VICAV 3.0: Zooming in on Lexical Resources

Karlheinz Moerth
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
karlheinz.moerth@oeaw.ac.at

Daniel Schopper
Austrian Centre for Digital Humanities
Austrian Academy of Sciences
daniel.schopper@oeaw.ac.at

Abstract

The paper outlines the language documentation platform VICAV which pools information on spoken varieties of contemporary Arabic. It gives a general outline of VICAV's background and its scope, touches on a number of methodological issues concerning the text-technological setup and discusses conceptual questions focusing on aspects of eLexicography, in particular on practical issues dealing with digital data and tools used to build it. The paper explains in detail the involved language resources and deals with issues pertaining to standards, formats and interoperability. Ample detail is furnished on tools developed as part of VICAV's evolution, in particular the dictionary editor and the web-interface. The paper is based on the presentation given at ÖLT in December 2019 in Salzburg and has been supplemented with information on recent developments achieved in the course of 2020.

Keywords: eLexicography, standards, digital tools, lexical data

1 Preliminaries

The *Vienna Corpus of Arabic Varieties* is a platform that has been pursued for several years by researchers of the Institute of Near Eastern Studies of the University of Vienna and the Austrian Centre for Digital Humanities and Cultural Heritage of the Austrian Academy of Sciences. It was set up with two main purposes in mind: to create a virtual research platform targeting the particular needs of Arabic dialectology by creating a collection of various digital language resources which documents varieties of spoken Arabic and to serve as a test bed for experiments in text-technological methods and tools. Integrating approaches from language technology and the wider field of text-oriented digital humanities, VICAV aims to address issues of representing heterogeneous data by providing a flexible yet sustainable technical environment based on a modular data architecture. All of these efforts have been deeply rooted in the ACDH-CH's strong commitment to the activities of the European research infrastructures CLARIN-ERIC and DARIAH-EU, also aiming at more digital language resources for lesser resourced linguistic varieties.

1.1 The linguistic background

The linguistic situation in the Arabic world has been characterised by a comparatively high degree of multilingualism, ranging from bilingual to rather complex linguistic setups in which one variety mostly limited to written and formal contexts, namely Modern Standard Arabic (MSA), stands next to one or more spoken varieties. These varieties are not only remarkably different from the written standard but also vary considerably from place to place. While MSA is comparatively homogeneous throughout the Arabic speech community, the local vernaculars may differ to a degree that makes them quite incomprehensible to speakers of other varieties. This basically diglossic situation gets more complex under the influence of languages other than Arabic, which is mainly true of the western parts of the area. There, French as the local prestige variety has kept playing an important role as a means of communication among the educated classes and has been influencing especially the spoken language. In some parts, this triglossia has evolved further, combining MSA, spoken varieties of Arabic, French and Berber into complex linguistic biotopes, where all these languages coexist in functional complementarity at the same spot. Similarly, entangled scenarios can also be found in the east where MSA, Arabic vernaculars, Turkish, Kurdish and Aramaic varieties are used in overlapping spheres.

The following illustration (Figure 1) gives the translations of the MSA sentence *Māḏā turīd?* 'What do you want?' in a number of local varieties:

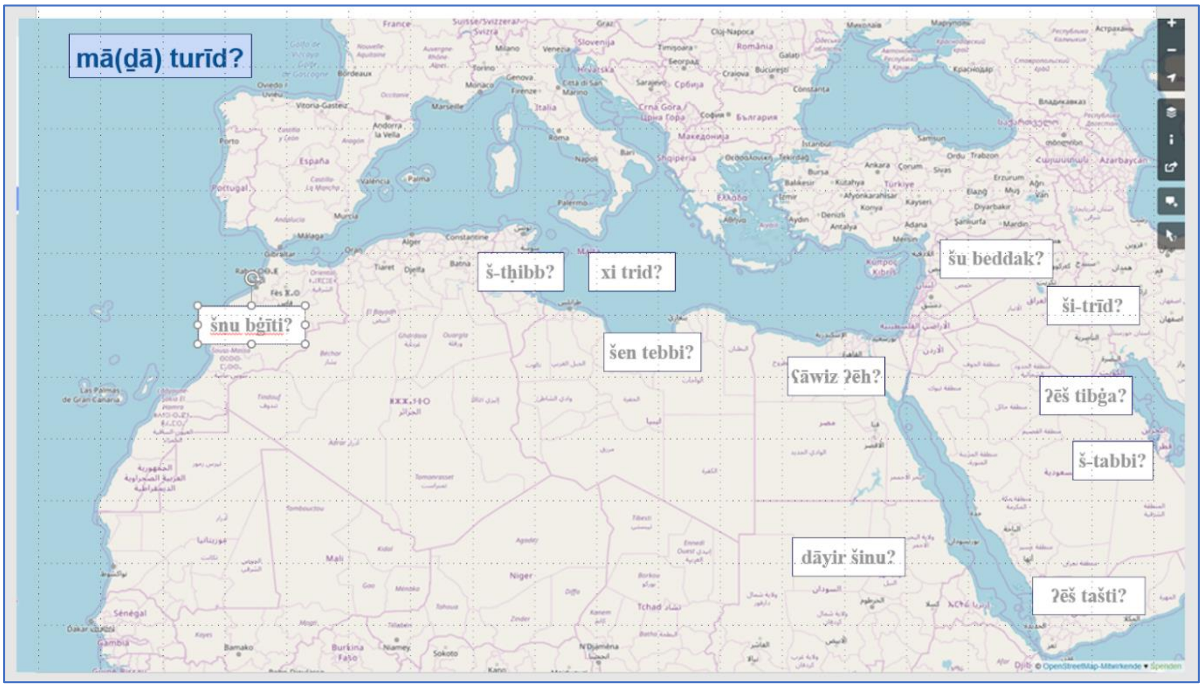


Figure 1: MSA question *Māḏā turīd?* 'What do you want?' with translations in a number of local varieties

1.2 Documenting spoken Arabic

Being located at the border between areal and corpus linguistics, the linguistic aim has been to gather a range of digital language resources for different localities. The description of the different varieties hinges on language profiles (i.e. concise and uniformly structured form sheets that offer information on the research history, available literature and other relevant information), feature lists and sample texts of particular varieties. In addition to that, VICAV also makes accessible an extensive research bibliography and dictionaries of selected locations.

In view of the distinctive digital humanities component of the undertaking, the importance of standards, best practices and open access to both data and tools have played an important role in all these efforts. VICAV was designed as a means to promote the efficient exchange of ideas and experiences as well as a technical platform for an active international community which is increasingly producing digital data but still lacks the infrastructure to make it widely available.

The ACDH-CH's research interests in developing VICAV have been related to issues of digital lexicography, visualisation of digital language resources in a multilingual environment and the application of de-facto standards in the creation of digital language resources. With respect to the linguistically focused goals, the main target groups are researchers in the field of variety linguistics (with a focus on diatopic research questions) and students of modern Arabic varieties. To make its content accessible to both expert users and students alike, the VICAV interface combines a map-based approach for intuitive data exploration with advanced text-based search tools which offer a more traditional, structured access mode.

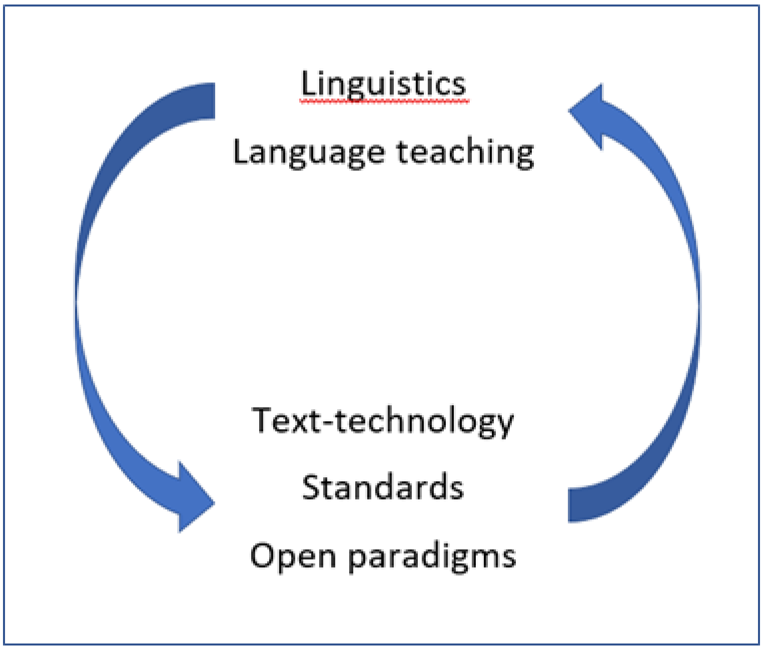


Figure 2: Interdependencies of research domains in VICAV

1.3 A modularly built network of projects

VICAV has never been conducted as a formal project but it has rather served as an umbrella for third-party funded research endeavours. Without constant financing and personal resources, the main proponents have strived to draw up projects, as part of which they could produce data and develop particular components of the required infrastructure. Data production relied largely on enthusiasts, student assistants and students financed by the Faculty of Philological and Cultural Studies of the University of Vienna. So far, three third-party funded projects could be devised and started, one of these having been successfully finished already in 2016. These projects have allowed to further develop the supporting infrastructure, and the synergies have created a lively biotope, which in the future will hopefully help to continue these activities.

