Laying the Foundations for a Diachronic Dictionary of Tunis Arabic: a First Glance at an Evolving New Language Resource

Karlheinz Mörth[1], Stephan Procházka[2], Ines Dallaji[2]
[1]Institute of Corpus Linguistics and Text Technology (Austrian Academy of Sciences)
[2]Department of Oriental Studies (University of Vienna)
Karlheinz.Moerth@oeaw.ac.at, stephan.prochazka@univie.ac.at, ines.dallaji@univie.ac.at

Abstract

Arabic lexicography has a long tradition. However, at the time of writing this report, there exist only a very few digital products, let alone products documenting Arabic dialects. Our paper presents the TUNICO project (Linguistic Dynamics in the Greater Tunis Area) and digital language resources which are being produced as part of the project. The TUNICO working group is working on a digital diachronic dictionary of Tunis Arabic which is being compiled as part of a larger linguistic endeavour to document the variety of the Tunisian capital. One of the interesting features of the project is that it draws on a number of heterogeneous sources: text books, grammatical descriptions and a corpus of spoken youth language which is currently being compiled. In this project, the dictionary is used as an analytical tool, as a research instrument, by integrating the various sources into one new coherent language resource thus allowing researchers to gain unprecedented insights in material that partly has been available for quite some time.

Keywords: eLexicgraphy; diachronic lexicography; lexicography tools

1 Introduction

The compilation of the diachronic dictionary of Tunis Arabic has been started as part of a larger project investigating the linguistic dynamics caused by recent demographic and social changes in the metropolitan area of Tunis (hence Tunis Arabic and not Tunisian Arabic). The TUNICO project (funded by a three year grant of the *Austrian Science Fund*[1]) will produce two digital language resources: (a) a corpus of unmonitored speech (dialogues as well as narratives) and (b) a dictionary based on this corpus and on other historical sources published in print form.

TUNICO in turn has grown out of an ongoing cooperative project which goes by the name *Vienna Corpus of Arabic Varieties* (VICAV). VICAV has been already started several years ago and is being run with a twofold perspective in mind: proceeding from linguistic research questions VICAV has been designed

[1] Project number P 25706-G23

as a forum for collecting, producing and making available digital language resources of a wide range of spoken Arabic varieties. In addition, the project also pursues text technological interests investigating relevant standards, developing tools and workflows. At the heart of the VICAV collection there are so-called language profiles. This type of text consists in concise sketches of spoken linguistic varieties. The intention has been to proceed in a complementary manner to similar endeavours (such as the *Encyclopedia of Arabic Language and Literature* which is a standard reference work in the field). For the time being, the concept does not foresee the production of detailed grammatical descriptions. The focus is rather put on general information, research histories, relevant literature and sample texts. Another language resource represented in the collection are lists of salient grammatical features. The working group has attempted to identify particular linguistic items that are repeatedly used elsewhere in comparative Arabic investigations and make them comparable in example sentences that are the same across the various linguistic varieties. There are also digital dictionaries, texts, bibliographies, descriptions of relevant workflows and best-practises such as thorough encoding guidelines that can be reused for similar purposes in other projects. VICAV is intended as a cross-disciplinary platform for researchers in the field enabling them to exchange data, to collaborate effectively on new digital resources and to publish their findings, tools and data.

Both projects, TUNICO and VICAV, are joint initiatives of the University of Vienna (Department of Oriental Studies) and the Austrian Academy of Sciences (Institute for Corpus Linguistics and Text Technology). They are typical examples of a new brand of research that understands itself as digital humanities pursuing research with innovative methods and in accordance with new paradigms such as collaborative work, transdisciplinarity and open humanities.

2 A New Digital Dictionary

In the history of lexicography, dictionaries documenting Arabic dialects are a rather recent phenomenon. While the situation with respect to print dictionaries has improved for many areas, there are almost no digital Arabic dictionaries available so far, let alone dictionaries that come in a digitally reusable form, live up to modern standards or cover varieties other than Modern Standard Arabic.

With respect to Tunis, the situation is no different. There exists no comprehensive dictionary of the Arabic dialect of Tunis. Nicolas 1911 can be regarded as a good basis for diachronic research. However, it is – by and large – outdated. Other sources for lexicographic data are the works of Beaussier/Lentin (2006, a fusion of the 1958 edition and the 1959 supplement) which also include data on Tunis, Quéméneur (1961 and 1962) who provided useful lists of lexicographical items and Abdellatif 2010 who produced a quite useful amateur glossary. The eight volume glossary compiled by Marçais/Guîga 1958-61 covering the vocabulary of the village of Takroûna (ca.100 km south-east of Tunis) is still of unmatched value for the documentation of the lexicon of the Arabic vernacular of the Tunis area. However, it reflects mainly rural speech and is based on material from the 1920s.

