and will help to improve other lexicographic resources.

The data and tools produced in the project are for the most part designed as Austrian contributions to the European infrastructure projects CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research infrastructure for the Arts and Humanities) and will be freely available. The TEI version of Wikipedia Masri will also be made available through the ICLTT Language Resources Portal.

7 References

Al-Sabbagh, R., Girju R. (2010). Mining the Web for the Induction of a Dialectical Arabic Lexicon. In *LREC 2010*.

Buckwalter, T., Parkinson, D. B. (2011). *A frequency dictionary of Arabic. Core vocabulary for learners.* London.

Budin, G., Majewski, S. & Moerth, K. (2012). Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. In *Journal of the Text Encoding Initiative (jTEI) 3*.

Davies, H. (2006). Dialect Literature. In *Encyclopedia of Arabic Language and Linguistics*, I, pp.597-604.

Duh, K., Kirchhoff, K. (2005). POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*. Ann Arbor.

Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool.

Hinds, M., Badawi, S. M. (1986). *A Dictionary of Egyptian Arabic. Arabic-English*. Beirut.

Hoogland, J., Versteegh, K. & Woidich, M. (2003). *Woordenboek Arabisch-Nederlands*. Amsterdam.

Jomier, J. (1976). *Lexique pratique français–arabe (parler du Caire)*. Cairo.

Landau, J. M. (1959). *A word count of modern Arabic prose*. New York, NY, USA.

Leuf, B., Cunningham, W. (2001): *The Wiki way. Quick Collaboration on the Web*. Boston, Mass., USA.

Moerth, K, Dorostkar, Niku & Preisinger, A. (2011). Gleaning micro-corpora from the internet: integrating heterogeneous data into existing corpus infrastructures. In *Actas del III Congreso Internacional de Lingüística de Corpus*, Valencia, pp. 111-118.

Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., Nichols, C. & Shareef, S. (2005). Parsing Arabic Dialects. Technical Report, the Johns-Hopkins University, 2005 Summer Research Workshop.

Rosenbaum, G. M. (2004). Egyptian as a written language. In *Jerusalem Studies in Arabic and Islam 29*, pp. 281–326.

Siam, O. (2013). *Ein digitales Wörterbuch der 200 häufigsten Wörter der Wikipedia in ägyptischer Umgangssprache - Corpusbasierte Methoden zur lexikalischen Analyse nichtstandardisierter Sprache.* Unpublished masters thesis. Vienna.