1.3.1 Linguistic dynamics in the Greater Tunis Area: a corpus-based approach (TUNICO; 2013-2016; FWF P-25706)

The first one of these projects was TUNICO, which was led by Stephan Procházka. It investigated the dialect of the Tunisian capital which had been regarded as one of the best-described urban dialects of the Arab world for a long time. The first linguistic descriptions go back to the late 19th century, i.e. to an era when the scientific interest in colloquial varieties of Arabic had just begun. The majority of publications on the dialect of Tunis have focused on sociolinguistics, phonological and morphological issues. In-depth studies of the syntax have remained scarce and there was no up-to-date dictionary available that would draw on authentic spoken data. In recent decades the language of the area has undergone fundamental changes, which were caused by demographic shifts in the country, the variety of Arabic spoken by most inhabitants of the capital having become a *koiné* that has not only spread to the vicinity of the city but is widely used throughout Tunisia.

The focus of TUNICO was on contemporary language documented by recordings made with young speakers, who grew up in the city of Tunis but descended from parents, who for the most part had come to the capital from other regions. The project was designed to combine linguistic~~al~~ methods with modern text-technological approaches. As part of the project, two digital language resources were created: a corpus of unmonitored speech that contains both conversations and narratives as well as a dictionary based on this corpus and on other previously published resources. A particular point of concern was the dictionary-corpus interface. The work on this issue resulted in a web application allowing researchers to navigate from the corpus to the dictionary and vice versa. Like all VICAV projects, TUNICO was conducted in the spirit of open source and open access: the corpus and the lexicographical data of the project were made available to the scientific community through a web interface and via the ACDH-CH's long-term repository ARCHE, which enables fellow researchers to do further analyses on the material.

1.3.2 The linguistic terra incognita of Tunesia (TUNOCENT; FWF P-31647; 2019-23)

TUNOCENT, a project led by Veronika Ritt-Benmimoun of the University of Vienna aims at providing up-to-date linguistic data for the hitherto almost unknown Arabic varieties spoken in the seven Tunisian governorates of Jendouba, Beja, Kef, Siliana, Kasserine, Sidi Bouzid, and Gafsa. The varieties under investigation share a number of socio-linguistic, socio-historical, socio-economic, and topographical features. After extensive fieldwork, recording and collecting linguistic data, a corpus of transcribed and translated narrative and ethnographic texts as well as conversations is being compiled and encoded which will serve as a basis for further linguistic research. Much of the current technical development on VICAV is driven by TUNOCENT's needs. The TUNOCENT branch of VICAV, i.e. a dedicated customization of the interface, is planned to go public within 2021.

1.3.3 The Shawi-type Arabic dialects spoken in South-eastern Anatolia and the Middle Euphrates region (SHAWI; 2021-25; FWF P-33574)

While TUNOCENT is a regional continuation of the TUNICO project, SHAWI, whose PI is again Stephan Procházka, is focussing on a different corner of the Arab world. It continues research in the Bedouin element in Arabic dialects, which also occupies a central role in the TUNOCENT project. The project will produce a corpus, a grammar and a digital dictionary.

2 A wide range of data: grappling with heterogeneity

One of the central goals of VICAV has been the creation of digital language resources, which have come to play a central role in many data-based and data-driven research approaches. In VICAV, we have worked on a very broad concept of digital language resources (LRs) which is largely based on Gary F. Simons and Steven Bird's (2008, p.88) definition:

> *A language resource is any physical or digital item that is a product of language documentation, description, or development or is a tool that specifically supports the creation and use of such products.*

In our understanding, digital LRs are conceived of as a triad of (a) data, (b) tools and (c) standards resp. best practices as a means to foster interoperability and usability. Especially the last item has to be seen as a key element, since one of the main challenges of VICAV lay in the heterogeneity of the various types of data. In the following paragraphs, we provide a concise outline of the main types of text represented in VICAV.

2.1 The VICAV bibliography

At the heart of the collection, there lies a meanwhile quite sizeable digital bibliography of studies dealing with Arabic varieties, which has been compiled from different sources since the outset of VICAV. Each entry is furnished with an elaborated system of keywords developed for the particular purpose, a system that allows the bibliographic items to be easily integrated with the other types of text of the dataset. The bibliography contains research articles, as well as other language related material, such as dictionaries or textbooks. At the end of 2020, the database contained over 4684 bibliographical records.

2.2 Language profiles

The *VICAV language profiles* consist in concise descriptions of particular linguistic varieties that – by means of a harmonised structure of this type of text – are meant to facilitate comparison between them. The selection of locations did not follow a strict plan and was determined largely by the availability of young researchers interested in these varieties. The articles were intended to be written in a style accessible both to specialists from different disciplines as well as to the interested general public. Each article starts with relevant glottonyms, giving Modern Standard Arabic terms as well as the name in the particular variety, if possible. This part is followed by a general categorisation of the described variety according to a formalised typology. The taxonomy applied furnishes general information about the location (western vs. eastern), the linguistic subgroup (e.g. *gilit* vs. *qeltu*) and/or sociolinguistic categorisations such as e.g. sedentary vs. Bedouin. Further, short prosa-style texts position the varieties in a wider context offering a concise outline of the research history, a short bibliography of relevant literature and available audio data. If available, there is also a section on didactic materials such as textbooks, grammars and dictionaries. For the time being, it is not planned to give detailed grammatical descriptions. The intention is to work in a complementary manner to comparable endeavours such as the *Encyclopedia of Arabic Language and Linguistics* (Versteegh 2006-2009). Currently, VICAV offers over 80 profiles.

2.3 Linguistic features

VICAV linguistic features are lists of salient linguistic phenomena (mainly morphological and lexical phenomena) that can be used in comparisons between linguistic varieties. The list of distinctive features is to be regarded as tentative, will be further discussed in the community and contains mainly items which keep being referred to in the comparative description of Arabic varieties. We regard this as a first approximation only and hope that the feature lists will serve as a basis for future developments. The draft is designed in an extensible manner, open for any amendments, refinements and enhancements. In the main VICAV site, we currently work with a list of roughly one hundred sentences. However, the collected material can easily be regrouped according to different criteria. In principle, the plan was to furnish such lists for as many varieties as possible. The basic VICAV setup provides only 8 such lists: Ahwaz (Iran), Baghdad, Cairo, Damascus, Douz (Tunisia), Tunis and Urfa (Turkey). The TUNOCENT project, which extends the same list, has produced a much more granular picture of the area of Western Tunisia with 98 lists so far.

2.4 Sample texts

The *VICAV sample texts* are translations of a short MSA template which was composed in a manner supposed to reflect a number of relevant features for the comparison between different varieties. On the main VICAV site there are currently six such sample texts, whereas the current TUNOCENT branch currently under development already contains 161 of these texts.

2.5 Corpora

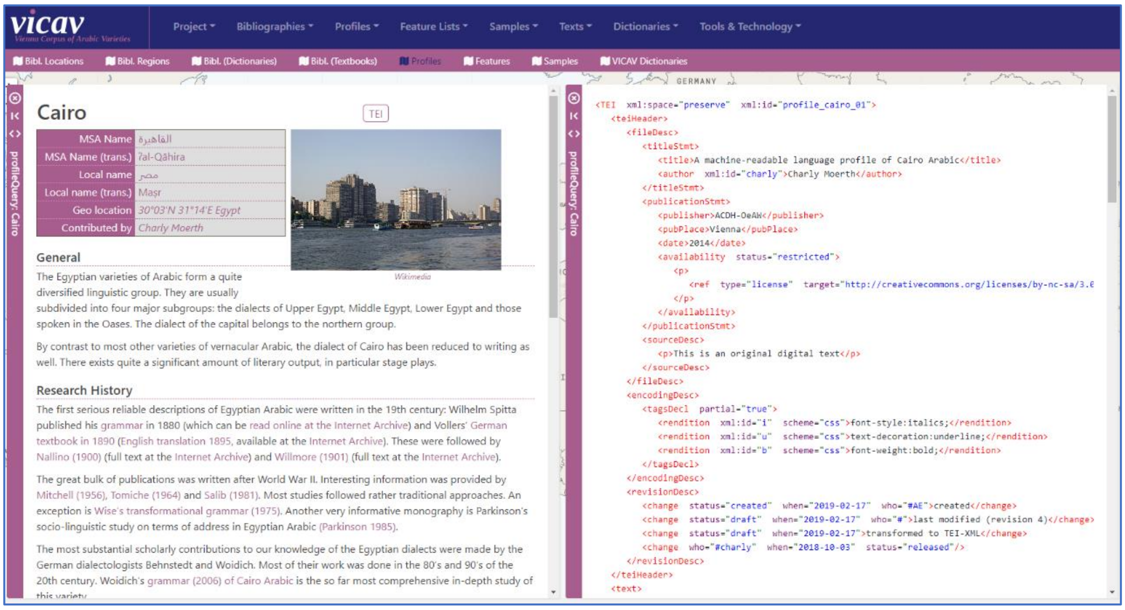
With a few exceptions, digital corpora of spoken Arabic have remained very scarce (v. Zaghouani 2014). The same is also true of VICAV, which – in spite of its name *Vienna Corpus of Arabic Varieties* – only contains a small number of actual digital texts making this part of the collection remaining to be expanded in the future. Except for one text from Morocco and one from Anatolia, there are 24 transcriptions of unmonitored speech from the TUNICO corpus, which at this moment are only accessible through the TUNICO interface (https://tunico.acdh.oeaw.ac.at/corpus.html). What makes this corpus a special resource is the fact that it was throughout interlinked with the TUNICO dictionary enabling the user to directly navigate in both directions: from the corpus to the dictionary and vice versa (Moerth et al. 2017).