Our project was designed to create up-to-date and easily accessible lexical information on Tunis Arabic, taking into account historical as well as contemporary data, by compiling a small, micro-diachronic and machine-readable dictionary of the variety. One of the many advantages of such a machine-readable dictionary is that queries in both directions (in our case Tunis Arabic – English/German/ French and vice versa) are possible. All hitherto published dictionaries except the outdated work by Nicolas (1911) are unidirectional Arabic – French.

3 Heterogeneous Sources

One innovative aspect of the project lies in the fact that it is not only drawing on contemporary data taken from a digital corpus. However, it will also incorporate various sources reaching back as far as the 19th century (Stumme 1893/1896a-b). By integrating both corpus data and historical sources, we will create a new language resource, a new dictionary. Technically, the intention has been to keep each bit of information added to the dictionary traceable to its origin, thus allowing coming generations of researchers to interpret the data in accordance with their particular needs.

The basis of the dictionary was laid by data taken from didactic materials that were compiled by participants in the project for university classes. The glossaries of this course of spoken Tunisian Arabic could be easily recycled for the purpose and transformed into digital dictionary entries. In the next phases of the project, this data will be enriched from three main additional sources: the newly created corpus, interviews with first language speakers, and historical publications on the linguistic variety under investigation.

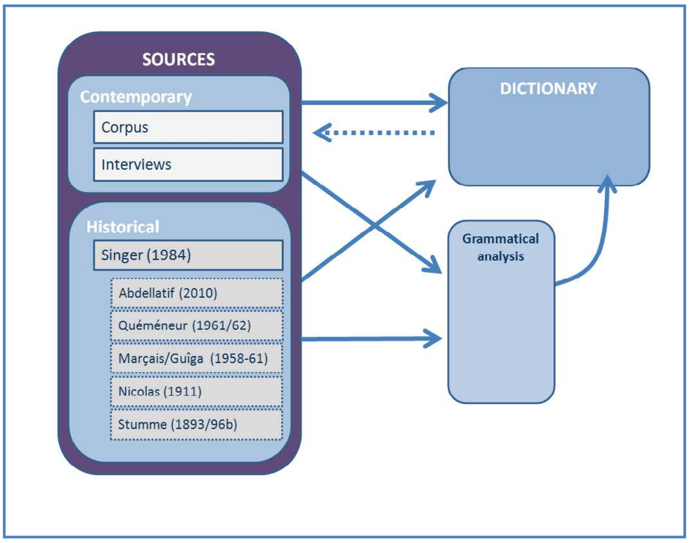


Figure 1: Basic chart of used sources.

3.1 Corpus of Spoken Youth Language

As can be seen in figure 1, the contemporary sources consist in a corpus of youth language by which we understand language produced by people under the age of 25. In Arabic dialectology, focussing on the language of the younger generation is a rather uncommon approach as traditionally linguistic interests were usually directed towards the past. Dialectological investigations often been focussed on the language of the older generation attempting to gain knowledge on older more conservative "pure" linguistic varieties. This has led to a situation (and not only in Arabic studies) where we often only know older forms of particular varieties and comparatively little about contemporary language. TUNICO (and also other language resources of the VICAV collection) has been designed to remedy this shortfall focussing on modern language, contemporary usage and lexical neologisms which, in the case of Tunis, are most often not of Arabic but French (or English) origin.

A first set of data was collected in August and September 2013 and is currently being transcribed. The field workers recorded some 33 hours of Tunis Arabic, the recordings contain data from approximately 90 different interviewees. The final version of the dictionary will contain the most frequent lemmas represented in the transcribed corpus, which will constitute the foundation of the dictionary's contemporary layer.

3.2 Additional Interviews

The second contemporary source, the interviews, will be created at a later stage of the project. As we expect numerous lacunae in the lemma list and constituents of the dictionary entries, the data gained from the corpus and also the historical sources will have to be completed with information gained from additional interviews conducted with Tunisian informants. This will be done during the third and the last field campaigns. The interviews will differ from those done in the first two campaigns as they will be semi-structured interviews that aim at the elicitation of lexical data absent in the corpus. Having collected plenty of dialogues, it is planned to also go for narratives in this phase of the project.

