Navigating linguistic diversity: modelling diatopic and bibliographic information with TEI Lex-0

Veronika Engler[1], Karlheinz Mörth[1], Stephan Procházka[2], Michaela Rausch-Supola[1], Daniel Schopper[1]

[1] Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Bäckerstraße 13, 1010 Vienna, Austria

[2] University of Vienna, Department of Near Eastern Studies, Spitalgasse 2, Court 4.1, 1090 Vienna, Austria

E-mail: {veronika.engler, karlheinz.moerth, michaela.rausch-supola, daniel.schopper} @oeaw.ac.at, stephan.prochazka@univie.ac.at

Abstract

The Vienna Corpus of Arabic Varieties (VICAV) is a digital language documentation platform that integrates methods from language technology and the digital humanities. It provides a modular, sustainable infrastructure for representing heterogeneous data, with a strong focus on standards, data modelling, and TEI-based encoding. Alongside different data types such as a growing digital bibliography and corpora, VICAV offers dictionaries of different Arabic varieties, covering Baghdad, Cairo, Damascus, Tunis, and Modern Standard Arabic. The dictionaries are compact lexical databases with structured entries and English translations, some also including German, French, or Spanish. A sixth dictionary, the SHAWI Dictionary, will be published as a beta version in the end of 2025 and is the first in the series to be encoded in TEI Lex-0, a community-driven baseline for lexicographic data. This paper discusses some modelling decisions behind the SHAWI Dictionary, including the encoding of diatopic variation —that is language variation across different geographical spaces—and integration of bibliographic sources. It also presents our TEI Lex-0 customization and addresses challenges in modelling grammatical features of Arabic varieties, particularly where existing standards fall short in representing Semitic structures.

Keywords: TEI Lex-0; lexicographic modelling; dialect dictionaries; Arabic dialectology

1. VICAV and the SHAWI Project

The Vienna Corpus of Arabic Varieties (VICAV) is a growing language documentation platform that hosts a diverse array of digital language resources. By integrating methodologies from language technology and the broader domain of text-oriented digital humanities, the project seeks to address the challenges of representing heterogeneous linguistic data by providing a flexible, yet sustainable technical infrastructure grounded in a modular data architecture. A core emphasis of the project lies in its commitment to open access, adherence to standards and best practices, as well as robust data modelling. From its inception, VICAV has adopted a text-centric approach, firmly rooted in the application of the *Guidelines* of the Text Encoding

Initiative (TEI Consortium, 2025). This infrastructure is designed to ensure consistent encoding practices across projects, while also supporting sustainable workflows for the creation, management, and publication of linguistic data.

VICAV provides a varied range of structured linguistic resources: (1) a bibliography of relevant research literature, (2) concise linguistic profiles, offering standardised, brief descriptions of individual language varieties; (3) structured inventories of linguistic features; (4) annotated sample texts, (5) digital corpora of transcribed speech and (6) digital dictionaries (Procházka & Moerth, 2015). These components are meant to be used both for descriptive and comparative linguistic research, while adhering to principles of transparency, interoperability, and long-term accessibility. Most of the data can be accessed as text and/or visualised on a map within the VICAV web application (https://vicav.acdh.oeaw.ac.at).

VICAV has served as an umbrella for third-party funded research endeavours. Without constant financing and substantial 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. Initially, data production relied largely on enthusiasts, student assistants and students financed by the Faculty of Philological and Cultural Studies of the University of Vienna. Meanwhile, several projects have allowed to further develop the platform creating a lively biotope which will hopefully lead to continued activities. One of these projects is the SHAWI project (*The Shawi-type Arabic dialects spoken in South-eastern Anatolia and the Middle Euphrates region;* 2021-27), a joint endeavour of the University of Vienna and the Austrian Academy of Sciences financed by the Austrian Science Fund FWF (P-33574). It investigates the Arabic dialects spoken by (formerly) goat- and sheep-breeding Bedouin tribes in various regions of Turkey, Syria, Lebanon and Iraq. These so-called Shawi-type dialects have been little investigated compared to other dialects of the region. The primary outcomes of the project, which will end in early 2027, are based on newly obtained field data, and are as follows: (1) the development of a text corpus based on extensive free-speech audio recordings, (2) the compilation of a comprehensive grammatical description, and (3) the creation of a digital dictionary, the SHAWI Dictionary.

The second VICAV project that is relevant for this paper due to its rather close link to the SHAWI project, both thematically and technologically, is the WIBARAB project (*What is Bedouin-Type Arabic?* 2021-2026; ERC 101020127-WIRARAB). Like the SHAWI project, WIBARAB focuses mainly on Arabic varieties spoken by (formerly) nomadic communities. WIBARAB is the largest of the VICAV projects in terms of resources, duration and scope, and it investigates the linguistic and socio-historical realities behind the millennia-old dichotomous concept of nomadic and sedentary Arabic speaking people in the Middle East and North Africa (Procházka, 2024). In addition to several text corpora, the newly collected data gathered in extensive fieldwork in Saudi Arabia, Kuwait, Jordan, Lebanon, Sudan, Tunisia and Morocco, together with previously published data is systematically collected in a database of

linguistic features that has been available since the beginning of the project (https://github.com/wibarab/featuredb/), and which will eventually be integrated into the VICAV framework.