2.6 Dictionaries

While the initial intention behind VICAV was to build and collect digital corpora, its main focus has shifted over the years to lexical information, including a series of digital bilingual dictionaries. Born-digital lexical data for Arabic varieties are also scarce (Moerth 2018, p.512), and VICAV has become one of the few places on the Internet to find such information. These dictionaries have been compiled with three main purposes in mind: to provide support for classes of spoken Arabic, to create a virtual research environment for comparative lexical studies, and to promote text-technological developments for lexical data.

2.7 Paratexts

An important objective of VICAV has also been to offer information about how the data were created. In this sense, VICAV can be understood as an infrastructure component that provides knowledge transfer tools and enables other researchers to accomplish similar tasks. This is why the VICAV team has tried to establish transparent and traceable workflows creating an environment that allows to reuse both data (https://github.com/acdh-oeaw/vicav-content) and software (https://github.com/acdh-oeaw/vicav-app). Part of this ‘didactic’ mission is the TEI viewer, which is integrated with most data on the website offering direct access to the TEI encoding of the respective piece of information.



Cairo

MSA Name	القاهرة
MSA Name (trans.)	ʔal-Qāhira
Local name	مصر
Local name (trans.)	Maṣr
Geo location	30°03'N 31°14'E Egypt
Contributed by	Charly Moerth

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.

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 particular stage plays.

Research History

The first serious reliable descriptions of Egyptian Arabic were written in the 19th century: Wilhelm Spitta published his grammar in 1880 (which can be read online at the Internet Archive) and Vollers' German textbook in 1890 (English translation 1895, available at the Internet Archive). These were followed by Nallino (1900) (full text at the Internet Archive) and Willmore (1901) (full text at the Internet Archive).

The great bulk of publications was written after World War II. Interesting information was provided by Mitchell (1956), Tomiche (1964) and Salib (1981). Most studies followed rather traditional approaches. An exception is Wise's transformational grammar (1975). Another very informative monography is Parkinson's socio-linguistic study on terms of address in Egyptian Arabic (Parkinson 1985).

The most substantial scholarly contributions to our knowledge of the Egyptian dialects were made by the German dialectologists Behnstedt and Woidich. Most of their work was done in the 80's and 90's of the 20th century. Woidich's grammar (2006) of Cairo Arabic is the so far most comprehensive in-depth study of this variety

```
<TEI xml:space="preserve" xml:id="profile_cairo_01">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>A machine-readable language profile of Cairo Arabic</title>
        <author xml:id="charly">Charly Moerth</author>
      </titleStmt>
      <publicationStmt>
        <publisher>ACDH-OeAW</publisher>
        <pubPlace>Vienna</pubPlace>
        <date>2014</date>
        <availability status="restricted">
          <p>
            <ref type="license" target="http://creativecommons.org/licenses/by-nc-sa/3.0
          </p>
        </availability>
      </publicationStmt>
      <sourceDesc>
        <p>This is an original digital text</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <tagsDecl partial="true">
        <rendition xml:id="i" scheme="css">font-style:italics;</rendition>
        <rendition xml:id="u" scheme="css">text-decoration:underline;</rendition>
        <rendition xml:id="b" scheme="css">font-weight:bold;</rendition>
      </tagsDecl>
    </encodingDesc>
    <revisionDesc>
      <change status="created" when="2019-02-17" who="#AE">created</change>
      <change status="draft" when="2019-02-17" who="#">last modified (revision 4)</change>
      <change status="draft" when="2019-02-17">transformed to TEI-XML</change>
      <change who="#charly" when="2018-10-03" status="released"/>
    </revisionDesc>
  </teiHeader>
  <text>
```

Figure 3: VICAV Language Profile of Cairo with parallel TEI view

The most comprehensive document that has been made available on the VICAV website is the *VICAV DICTIONARY Encoding Guidelines* (https://vicav.acdh.oeaw.ac.at/docs/vicav_dict_encoding __v003.html). In addition, the website also offers a general section, which presents an overview of digital language resources useful to Arabic dialectologists.

2.8 Standards and formats

A substantial part of the technological research undertaken in VICAV and the projects making use of its infrastructure relates to standards and workflows. In all digital endeavours, standards necessarily play a key role, in particular when several research projects are expected to operate in a common infrastructure and need to interact seamlessly. While many industry standards have been developed by bodies outside academia such as ISO, W3C, IETF, etc., lots of developmental work concerning standards, workflows, controlled vocabularies, etc. has also been driven ahead inside the research community. Furthermore, the accelerated speed of the digital turn has led to increased awareness of the importance of standards for research and consequently to the active participation of researchers in these endeavours. The significance of standards lies in two keywords: reusability and interoperability, both playing an important part in the technical agenda of our project.

Any digital research project makes use of a host of standards and norms. With respect to text-related activities, the main name to be mentioned here are the *Guidelines for Electronic Text Encoding and Interchange* of the Text Encoding Initiative (TEI Consortium 2020), which have reached a high degree of acceptance and consensus in the humanities. The use of these guidelines can – in good conscience – be regarded as best practice today. The system of the TEI has been developed over many years by a large community of practitioners. VICAV has been committed from the very beginning to using TEI P5 in encoding all types of text: the profiles, bibliographies, feature lists etc. (Budin et al. 2012).

The adoption of an adequate system of transcription has been a well-known challenge in dialectological projects spanning several areas, both in terms of dialects, disciplines and research communities. Following a widespread convention in Arabic dialectology, VICAV has adopted a system that represents the speech sounds in a broad phonological transcription, not usually indicating allophones. It corresponds by and large to the system employed in the *Encyclopedia of Arabic Language and Linguistics* (Leiden: Brill, 2006-2009). The details can be looked up in the *VICAV Dictionary Encoding Guidelines:*
(https://vicav.acdh.oeaw.ac.at/docs/vicav_dict_encoding__v003.html#characterEncoding)

As stated before, text encoding in VICAV has been built entirely on the TEI Guidelines. Although the TEI provides a vast inventory of markup covering most conceivable dimensions of how to think of a text, this richness is not needed in all projects nor desirable for special-purpose text classes as the ones in VICAV. To tackle this problem, the TEI provides a mechanism called ODD ("One Document Does it all", cf. TEI Guidelines, Chapter 22 "Documentation Elements") for narrowing down the tagset to the subset needed for a project, thus providing documentation on the data, formalising the encoding rules, which govern data curation, and easing editing and curation processes.

Given the wide range of structurally distinct language resource types in VICAV, we have defined one ODD for each of them. These ODDs define rules which TEI constructs may (or must) be used, e.g. in a language profile or a dictionary entry, so that the software for rendering and querying the data can rely on the information provided by encoders. Since the language resources are tightly interlinked, referential integrity is an important issue here (e.g. links between a language profile and the bibliography). Internal consistency plays an important role with respect to our dictionary data: in the VICAV dictionaries, entries and examples are kept separate in order to enable one example to be integrated in several entries: An ODD can ensure that such relations are consistent and broken links are immediately identified.

Controlled vocabularies, i.e. lists of words or phrases used in categorising and accessing digital data, are still an often overlooked and yet crucial component of digital humanities applications. Especially in the humanities, where concepts and interpretations of concepts play a central role, developments towards formalising and publishing this type of language resource have been remarkably slow. Many features cannot be implemented digitally without such controlled vocabularies, and while published and standardised controlled vocabularies have become an indispensable part of best practices in the natural sciences, the creation of such resources has gained pace only slowly. In many smaller disciplines, there is still a complete deplorable lack of digitally available vocabularies, not to speak of community based and agreed upon products.

VICAV has compiled two vocabularies, which are still being worked on: a list of concepts used in categorising the relevant linguistic varieties and a list of linguistic terms used in the categorisation of the feature lists and in the dictionaries. For word classes and morphological categories, the team has tried to proceed from existing standard vocabularies such as ISOcat (a data category registry supported by the ISO Technical Committee 37) and later the CLARIN Concept Registry (https://www.clarin.eu/ccr). However, a number of lacunae, such as *count plural*, *construct state*, *collective noun* and others typically used in Arabic linguistics, had to be added to the customised new list. These vocabularies have been integrated in the VICAV system as TEI feature (value) libraries.

This is actually the only part where a system other than TEI has been used of late, the data being modelled in SKOS (Simple Knowledge Organization System) and being edited with the ACDH-CH's *Vocabs Editor* (https://vocabseditor.acdh.oeaw.ac.at/). While the taxonomy of grammatical terms is still under revision, the *VICAV Taxonomy of Arabic Dialects* can be accessed via the ACDH-CH's *Vocabs* services website (https://vocabs.acdh.oeaw.ac.at/en/).