3.3 Historical Data

The historical aspect will be introduced by way of lexicographic items excerpted from print publications, especially the very rich lexical material contained in Singer's monumental grammar (1984; almost 800 pages) of the Medina of Tunis, which hitherto has been difficult to use which is mainly due to a lack of an index[2]. Singer's data will be evaluated systematically and integrated into the dictionary (the material will, of course, be indicated by reference to his book; however on account of the unclear

[2] It is important to note that Singer's study is based on fieldwork carried out in the early 1960s (Singer 1984: VIII); the texts and the glossary which were advertised in the foreword (p.X) have never been published.

copyright status it is not planned to create a digital version of the book itself). Additional resources (Nicolas 1911, Marçais/Guîga 1958-61, Quéméneur 1962, Abdellatif 2010) will also be consulted in order to verify and complete the collected data. The diachronic dimension will help to better understand processes in the development of the lexicon.

The rich material gathered from young people whose parents are often not natives of the city of Tunis but have migrated to the capital from rural regions or other cities of Tunisia will hopefully enable us to analyse recent developments in the lexicon, particularly semantic changes including semantic shift, semantic reduction, and semantic extension of lexical items. Our diachronic approach will also make possible to determine the influence of other dialectal varieties of Arabic which in many cases are a result of the pan-Arabic satellite channels. One focus of the interviews carried out during our fieldwork is to gain newly coined vocabulary that appeared during and after the revolution of 2011, which had an immense impact on Tunisian society and hence also on Tunisian language. This vocabulary is, however, different from real youth slang that is often only used in in-group conversation. Studies on this particular field of the lexicon of Arabic dialects are extremely rare (the best publication on this topic is Caubet 2004 dealing mostly with Morocco).

As we are not mainly interested in the "pure and real" dialect we will also pay attention to the incorporation of foreign elements into the Arabic language as spoken in Tunis both with regard to semantics and morphology. The latter is characterized by a high degree of integration into the morphological frame of Arabic. Particularly verbs of foreign origin have to be adapted to Arabic patterns for the sake of inflection. A similar development is often found with pluralisation of nouns and adjectives.

The direct connection of corpus and dictionary (see figure 1) will facilitate research on phrases, idioms and collocations. Apart from some very well-studied varieties of Arabic (especially Egyptian and Moroccan Arabic) phraseology and related subjects such as collocation have been largely neglected, mainly because of a lack of text corpora sizeable enough for these purposes. The linkage of dictionary and corpus will enable users to investigate in which way given lexical items are connected to one another.

4 Modelling the Dictionary Entries

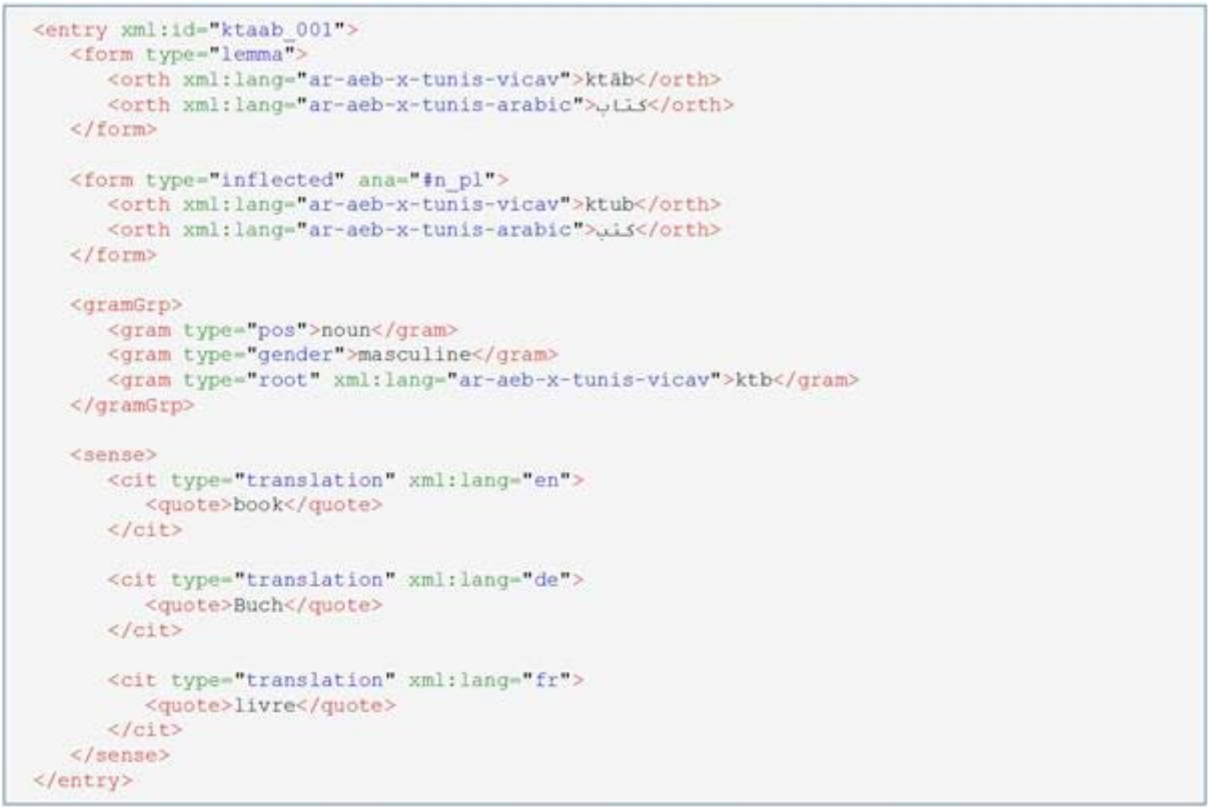
In the world of digital dictionary production, a considerable number of competing formats co-exist. We are far from any real standardisation in the field and our paper will not resume the discussion as to which format is best suited for which task (cf. Budin et al. 2012). Let it suffice to state, that using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. However, it has been shown that the TEI dictionary module is also usable for NLP purposes (Budin 2012). Data modelling for our project has been undertaken with two perspectives in mind: (a) to achieve a high degree of interoperability with comparable dictionaries of other varieties of Arabic (already existing at the same department) and (b) to stay as compliant as possible with the ISO standard LMF (Romary 2013).