In spite of its name, *Vienna Corpus of Arabic Varieties*, VICAV only contains a very 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[1] corpus, which at this point of time 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., 2015). The SHAWI project, which is being built upon a modernised version of the VICAV software framework, will finally add sizeable digital corpora to the collection with a similar structure. Next to this, also TUNOCENT[2] (https://tunocent.acdh.oeaw.ac.at) extends the collection of transcriptions by 24 texts.

2. The SHAWI Dictionary

An essential part of the VICAV platform are its dictionaries. The VICAV dictionaries published so far cover five linguistic varieties: Baghdad, Cairo, Damascus, Tunis and Modern Standard Arabic, the latter one mainly serving as a point of reference for the others. The Baghdad dictionary is the smallest one with approximately 1100 entries. The *Digital Dictionary of Damascene Arabic* was based on the vocabulary used in the 2 volume *Textbook of Syrian Arabic* (Aldoukhi et al., 2014-15). The largest of the VICAV dictionaries is the micro-diachronic *TUNICO Dictionary* which was compiled on the basis of a corpus of recorded conversations from Tunis. In addition to the TUNICO corpus, this resource also incorporates lexical material extracted from printed sources (Procházka & Moerth, 2015). Generally, these dictionaries are all small in size, none of them containing more than 8000 entries, and constitute lexical databases with structured lexicographic information (Moerth & Schopper, 2021). They are all provided with English translation equivalents (some also have additional translations into German, French, Turkish or Spanish), and are all built on the same technology and follow a similar encoding schema.

The most recent addition to the series of VICAV dictionaries so far, is the SHAWI Dictionary. It will contain the entire vocabulary of the afore-mentioned digital corpus of the SHAWI project, and as with previous VICAV projects, its compilation is closely linked to the corpus creation process. On the one hand, the dictionary serves as a tool

[1] *Linguistic dynamics in the Greater Tunis Area: a corpus-based approach* (TUNICO; 2013-16; FWF P-25706). TUNICO was the first VICAV projects with third-party funding.

[2] *The terra incognita of Tunisia* (TUNOCENT; 2019-23; FWF P-31647). TUNOCENT is another project conducted within the VICAV framework.

to enrich the linguistic annotation of the corpus, algorithmically creating wordforms which are used in pre-processing the token-level annotations of the corpus, which in a second step are manually verified. On the other hand, the corpus serves as input for the dictionary (Moerth et al., 2013).

Unlike the other lexcial resources in VICAV, the SHAWI Dictionary has to deal with a rather high degree of internal lexical variation reflecting the nomadic background of its speaker community. This variation correlates both with the social identity of the speakers (esp. a speaker's membership to a given tribe) and with geographic distribution. In order to adequately model this situation, it is necessary to enhance the structure of the VICAV dictionaries with diatopic[3] information at several positions. Together with its hybrid provenience (both research literature and the corpus) and a series of other special cases, this leads to the necessity of extending the underlying TEI Lex-0 encoding scheme—as we will show below.

2.1 From TEI to Lex-0

While many industry standards have been established by organisations outside academia—such as ISO (International Organization for Standardization), W3C (World Wide Web Consortium), and IETF (Internet Engineering Task Force)—academic research has also played a significant role in shaping standards, workflows, and controlled vocabularies. The growing recognition of the importance of standards in research has been largely driven by the rapid pace of the digital transformation. As researchers more and more operate within shared infrastructures and collaborate across projects and disciplines, the need for common knowledge representation practices has become increasingly evident. Standards now serve critical functions in ensuring reusability, interoperability, preservation, and accessibility of research outputs, especially fostered by European research infrastructures like CLARIN or DARIAH.

One of the methodological aims of the VICAV projects has been the modelling of lexical data and the documentation of development workflows, both for research and didactic purposes. A key objective has been to offer guidance to others working in the field—particularly regarding relevant standards and best practices—while also sharing insights gained through practical experience. To support this goal, the developmental framework of VICAV—and especially its dictionaries—has been built entirely on the *Guidelines* of the Text Encoding Initiative. TEI has emerged as the widely accepted community standard for academic text encoding, addressing many of the challenges faced in this domain. Over time, the TEI *Guidelines* have become deeply integrated into the project's infrastructure. Today, nearly all components of the VICAV system are based on TEI, and all text types are modelled in compliance with its specifications.

[3] The term *diatopic* refers to variations in language or other phenomena that occur across different geographical locations or spaces.