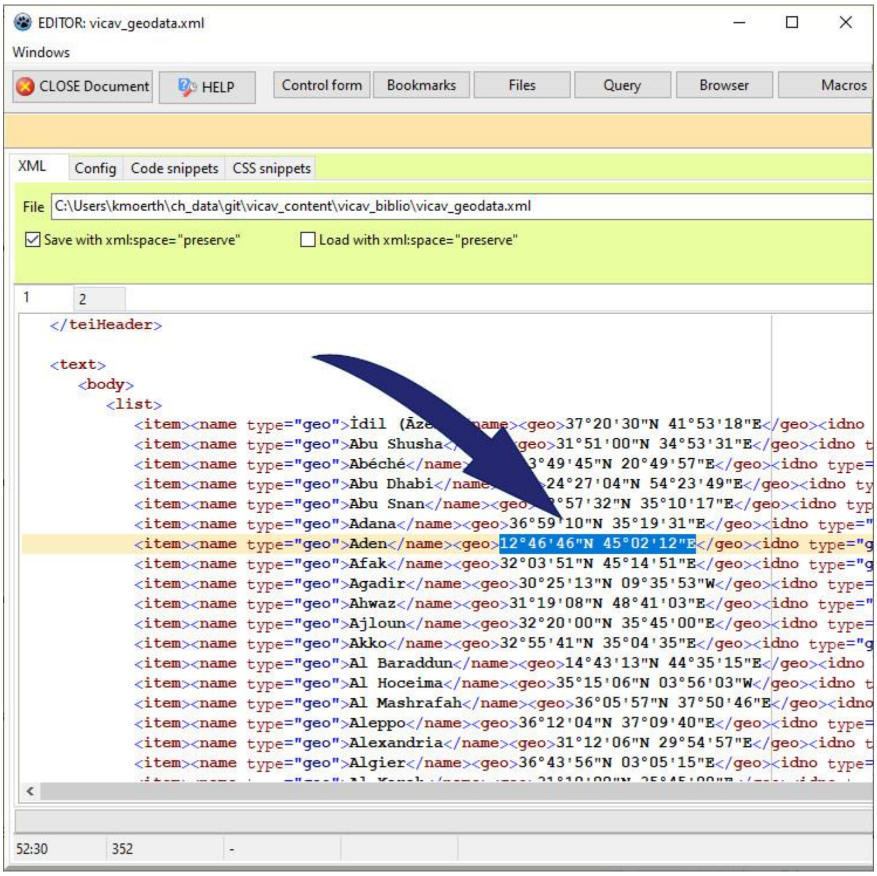
3 Experimental tool design

The issue concerning tools is highly relevant in many digital humanities projects, especially when producing data in domain niches. Standard tools – more often than not – do not provide all that is needed in such setups. In VICAV, the teams have been confronted with text-technological challenges over and over again, which ultimately resulted in developments that, step by step, have been moulded into a set of re-usable infrastructure components. In cooperation with the group of developers at the ACDH-CH, the VICAV team has been working on several digital tools.

3.1 Shaping the data creation workflow

In tackling the software issue, the VICAV team has pursued a two-pronged approach, by making use of ideally free standard software when available, and by developing their own solutions whenever existing software did not provide what was needed. Four main areas can be identified: text production, collection of bibliographical data, creation of lexical data and online publication.

Important tools that were used in workflow steps involving data creation and curation are the (commercial) XML editor oXygen (https://www.oxygenxml.com/) and the very popular and freely available bibliography tool Zotero (https://www.zotero.org/). Up until know, several other tools have been developed and used for data processing. The latest one is the *<TEI>Enricher* tool, an experimental general-purpose XML editor geared towards the easy production of TEI documents. It comes with some built-in features that ease the process of composing comparatively large text documents and to visualise them. Functionalities provided by this tool allow to work in TEI encoded documents with geo-coordinates. It has a special map-component that interfaces with the geonames API. It can be used to enrich the TEI export of the VICAV bibliography with data from a list of geolocations.



File C:\Users\kmoerth\ch_data\git\vicav_content\vicav_biblio\vicav_geodata.xml

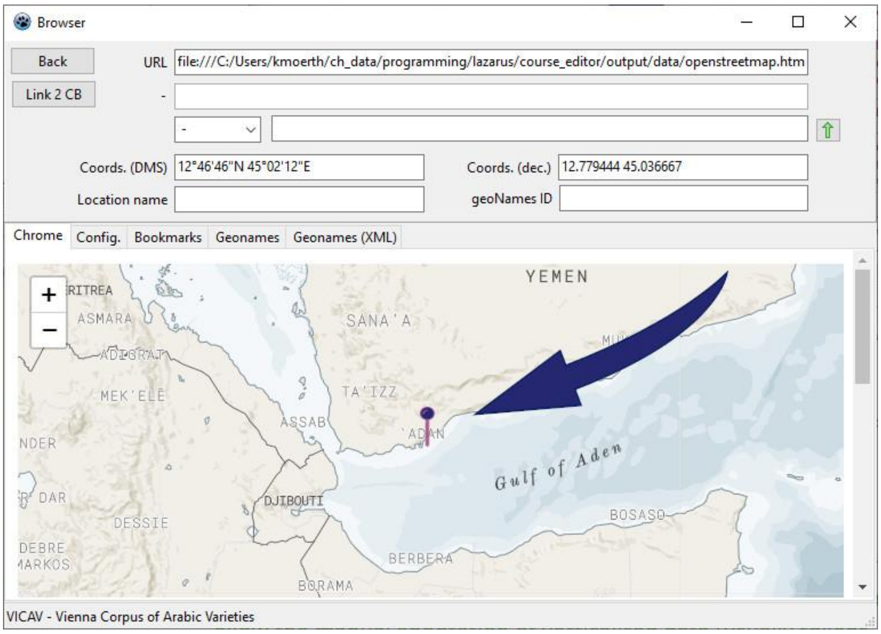
☑ Save with xml:space="preserve"   □ Load with xml:space="preserve"

```
</teiHeader>

<text>
   <body>
      <list>
         <item><name type="geo">İdil (Âze       ame><geo>37°20'30"N 41°53'18"E</geo><idno
         <item><name type="geo">Abu Shusha</    geo>31°51'00"N 34°53'31"E</geo><idno t
         <item><name type="geo">Abéché</name>   3°49'45"N 20°49'57"E</geo><idno type=
         <item><name type="geo">Abu Dhabi</name    24°27'04"N 54°23'49"E</geo><idno ty
         <item><name type="geo">Abu Snan</name><geo  °57'32"N 35°10'17"E</geo><idno typ
         <item><name type="geo">Adana</name><geo>36°59'10"N 35°19'31"E</geo><idno type="
         <item><name type="geo">Aden</name><geo>12°46'46"N 45°02'12"E</geo><idno type="g
         <item><name type="geo">Afak</name><geo>32°03'51"N 45°14'51"E</geo><idno type="g
         <item><name type="geo">Agadir</name><geo>30°25'13"N 09°35'53"W</geo><idno type=
         <item><name type="geo">Ahwaz</name><geo>31°19'08"N 48°41'03"E</geo><idno type="
         <item><name type="geo">Ajloun</name><geo>32°20'00"N 35°45'00"E</geo><idno type=
         <item><name type="geo">Akko</name><geo>32°55'41"N 35°04'35"E</geo><idno type="g
         <item><name type="geo">Al Baraddun</name><geo>14°43'13"N 44°35'15"E</geo><idno
         <item><name type="geo">Al Hoceima</name><geo>35°15'06"N 03°56'03"W</geo><idno t
         <item><name type="geo">Al Mashrafah</name><geo>36°05'57"N 37°50'46"E</geo><idno
         <item><name type="geo">Aleppo</name><geo>36°12'04"N 37°09'40"E</geo><idno type=
         <item><name type="geo">Alexandria</name><geo>31°12'06"N 29°54'57"E</geo><idno t
         <item><name type="geo">Algier</name><geo>36°43'56"N 03°05'15"E</geo><idno type=
```

52:30 352 -

Figure 4: TEI conformant list of place names with geolocations



URL file:///C:/Users/kmoerth/ch_data/programming/lazarus/course_editor/output/data/openstreetmap.htm

Coords. (DMS) 12°46'46"N 45°02'12"E   Coords. (dec.) 12.779444 45.036667

Location name   geoNames ID

VICAV - Vienna Corpus of Arabic Varieties

Figure 5: Mapping a geo-coordinate on the map with <*TEI*>*Enricher*

The <*TEI*>*Enricher* also allows to highlight geo-coordinates in TEI texts and to verify them by making use of the leaflet open-source JavaScript library for interactive maps.

Over the years, VICAV has been experimenting with tokenisers and tools to annotate tokens in digital texts. The <*TEI*>*Enricher* is the last step in these developments, implementing functionalities that allow the manual annotation of verticalised data (usually tokens, used in feature lists and texts). It also provides an interface to comfortably edit ID-based standoff data in TEI documents.

3.2 Publishing data: the dual approach

Visualising heterogeneous and complex language data for comparative purposes in an easily accessible and inuitive form has remained a conceptual and technological challenge. In the course of developing the VICAV platform, a lot of experimenting has been undertaken and the website has undergone numerous modifications so far (Procházka et al. 2015). All in all, there were three major releases: VICAV 1.0, 2.0 and 3.0, the latest one being due to supersede its predecessor on the Internet in early 2021.

One of the first requirements that were specified for VICAV was the capability of the web-interface to visualise the data on geographical maps. This feature had to be accomplished for all types of data from the bibliography via the profiles to the dictionaries. After the first clumsy attempts that made use of a static map which did not really serve the purpose, VICAV 2.0 has offered a leaflet-based map solution that has been retrieving data dynamically from BaseX (https://basex.org/), a native XML database. This version has been up and running for several years now. The main conceptual idea behind its setup was to visualise all the data in a dual manner, the basic access mode being the map, which displays the spatial relation of the underlaying data. By clicking on the symbols on the map, users are taken to respective items, which are displayed in dedicated content viewers. On the maps, two types of symbols can be found: arrows and circles. The arrows indicate particular locations (towns/cities, parts of towns, villages), the circles are used for regions that can but do not necessarily correlate with countries. In the first half of 2021, VICAV 3.0 will be launched introducing a third category of markers in form of text labels. These labels are used for large dialect regions such as Maghreb, Mesopotamia or the Gulf area.