A major issue in this endeavour is how to represent and how to harmonise the diverging systems of transcription and transliteration found in the historical sources. As in comparable other projects, researchers in our project try to reduce the rich inventory of combinations of diacritics and characters by applying a basically phonemic transcription. Following a widespread convention in Arabic dialectology, the data is presented in a broad phonological transcription that does not usually indicate allophones. Basically, the set of characters used in the dictionary follows widely used conventions in Arabic studies. It is by and large the system used in standard reference works such as the *Encyclopedia of Arabic Language and Linguistics* (Leiden: Brill, 2006-2009). In the future, it is planned to provide the data in IPA-transcription too.

As can be seen in figure 2 we indicate the so-called root for each lexical item in the dictionary. The root is an intrinsic feature of Arabic word formation. In all layers of Arabic the bulk of the vocabulary is built on the principle of root and pattern. To express certain semantic terms, i.e. words, a purely consonantal root carrying the basic semantic information is combined with a limited set of patterns utilizing a fixed sequence of consonants, vowels, and optional prefixes and suffixes. To make comparative cross-dialectal search possible we have decided to indicate the root in a strictly etymological way. This means, each root reflects the corresponding root of Standard Arabic wherever possible. We are convinced that this approach does not reduce the usability of the on-line dictionary because, for users familiar with Arabic morphology, it is easy to detect the dialectal root in question. The main advantage of this approach lies in the possibility to find the reflexes of a certain Standard Arabic root in all dictionaries simultaneously.

4.1 Basic Schema

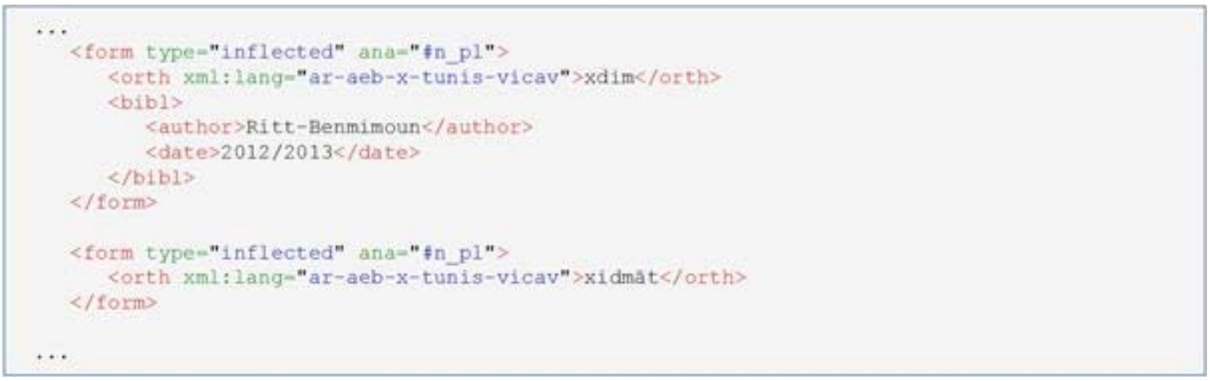
The schema applied in the compilation of the new dictionary has been used before in other projects for various languages and serves as the structural foundation of the dictionary entries. It imposes a number of very strict structural constraints on the TEI elements to ensure a high degree of interoperability with other components of the existing dictionary infrastructure at the ICLTT (Budin 2012). These constraints are defined by means of an XML Schema which only allows the use of a small subset of TEI elements and only a very few combinations thereof. The typical, slightly simplified basic structure of an entry taken from the Tunis dictionary is shown below. The entry begins with the lemma, this is followed by morphological forms and grammatical information. The system provides for translations in several languages. In the Tunis dictionary, we intend to offer German and English translations. Resources permitting, we will also add French.



```
<entry xml:id="ktaab_001">
   <form type="lemma">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktāb</orth>
      <orth xml:lang="ar-aeb-x-tunis-arabic">كتاب</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
      <orth xml:lang="ar-aeb-x-tunis-arabic">كتب</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="gender">masculine</gram>
      <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">ktb</gram>
   </gramGrp>

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

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

      <cit type="translation" xml:lang="fr">
         <quote>livre</quote>
      </cit>
   </sense>
</entry>
```

Figure 2: Basic encoding of a typical dictionary entry.

As can be seen in the example above, *sense* elements can have multiple translations. In a similar manner, every *entry* can contain an unspecified number of form elements. These can represent different morphological forms, variants such as for instance competing plurals or varying phonological representations. All these cases are treated similarly. Hierarchies are avoided, all *form* elements are placed directly inside the *entry* element.



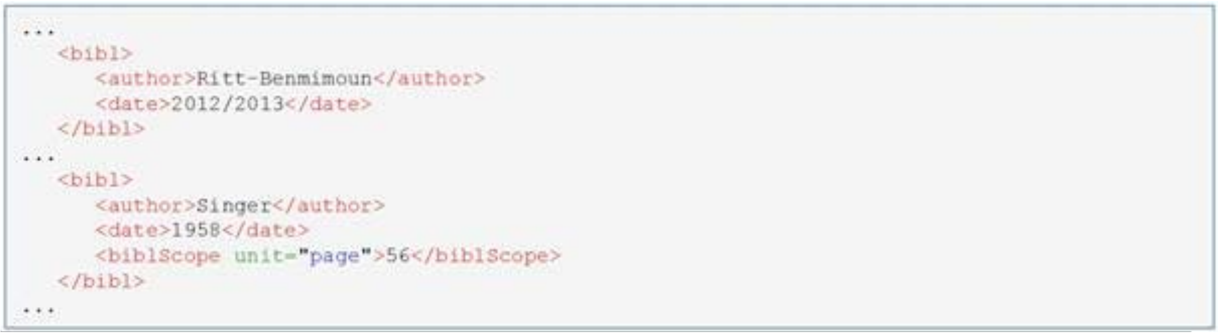
```
...
   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">xdim</orth>
      <bibl>
         <author>Ritt-Benmimoun</author>
         <date>2012/2013</date>
      </bibl>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">xidmāt</orth>
   </form>

...
```

Figure 3: Overabundance in plural forms.

The two inflected forms represent both common plurals. In cases of a clear distribution across registers, labels can be used to assign information regarding the register of the particular form. However, for the time being such forms are merely collected without adding information concerning the formality scale. Once the corpus is available, we intend to add frequency data to the lemmas and the inflected forms. As to the encoding of the frequency information, discussions concerning data modelling are still ongoing.[3]