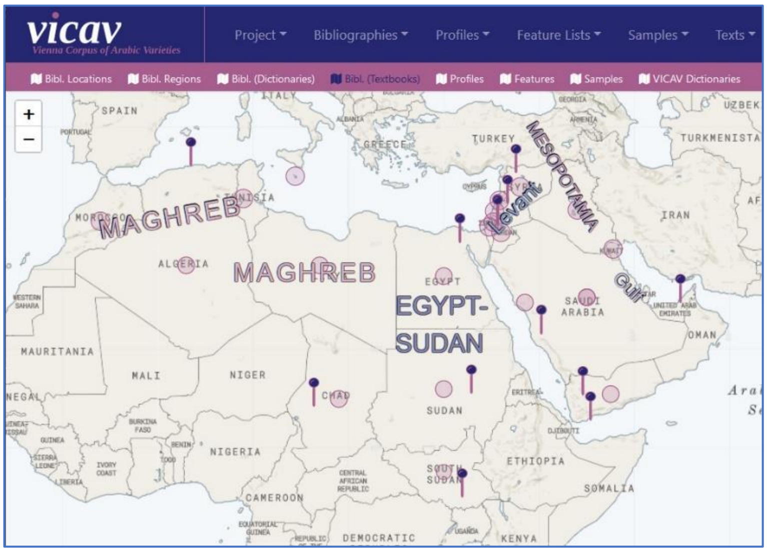
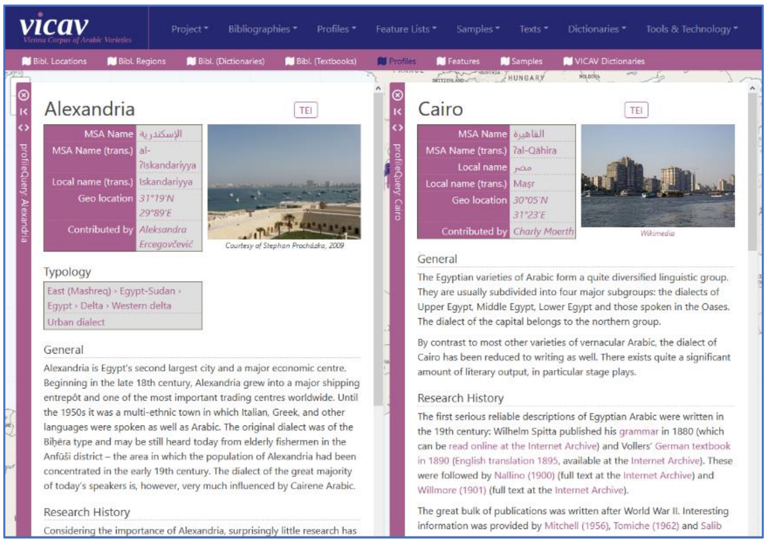


Figure 6: Visualisation of the five main dialect regions

All detailed search results are displayed in collapsible panels which is meant as a method to allow juxtaposing similar datasets and thus facilitate the comparison between linguistic phenomena across various varieties.



Alexandria

MSA Name	الإسكندرية
MSA Name (trans.)	al- ʔIskandariyya
Local name (trans.)	Iskandariyya
Geo location	*31°19'N* *29°89'E*
Contributed by	*Aleksandra Ercegovčević*

Typology

East (Mashreq) › Egypt-Sudan › Egypt › Delta › Western delta

Urban dialect

General

Alexandria is Egypt's second largest city and a major economic centre. Beginning in the late 18th century, Alexandria grew into a major shipping entrepôt and one of the most important trading centres worldwide. Until the 1950s it was a multi-ethnic town in which Italian, Greek, and other languages were spoken as well as Arabic. The original dialect was of the Biḥēra type and may be still heard today from elderly fishermen in the Anfūši district – the area in which the population of Alexandria had been concentrated in the early 19th century. The dialect of the great majority of today's speakers is, however, very much influenced by Cairene Arabic.

Research History

Considering the importance of Alexandria, surprisingly little research has

Cairo

MSA Name	القاهرة
MSA Name (trans.)	ʔal-Qāhira
Local name	مصر
Local name (trans.)	Maṣr
Geo location	*30°05'N* *31°23'E*
Contributed by	*Charly Moerth*

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.

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 particular stage plays.

Research History

The first serious reliable descriptions of Egyptian Arabic were written in the 19th century: Wilhelm Spitta published his grammar in 1880 (which can be read online at the Internet Archive) and Vollers' German textbook in 1890 (English translation 1895, available at the Internet Archive). These were followed by Nallino (1900) (full text at the Internet Archive) and Willmore (1901) (full text at the Internet Archive).

The great bulk of publications was written after World War II. Interesting information was provided by Mitchell (1956), Tomiche (1962) and Salib

Figure 7: Juxtaposing linguistic profiles

4 Zooming in on lexical data

As mentioned before, digital lexical data for Arabic have remained scarce so far, and even more so for spoken varieties of Arabic. Over the years VICAV has shifted towards lexical data and has served as an environment for the publication of several dictionaries, which are all encoded according to a unified TEI conformant schema.

4.1 The VICAV dictionaries

These dictionaries were started with three main purposes in mind: to provide support for teaching spoken Arabic, to facilitate comparison across linguistic varieties and to promote technological developments for digital lexical data. As the Department of Near Eastern Studies at the University of Vienna runs courses for several Arabic varieties on a regular basis (Morocco, Tunis, Cairo, Damascus and Baghdad), and in view of the lack of other such resources, these dictionaries have been worked out in tandem with respective teaching activities.

Although all of these lexical resources are comparatively small – none of them exceeds 8000 entries – they offer structured information that encompasses detailed lexical data. They are not simple look-up dictionaries with single sense-to-sense-relations. This is why we go for the term 'dictionary' and not 'glossary' as they constitute databases with structured lexicographic information (Moerth et al. 2014), which concerns in particular those parts of the dictionary entries providing the semantic information of the lemma.

The *Digital Dictionary of Cairo Arabic* – the first dictionary in the collection – was compiled by Karlheinz Moerth and Tarek Eltayeb. Cairo Arabic was also the first Arabic variety to be taught regularly at the University of Vienna. The initial data for the dictionary was extracted from digitally available course materials, a strategy followed also in our later dictionary projects.

The *Digital Dictionary of Damascene Arabic*, compiled by Carmen Berlinches Ramos and Stephan Procházka was also based on material taken from didactic sources, the initial input having been derived from the vocabulary lists contained in Procházka's textbook of Syrian Arabic (2014/15). Parts of the data were enhanced with audio recordings of lemmas and inflected forms, which have not been integrated into the public interface yet.

The largest dictionary in the series is the *TUNICO Dictionary*, which was compiled as part of the project *Linguistic dynamics in the Greater Tunis Area*. It was based on various sources, containing large parts of the lexical material of the corpus that was prepared in the same project, as well as data taken from interviews with young Tunisians and some diachronic information extracted from various sources published in the middle of the 20th century and earlier. For these reasons TUNICO has been described as a micro-diachronic dictionary. The most important reference work was Hans-Rudolf Singer's monumental grammar (1984; almost 800 pages) of the Medina of Tunis. Singer's data was systematically evaluated and integrated into the dictionary, all with explicit reference to the original printed edition. In order to verify and to complete the contemporary data, other resources (including Nicolas 1911, Marçais/Guîga 1958-61, Quéméneur 1962) were consulted as well. The diachronic dimension also helped to improve the understanding of processes in the development of the lexicon (for more details see Moerth et al. 2014).

The most recent one is the *VICAV Baghdad Dictionary*, which is part of VICAV 3.0. It has been worked out together with the textbook of Iraqi Arabic (Leitner et al. 2021) and has been compiled as part of the support activities of the Iraqi Arabic course at the University of Vienna.

All of the previously mentioned resources deal with varieties spoken in large urban centres (Baghdad, Cairo, Damascus and Tunis), which makes the latest one an outlier. The *Shawi Dictionary*, which will probably go online in 2021, will cover varieties spoken in the so-called Ǧazīra of north-eastern Syria and adjacent regions across the border of Turkey, particularly in the Turkish province of Şanlıurfa.

Finally, there exist several other datasets in the collection, including a *Modern Standard Arabic Dictionary*, which has also been published on the VICAV website. It is also based on course material that was developed at the Department of Near Eastern Studies of the University in Vienna. This latter one serves primarily comparative purposes, since a mid-term goal of VICAV has been the creation of an integrated lexicographic system, a *Comparative Dictionary of Arabic Varieties*, which is expected to allow researchers to query across a number of dictionaries and to obtain integrated results (Moerth et al. 2015).

All these dictionaries are bilingual or trilingual. The common denominator are the English translation equivalents, some also offering additional translations into German, French or Spanish.

4.2 Collecting lexical knowledge

The dictionaries have all been created by making use of the same technology: VLE for editing, TEI P5 for the encoding, X-technologies (XSLT, XPath, XQuery) and XHTML for visualising and publishing the data. The encoding is based on a shared schema, which has been adapted several times in accordance with necessities of the contributing projects.