Modelling Diachrony

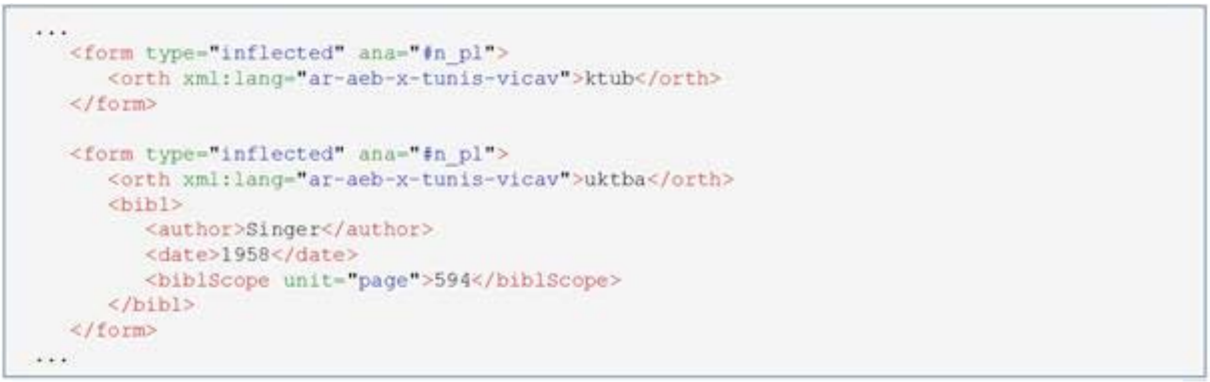
Diachrony, or as some might insist micro-diachrony (as we are only talking about a time-span of roughly a century), is represented in the dictionary by indicating the source from which the data was taken. To this end, we make use of the *bibl* (bibliographic citation) element. This is a loosely-structured element the sub-components of which may or may not be explicitly tagged (TEI Guidelines 2013).



```
...
   <bibl>
      <author>Ritt-Benmimoun</author>
      <date>2012/2013</date>
   </bibl>
...
   <bibl>
      <author>Singer</author>
      <date>1958</date>
      <biblScope unit="page">56</biblScope>
   </bibl>
...
```

Figure 4: Bibliographic citations in TEI (P5).

Diachrony is established by adding these *bibl* elements to *form* and/or *sense/cit* elements. As stated before, any entry can have multiple forms and also can have multiple instances of the same morphological form. The absence of a *bibl* element indicates that the form has been entered from contemporary sources. In this manner, each element can be historically classified. In the following example, the *entry*, a noun, has two plural forms. By contrast to the example above which displays synchronous data (*xdim* vs. *xidmāt* are both still in use) the second form here represents evidence of a historic form. An analogous contemporary form could so far not be verified.



```
...
   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">uktba</orth>
      <bibl>
         <author>Singer</author>
         <date>1958</date>
         <biblScope unit="page">594</biblScope>
      </bibl>
   </form>
...
```

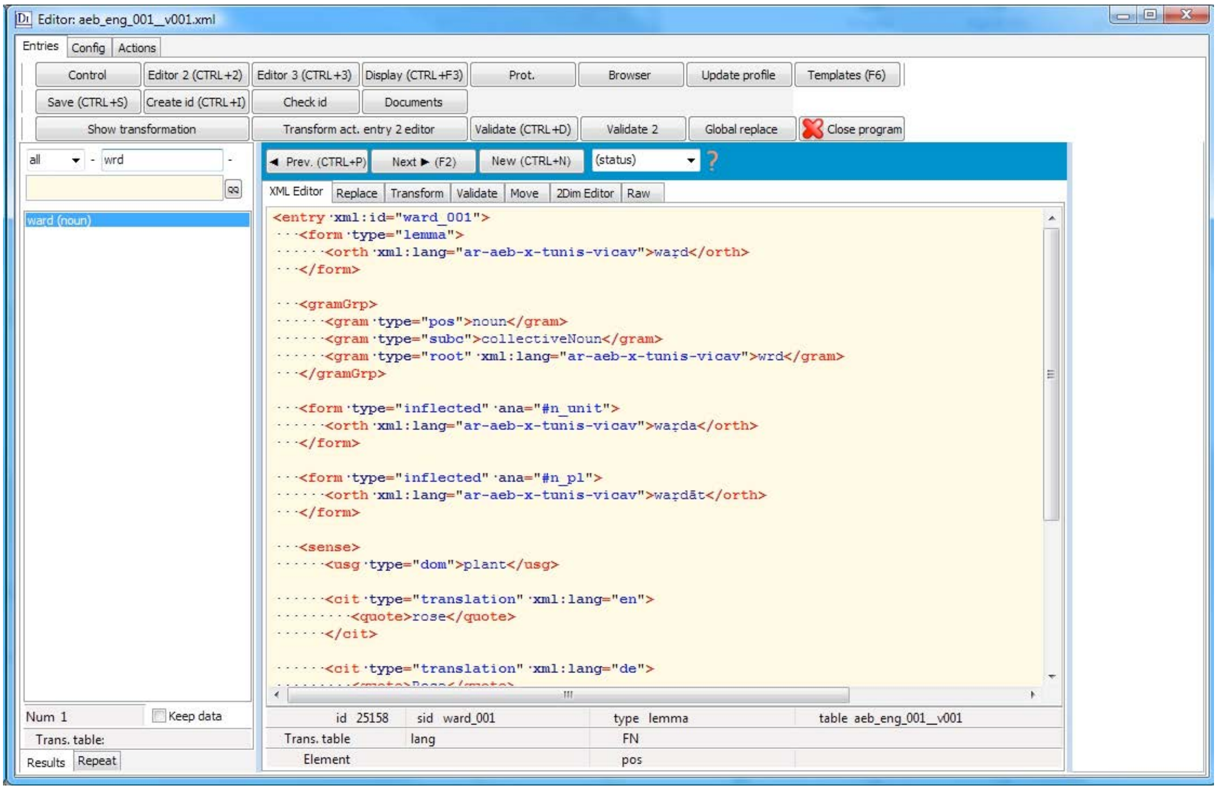
Figure 5: *ktub* vs. *uktba*.