The Viennese Lexicographic Editor (VLE) is an XML editor providing specialised functionalities to streamline lexicographic editing procedures. It is a standalone Windows application that allows to work collaboratively in internet-based settings. It builds on XML and cognate technologies such as XPath, XQuery, XSLT and XML Schema. While, in principle, it can process any XML-based format, it has a number of features that are geared towards the use of data, which are encoded according to the *Guidelines of the Text Encoding Initiative* (TEI P5). It was developed in several lexicographic projects[1] and has a number of specialised functions that are typically used when compiling digital dictionaries.

VLE provides several special modules which have emerged from particular needs in projects. There is, for example, an integrated 'book reader', allowing to efficiently work with books in form of scanned images and thus to navigate sources used during the dictionary compilation. Another similar tool is an integrated Internet browser that allows direct access to external sources (corpora, other

[1] Cf. also the DLGenR-project (Katsikadeli/Klampfl/Slepoy, in this volume)

dictionaries, etc.) and has been heavily used to integrate example sentences in the dictionaries. VLE visualises all lexicographic data making use of freely configurable XSLT styles. It can check the integrity of the XML data (well-formedness) and also verify the validity of the input against XML schemas. It performs versioning on the entry level storing a time-stamped copy on the server every time a record is saved. Furthermore, recent versions have a module which can be used to comfortably create a web presence for one's dictionary. While VLE was functioning for several years as part of a client-server-based architecture making use of MySQL in the backend, recent versions of the editor are used in combination with the free and easy-to-use XML database BaseX. VLE is freely available and can be downloaded from the ACDH-CH website under the address https://www.oeaw.ac.at/acdh/tools/vle.

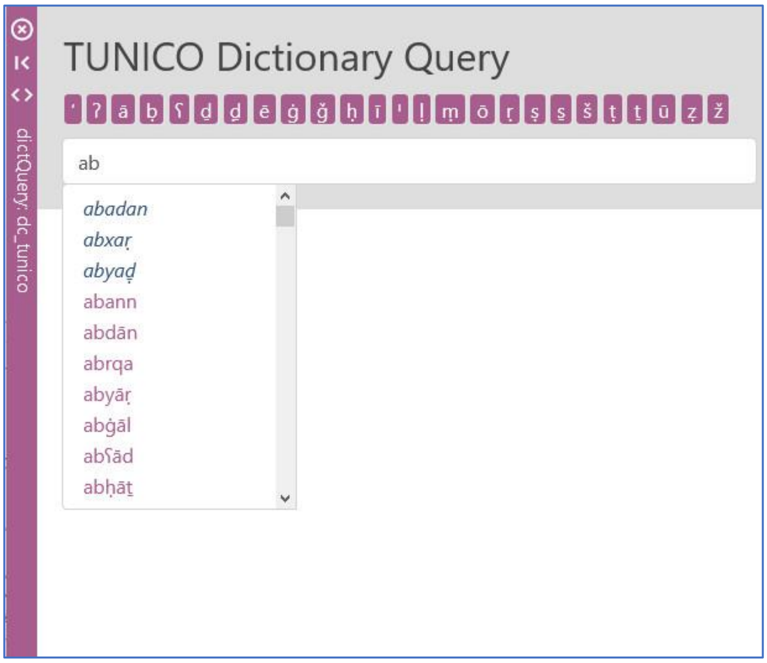
Lexical data constitute a special case in many respects. When it comes to standards, the situation is rather complex since there exist several competing systems. ISO (The International Organization for Standardization) has been working on a revision of the LMF (Lexical Markup Framework) standard in recent years. However, in academic dictionary editing, making use of the TEI dictionary module (TEI Consortium 2016: 275-313) in order to encode dictionaries has become a fairly common standard procedure. It has been shown repeatedly that the TEI dictionary module is also usable for NLP purposes and born digital material (Budin et al. 2012). The currently running H2020-funded project eLexis (European lexicographic infrastructure) has adopted as primary formats for their developmental work TEI Lex-0, a TEI customisation that has been established as a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources (Romary et al. 2020) and OntoLex-Lemon, the OntoLex format being the de-facto standard for representing lexical information (McCrae 2020).

4.3 Publishing digital dictionaries

The system has been developed at the Austrian Academy of Sciences for several years now and is being used also for other dictionaries. It is very flexible and eases the building of new web-applications. So far, the VICAV dictionaries have all been made available via the VICAV website, each with their own specialised interface. However, the technology behind VICAV allows also to create independent websites. This has been done on an experimental level for the TUNICO and Damascus dictionaries as the organisation of contents into panels proved to be the right solution for the research environment. Unfortunately, this way of data display does obviously not work on devices with small screens.

For these dictionaries the initial specifications contained six minimum requirements that had to be met: auto-completion, text-searching with wild-cards, the capability to limit queries to particular fields (i.e. parts of a dictionary entry identified by Xpath expressions), the capability to combine different fields in a query, to provide the interface with adequate instructions to allow also inexperienced users to use it, and to furnish input tools to allow queries with the adequate alphabets.

One of the frequent most unpleasant features in websites of academic dictionaries is the high entry threshold, that allows only experienced insiders who are familiar with the system to enter meaningful queries, while users without insider knowledge are left with no possibility to see any results. Auto-completion – i.e. a mechanism for letting the user dynamically browse an index in a dataset, is a remedy to that as that provides users with clues as to what they might be looking for.



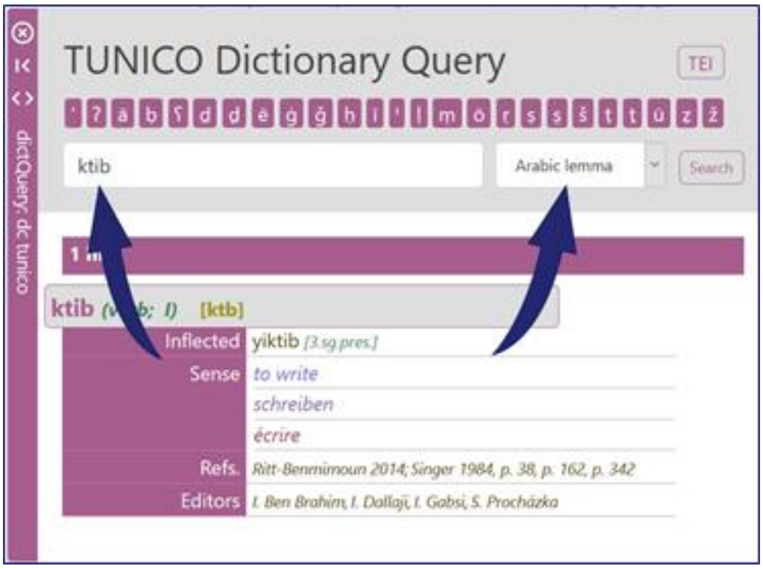
TUNICO Dictionary Query

ab

abadan
abxaṛ
abyaḍ
abann
abdān
abrqa
abyāṛ
abġāl
abʕād
abḥāṯ

Figure 8: Auto-complete control in the dictionary interface

Online dictionaries come in many different shapes. With respect to search capabilities most of what is available does not offer a wide range of functionalities and is rather simplistic. A determining factor is – of course – the structure of the underlying data. The highly structured TEI encoded VICAV dictionaries allow to formulate queries that retrieve data from all available fields (i.e. named parts of a TEI entry) and combinations thereof. There are two possibilities to select particular fields: either using the drop-down list next to the query input or by adding manually a prefix like *lemma*= to the query string.



TUNICO Dictionary Query

ktib | Arabic lemma

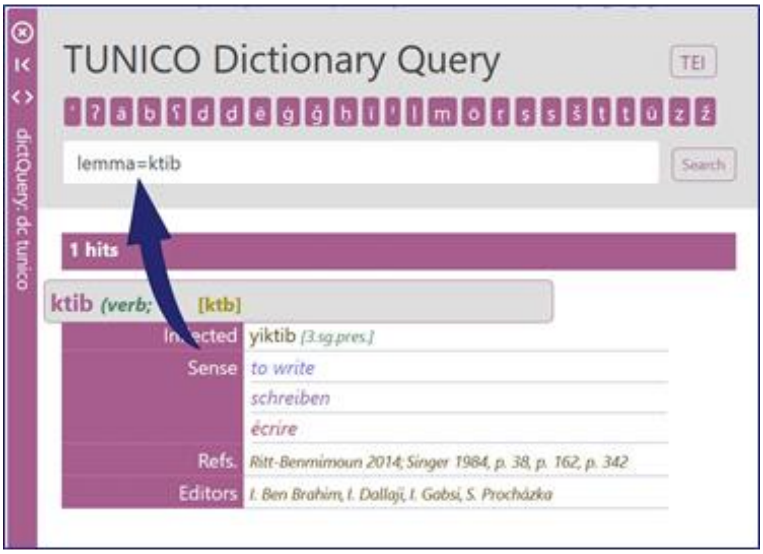
1

ktib *(verb; I)* [ktb]

Inflected	yiktib *[3.sg.pres.]*
Sense	*to write*
	schreiben
	écrire
Refs.	*Ritt-Benmimoun 2014; Singer 1984, p. 38, p. 162, p. 342*
Editors	*I. Ben Brahim, I. Dallaji, I. Gabsi, S. Procházka*