[3] Details of these discussions were presented in the TEI members Meeting 2014 (Rome) and have been submitted for publication in jTEI 8 (papers from the 2013 conference).

5 Tools

The dictionary entries are compiled making use of the *Viennese Lexicographic Editor* (VLE), a general purpose XML editor providing a number of functionalities typically needed in compiling lexicographic data. It allows to collaboratively work on lexicographic data. From the very beginning of its development, it was designed to process standard-based lexicographic and terminological data such as LMF, TBX, RDF or TEI. VLE can automate many editing procedures. Most of these functions can be applied both to single and/or multiple entries. VLE has been implemented as a standalone desktop application (for Windows). VLE is the client of a server-client architecture. In order to realise a working environment, a web-server is needed. The server-side scripts (*php* + *mysql*) are also freely available and easy to setup. The program can check the structural integrity (well-formedness) of input on the fly and can validate the data against XML Schemas.

One of the particular features of VLE is a special module optimising access to external language resources such as corpora, other dictionaries, word lists etc. which makes it particularly well suited for deployment in our project with its various digital resources.[4] The fact that VLE is a product of our institute and constantly being updated will ease the implementation of necessary interfaces between the corpus and the dictionary infrastructures.



```
<entry xml:id="ward_001">
   <form type="lemma">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛd</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="subc">collectiveNoun</gram>
      <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">wrd</gram>
   </gramGrp>

   <form type="inflected" ana="#n_unit">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛda</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛdāt</orth>
   </form>

   <sense>
      <usg type="dom">plant</usg>

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

      <cit type="translation" xml:lang="de">
         <quote>Rose</quote>
```

Num 1
Trans. table:

id 25158	sid ward_001	type lemma	table aeb_eng_001__v001
Trans. table	lang	FN	
Element		pos	

Figure 6: The dictionary editor.

[4] The tools can be downloaded from the *Language Resources Portal* (*CLARIN Centre Vienna*): http://clarin.oeaw.ac.at/ccv/vle.

The online publication of both the corpus and the dictionary will be undertaken by means of *corpus_shell*, a modular framework of reusable software components which has also been developed at the ICLTT over the past couple of years. It is used to access and publish heterogeneous and distributed language resources such as language corpora, dictionaries, encyclopaedic databases, prosopographic databases, bibliographies, metadata, and schemata. Its core functionality is encapsulated in self-contained components exposing well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components.[5] This software was used in several other projects before, it is the backbone of the *Language Resources Portal*[6] which is run at the Austrian Academy of Sciences.

6 Status and Outlook

The project is being conducted in the context of CLARIN-AT, the local branch of the European infrastructure consortium CLARIN-ERIC (Common Language Resources and Technology Infrastructure). Both the corpus and the dictionary were planned as in-kind contributions to the CLARIN network for the years to come. The build-up of the corpus, the compilation of the dictionary and the development of software are being undertaken in the spirit of open-access and open-source. So far, no binding decision has been made as to the licence under which the particular language resources will be available. However, there is a strong case for a Creative Commons licence, CC-BY being the favoured option, which has been used for comparable other projects of the department. Discussions with interested researchers and other stakeholders have shown that the permission to create derivative works is usually regarded as an important prerequisite in order to ensure reuse of data. The tools, workflows and specifications created in this project can potentially also be used for other languages and many other applications.