Figure 9: Querying in selected fields (elements) of the dictionary entries

Looking for a lemma might look as indicated in Figure 9, or in Figure 10 below:



TUNICO Dictionary Query

lemma=ktib

1 hits

ktib *(verb;* [ktb]

In_ected yiktib *[3.sg.pres.]*

Sense *to write*

schreiben

écrire

Refs. *Ritt-Benmimoun 2014; Singer 1984, p. 38, p. 162, p. 342*

Editors *I. Ben Brahim, I. Dallaji, I. Gabsi, S. Procházka*

Figure 10: Querying in the lemma field

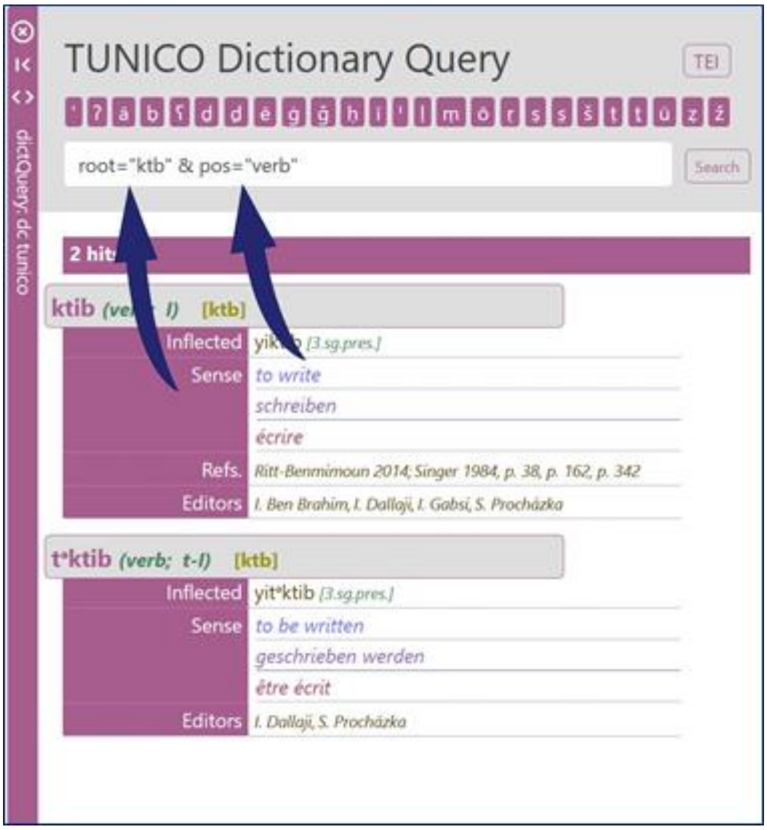
The options provided by the prefixes offer a higher degree of flexibility:

Prefix	Full form	Explanation
de|en|fr|es	German|English|French| Spanish	The iso values indicate in which translation equivalents the query should be performed.
infl	inflected form	Inflected forms include plurals, count nouns, third person singular forms of verbs, verbal nouns etc.
inflType	inflectional type	With this prefix, one can indicate morphological categories such as plural, feminine etc. As to the available labels have a look at the *VICAV Encoding Guidelines*.
lemma		This is the canonical form of the entry.
pos	part-of-speech	By indicating the wordclass the query looks for nouns, verbs, adjectives etc.
etym	etymology	This is mainly used to identify foreign loans which are particularly frequent in the *TUNICO Dictionary*.
etymLang	etymology language	This prefix allows to search for etymologies in particular languages.
root		The consonantal radicals that determine the basic semantics of words in the Semitic languages
stem		With this prefix it is possible to search for derivational verbal patterns. The system follows largely Woidich (2006, p.87sq.): I, II, t-II, etc.

subc	sub-category	With this prefix, one can query for phenomena such as *collective nouns* (subc=collectiveNoun) or *diminutives* (subc=diminutive).

Table 1

The system also allows to define constraints by combining more than one atomic query term making use of the operator & (=and). The following query contains two terms and retrieves all entries of the root *ktb* which carries the basic meaning of ‘writing’.



root="ktb" & pos="verb"

2 hit

ktib *(ve I)* [ktb]

Inflected	yik b *[3.sg.pres.]*
Sense	*to write*
	schreiben
	écrire
Refs.	*Ritt-Benmimoun 2014; Singer 1984, p. 38, p. 162, p. 342*
Editors	*I. Ben Brahim, I. Dallaji, I. Gabsi, S. Procházka*

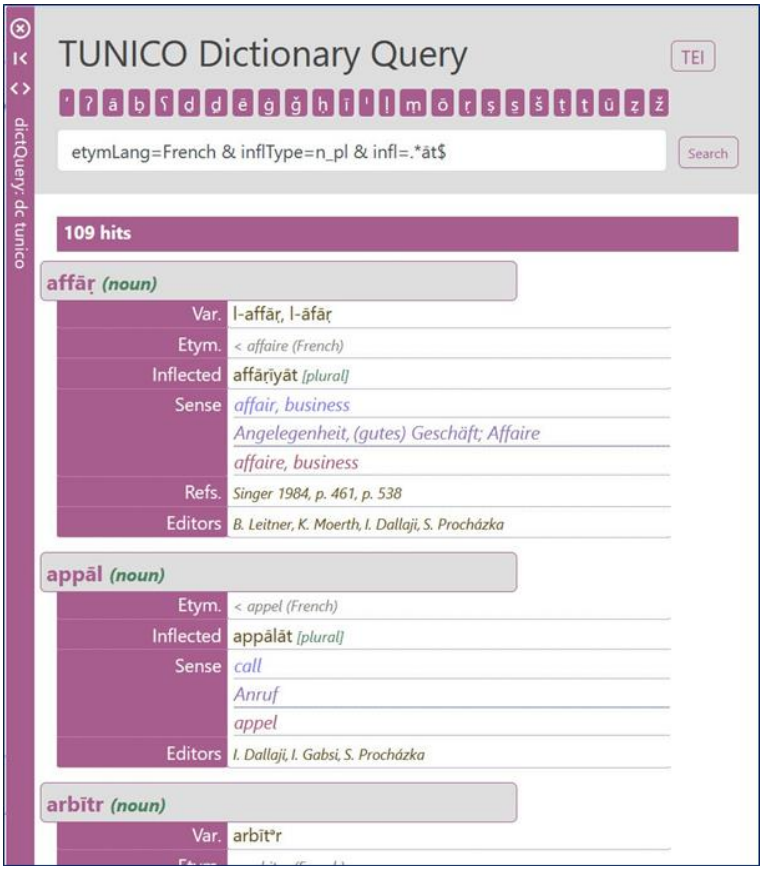
tᵊktib *(verb; t-I)* [ktb]

Inflected	yitᵊktib *[3.sg.pres.]*
Sense	*to be written*
	geschrieben werden
	être écrit
Editors	*I. Dallaji, S. Procházka*

Figure 11: Looking for all verbs with the root “ktb” in the TUNICO dictionary.

The number of terms may, of course, also be larger, like in the following example, in which you retrieve all Tunisian nouns of French origin whose plural ends in -*āt*.

As can be seen in the following example, the terms can also contain wildcards. The details are furnished in the help section of the TUNICO dictionary:



etymLang=French & inflType=n_pl & infl=.*āt$

109 hits

affāṛ *(noun)*

Var.	l-affāṛ, l-āfāṛ
Etym.	*< affaire (French)*
Inflected	affāṛīyāt *[plural]*
Sense	*affair, business* *Angelegenheit, (gutes) Geschäft; Affaire* *affaire, business*
Refs.	*Singer 1984, p. 461, p. 538*
Editors	*B. Leitner, K. Moerth, I. Dallaji, S. Procházka*

appāl *(noun)*

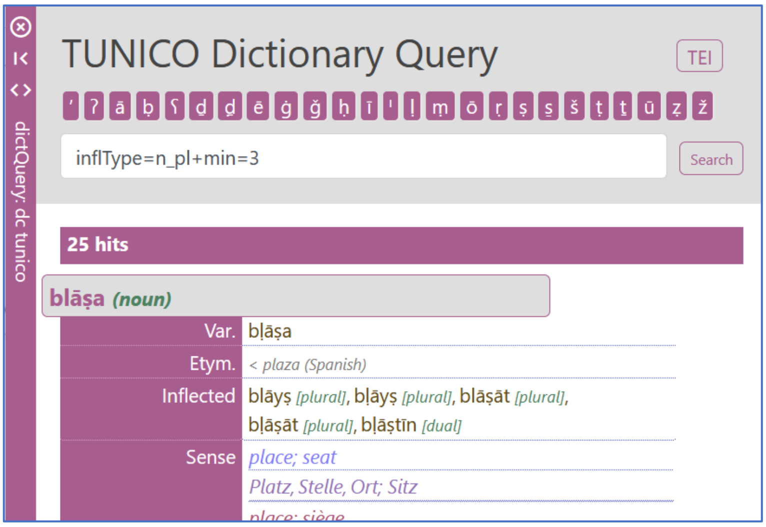
Etym.	*< appel (French)*
Inflected	appālāt *[plural]*
Sense	*call* *Anruf* *appel*
Editors	*I. Dallaji, I. Gabsi, S. Procházka*

arbītr *(noun)*

Var.	arbītᵊr

Figure 12: Query made up of three terms

In VICAV 3.0, the operator for morphological forms *inflType* can be combined with an operator to indicate a minimum occurrence number. The following query retrieves all nouns with at least three plural forms.



inflType=n_pl+min=3

25 hits

blāṣa *(noun)*

Var.	bḷāṣa
Etym.	*< plaza (Spanish)*
Inflected	blāyṣ *[plural]*, bḷāyṣ *[plural]*, blāṣāt *[plural]*, bḷāṣāt *[plural]*, bḷāṣṭīn *[dual]*
Sense	*place; seat* *Platz, Stelle, Ort; Sitz* *place; siège*

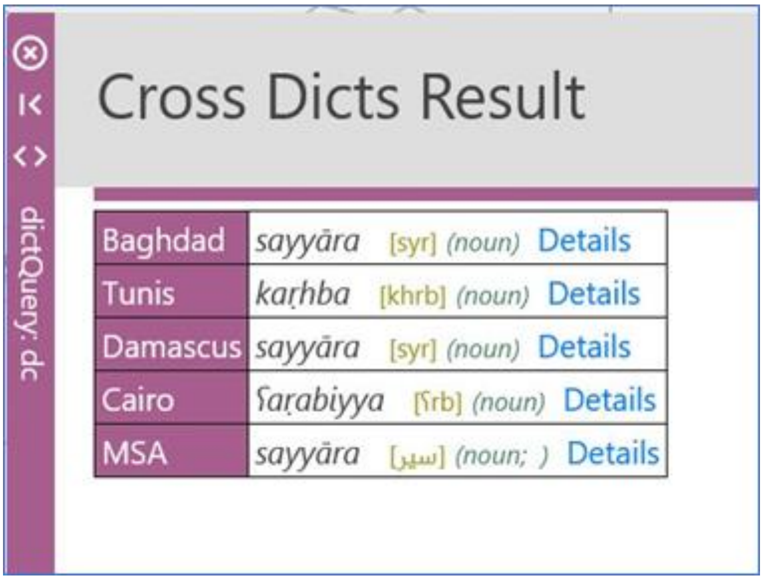
Figure 13: Looking for nouns with at least three plural forms

As mentioned before, comparative research into Arabic dialects was one of the main purposes of creating the VICAV dictionaries. A mid-term goal of this undertaking has always been a digital *Comparative Dictionary of Arabic Varieties* based on the existing dictionaries, a lexicographic system that allows to compare the different Arabic varieties. Several experiments towards integrating the various resources were undertaken over the past few years and VICAV 3.0 offers an interface that eases consulting all or selected dictionaries simultaneously.

In principle, all queries that can be performed in the dictionary specific interfaces can also be launched as cross-dictionary requests. Users can choose between two options to display search results: either juxtaposing the entries in separate panels or getting an integrated listing of relevant results extracted from the various dictionaries. This synoptic view contains links which enable users to directly inspect the detailed entries.

In order to get meaningful results, it will often be necessary to use combined query terms. One obvious option is to formulate queries for consonantal roots of words, which are used in combination with patterns for inflection and derivation, information every entry in the dictionaries contains. Here, it is noteworthy that, in assigning roots to the lemmas, the VICAV dictionaries do not proceed from the synchronic situation, but attribute corresponding Classical Arabic (CA) roots wherever a colloquial lexeme can be traced back to a CA cognate.

For many questions recurring to etymology may not be the way to go as common concepts do not necessarily share the same etymological roots. Another option is to match senses via translation equivalents. The following screenshot gives vernacular forms of English ‘car’, which is a good example of a concept, which has etymologically completely unrelated equivalents in the various varieties. Interestingly, three out of four can be traced back to CA words: *ṭumubil* (Salé), *kaṛhba* (Tunis) vs. *ʕaṛabiyya* (Cairo) vs. *sayyāra* (Damascus).



Cross Dicts Result

Baghdad	*sayyāra* [syr] *(noun)* Details
Tunis	*kaṛhba* [khrb] *(noun)* Details
Damascus	*sayyāra* [syr] *(noun)* Details
Cairo	*ʕaṛabiyya* [ʕrb] *(noun)* Details
MSA	*sayyāra* [سير] *(noun;)* Details

Figure 14: The result of a query across the five currently published VICAV dictionaries

5 Outlook

VICAV 3.0 is a first attempt at consolidating the infrastructure in a way that it can be easily customised to new necessities in projects to come. The next major steps on the current to-do-list will be the publication of consolidated ODDs. We are currently working on frontend updates implementing a reading mode and printable views.

With respect to lexicography, the work is being directed towards enhancing the number of dictionaries, but also on increasing the number of published lexical items in the existing dictionaries, a major concern being more parallel data in the various dictionaries. In addition, it is planned to provide standalone instances of the dictionaries in order to allow their use on handheld devices and thus to increase their usability.

References

Bird, Simon & Gary F. Simons. 2008. Toward a Global Infrastructure for the Sustainability of Language Resources. In: *22nd Pacific Asia Conference on Language, Information and Computation*, 87–100. (https://www.aclweb.org/anthology/Y08-1008.pdf)

Budin, Gerhard, Stefan Majewski & Karlheinz Moerth. 2012. Creating Lexical Resources in TEI P5. In *Journal of the Text Encoding Initiative (jTEI)* 3. (doi:10.4000/jtei.522)

Ferguson, Charles. 1959. Diglossia. In *Word* 15. 325-340.

Haywood, John A. 1960. *Arabic Lexicography. Its History and its Place in the General History of Lexicography*. Leiden: Brill.

Haywood, John A. 1991a. Arabic Lexicography. In F. J. Hausmann, O. Reichmann, H. E. Wiegand and L. Zgusta (eds): *Dictionaries. An International Encyclopedia of Lexicography* 5.3, 2438-2448. Berlin/New York: de Gruyter.

Hoogland, Jan. 2008. Lexicography: Bilingual Dictionaries. In K. Versteegh (ed): *Encyclopedia of Arabic Language and Linguistics III*, 21-30. Leiden/Boston: Brill.

Leitner, Bettina, Stephan Procházka & Fadi Yousuf. 2021. Lehrbuch des Irakisch-Arabischen: Praxisnaher Einstieg in den Dialekt von Bagdad. Wiesbaden: Harrassowitz.

Marçais, William & Abderrahmân Guîga. 1958-61. Textes arabes de Takroûna. II: Glossaire. 8 vol. Paris.

McCrae, John. 2020. Interoperable Interface for Lemon and TEI resources (D2.2). (https://elex.is/wp-content/uploads/2020/02/ELEXIS_D2_2_Interoperable_Interface_for_Lemon_and_TEI_resources.pdf)

Moerth, Karlheinz. 2018. Arabic lexicography in the Internet era. In Pedro A. Fuertes-Olivera (ed.): *The Routledge Handbook of Lexicography*, 503-517. London and New York: Routledge.

Moerth, Karlheinz, Stephan Procházka & Ines Dallaji. 2014. Laying the Foundations for a Diachronic Dictionary of Tunis Arabic. A First Glance at an Evolving New Language Resource. In A. Abel, C. Vettori and N. Ralli (eds): *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 377-387. Bolzano: EURAC research.

Moerth; Karlheinz, Daniel Schopper & Omar Siam. 2015. Towards a Diatopic Dictionary of Spoken Arabic Varieties: Challenges in Compiling the VICAV Dictionaries. In G. Grigore and G. Bițună (eds): *Arabic Varieties: Far and Wide. Proceedings of the 11th International Conference of AIDA*, 395-404. Bucharest.

Moerth, Karlheinz, Daniel Schopper & Omar Siam. 2017. Linking Instead of Lemmatising. 2017. Enriching the TUNICO Corpus with the Dictionary of Tunis Arabic. In V. Ritt-Benmimoun (ed.) *Tunisian and Libyan Arabic Dialects: Common Trends - Recent Developments - Diachronic Aspects*, 219-238. Zaragoza: Prensas de la Universidad de Zaragoza.

Nicolas, Alfred. 1911. Dictionnaire français-arabe: idiome tunisien and Dictionnaire arabe-français. Tunis.

Procházka, Stephan & Karlheinz Moerth. 2017. The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects. In Al-Hamad, M., Ahmed, R. and Aloui, H. (eds): *Lisan Al-Arab: Studies in Contemporary Arabic Dialects, Proceedings of the 10th International Conference of AIDA, Qatar University 2016*. Vienna: LIT Verlag, 176-183.

Procházka, Stephan, Rima Aldoukhi & Anna Telič. 2014-2015. *Lehrbuch des Syrisch-Arabischen: Praxisnaher Einstieg in den Dialekt von Damaskus*. Wiesbaden: Harrassowitz.

Quéméneur, Jean. 1962. Glossaire de dialectal. In *IBLA*, 1962. 325-67.

Romary, Laurent & Toma Tasovac. 2020. TEI Lex-0 — A baseline encoding for lexicographic data. (https://dariah-eric.github.io/lexicalresources/pages/TEILex0)

TEI Consortium. 2020. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Version 4.1.0. (www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf).

Versteegh, Kees. (ed.) 2006-2009. *Encyclopedia of Arabic Language and Linguistics (EALL)*. 4 vols, Leiden/Boston: Brill.

Woidich, Manfred. 2006. Das Kairenisch-Arabische. Wiesbaden: Harrassowitz.

Zaghouani, Wajdi. 2014. Critical Survey of the Freely Available Arabic Corpora. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (LREC 2014)*. 1-8. (http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-OSACT%20Proceedings.pdf)