At the time of writing this paper, the dictionary already contained roughly five thousand raw entries and several hundred edited entries. We are planning to make data available as soon as data from the corpus and historical sources have been added to the basic entries, i.e. already during the production phase which is meant to allow and to encourage other researchers in the field to contribute to this work. To our knowledge, this lexicographic undertaking is not only the first scholarly attempt to make available a digital dictionary of a spoken Arabic variety, but also the first attempt at creating a digital dictionary presenting diachronic data of a spoken Arabic variety.

[5] More details at clarin.oeaw.ac.at/ccv/corpus_shell.

[6] clarin.oeaw.ac.at/ccv/

7 Selected References

Abdellatif, K. (2010). Dictionnaire «le Karmous» du Tunisien. Accessed at: http://www.fichier-pdf.fr /2010/08/31/m14401m/ [06/04/2014].

Baccouche, T., Mejri, S. (2000). L'Atlas Linguistique de Tunisie: problématique phonologique. In *Revue Tunisienne de Sciences Sociales* 120, pp. 151-156.

Banski, P., Wójtowicz, B. (2009). FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries. In TEI-MM, Ann Arbor. Accessed at: http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz- TEIMM-presentation.pdf [06/04/2014].

Beaussier, M. (2006). *Dictionnaire pratique arabe-français: (arabe maghrébin); constitué du «Dictionnaire pratique arabe-français» de Marcelin Beaussier dans l'édition de Mohamed Ben Cheneb & de son «Supplément» par Albert Lentin,* Paris: Ibis Press.

Bel, N., Calzolari, N. & Monachini, M. (eds.) (1995). Common Specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B. Pisa.

Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries. In *Journal of the Text Encoding Initiative 3 (Special issue on TEI and linguistics).*

Hass, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz. Berlin, New York: W. de Gruyter.

Ide, N., Kilgarriff, A. & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In *Euralex 2000 Proceedings*, pp. 113-126.

Marçais, W. (1925). Textes arabes de Takroûna. I. Textes, Transcription et Traduction annotée. Paris.

Marçais, W., Guîga, A. (1958-61). *Textes arabes de Takroûna. II: Glossaire*. 8 vol. Paris.

Mörth, K., Budin, G. (2011). Hooking up to the corpus: the Viennese Lexicographic Editor's corpus interface. In *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011.* Bled (Slovenia), pp. 52-59.

Nicolas, A. (1911). Dictionnaire français-arabe: idiome tunisien and Dictionnaire arabe-français. Tunis.

Quéméneur, J. (1962). Glossaire de dialectal. In *IBLA* (1962), pp. 325-67.

Romary, L., Salmon-Alt, S. & Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on enhancing and using electronic dictionaries*. Coling 2004, Geneva.

Romary, L., Wegstein, W. (2012). Consistent Modelling of Heterogeneous Lexical Structures. In *Journal of the Text Encoding Initiative 3 (Special issue on TEI and linguistics).*

Romary, L. (2013). TEI and LMF crosswalks. In *Stefan Gradmann and Felix Sasaki (eds.), Digital Humanities: Wissenschaft vom Verstehen.* Humboldt Universität zu Berlin. Accessed at: http://hal.inria.fr/hal-00762664 [08/03/2014].

Singer, H. (1984). Grammatik der Arabischen Mundart der Medina von Tunis. Berlin-New York.

Sperberg-McQueen, C.M., Burnard L. & Bauman S. (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlotteville, Nancy. Accessed at: http://www.tei-c.org/Guidelines/P5/ [08/03/2014].

Stumme, H. (1893). Tunisische Märchen und Gedichte. Band I: Transcribierte Texte nebst Einleitung; Band II: Übersetzung. Leipzig.

Stumme, H. (1896a). Grammatik des tunisischen Arabisch nebst Glossar. Leipzig.

Stumme, H. (1896b). Neue tunisische Sammlungen. (Kinderlieder, Strassenlieder, Auszählreime, Rätsel, 'Arôbi's, Geschichtchen u.s.w.). Berlin (ZAOS II).

Stumme, H. (1898). Märchen und Gedichte aus der Stadt Tripolis in Nordafrika. Eine Sammlung prosaischer und poetischer Stücke im arabischen Dialekt der Stadt Tripolis, nebst Übersetzung, Skizze des Dialekts und Glossar. Leipzig.

Versteegh, K., Eid, M., Elgibali, A., Woidich, M & Zaborski, A. (eds.) (2005-2009). *Encyclopedia of Arabic Language and Linguistics*. 5 vols. Leiden, Boston: Brill