Lexical data present an interesting case when it comes to standardisation, characterised by a complex landscape of competing systems. While the International Organization for Standardization (ISO) has been revising the Lexical Markup Framework (LMF) not long ago (Frontini et al., 2023), the TEI dictionary module has become a widely adopted standard in academic lexicography. Originally developed for the digitisation of historical dictionaries, the TEI module has also proven very effective for natural language processing (NLP) applications and born-digital lexical resources (Budin et al., 2012).

In the Horizon 2020-funded ELEXIS project[4], which worked on integrating NLP methods with lexicographic workflows, two primary formats were used for development: TEI and OntoLex-Lemon (McCrae, 2020). A more recent initiative is TEI Lex-0, a streamlined customisation of the TEI dictionary module. Developed by a group of lexicographers and TEI experts, TEI Lex-0 aims to enhance the usability of TEI for lexicographic purposes by introducing a set of constraints on the use of TEI elements as well as enhanced documentation, offering an important resource for incentivising the harmonisation of lexicographic TEI-encoded data. In close collaboration with the lexicographic community, this initiative seeks to establish a baseline encoding model and a target format to improve the interoperability of lexical resources that have been encoded using diverse approaches (Tasovac et al., 2018ff.).

> *"TEI Lex-0 should not be thought of as a replacement of the Dictionaries Chapter in the TEI Guidelines or as the format that must be necessarily used for editing or managing individual resources, especially in those projects and/or institutions that already have established workflows based on their own flavors of TEI. TEI Lex-0 should be primarily seen as a format that existing TEI dictionaries can be unequivocally transformed to in order to be queried, visualised, or mined in a uniform way. At the same time, however, there is no reason why TEI Lex-0 could not or should not be used as a best-practice example in educational settings or as a foundation of new TEI-based projects. This is especially true considering the fact that TEI Lex-0 aims to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF (Lexical Markup Framework) standard." (Tasovac et alt, 2018ff.)*

With lexicographers increasingly adopting TEI Lex-0 (Salgado et al., 2019; Tasovac et al., 2020; Moerth, 2024), it is to be expected to gain importance in more and more upcoming dictionary projects. Among the VICAV dictionaries, the SHAWI Dictionary is the first one being encoded natively in TEI Lex-0.

[4] *European Lexicographic Infrastructure* (ELEXIS; 2018-22; https://doi.org/10.3030/731015). Cf. https://project.elex.is

2.2 Encoding the macrostructure

Initially, the system applied in VICAV to encode dictionaries was a customised version of TEI's Dictionary Module. Very much like Lex-0, the VICAV customisation imposed a number of constraints and was designed to keep it compatible across the various dictionaries of the collection. Actually, some of the developments in Lex-0 have been very similar to what has been done previously in the encoding of the VICAV dictionaries.

In line with the TEI *Guidelines*, dictionaries are conceptualised as a specific type of text and are therefore encoded using standard <text> elements. Each dictionary begins with a <teiHeader> element, which contains the metadata describing the resource.

The <body> of every VICAV dictionary follows a consistent macrostructure, comprising two main divisions within the <body> element:

- The division contains all dictionary entries, including both single-word and multiword units.
- The division holds all example sentences.

The decision to separate example sentences from the entries themselves is intentional. It allows for the reusability of examples across different entries, creating a more modular and flexible structure. This approach ensures consistency across dictionaries and supports both human readability and machine processing.

```
<TEI version="5.0">
   <teiHeader>
      ...
   </teiHeader>

   <text>
      <body>
         <div type="entries">
            <entry>...</entry>
            <entry>...</entry>
            <entry>...</entry>
            ...
            ...
            ...
         </div>

         <div type="example">
            <cit type="example">...</cit>
            <cit type="example">...</cit>
            <cit type="example">...</cit>
            ...
            ...
            ...
         </div>
      </body>
   </text>
</TEI>
```

Figure 1: Dictionary structure

2.3 The microstructure: entries and examples

Each entry consists of at least three components:

- morphological information regarding the lemma,
- grammatical information, and
- semantic information, typically in the form of senses with translation equivalents.

```
<entry xml:lang="ar-acm-x-shawi-vicav" xml:id="DShaAr.gbile_00000">
  <form type="lemma">
     <orth>ǧbile</orth>
     ...
  </form>

  <gramGrp>
     <gram type="pos">noun</gram>
     ...
  </gramGrp>


  <sense xml:id="DShaAr.s_gbile_00000_0">
     <cit type="translationEquivalent" xml:lang="en">
        <form><orth>mountain</orth></form>
     </cit>

     <cit type="translationEquivalent" xml:lang="de">
        <form><orth>Berg</orth></form>
     </cit>

     <cit type="translationEquivalent" xml:lang="tr">
        <form><orth>dağ</orth></form>
     </cit>

     <cit type="translationEquivalent" xml:lang="fr">
        <form><orth>montagne</orth></form>
     </cit>

  </sense>

  ...
</entry>
```

Figure 2: Structure of an entry

Multiword units follow the same structure as single word entries, with only the @subtype attribute on the lemma indicating the different type of the headword. This is a slight variation of the Lex-0 specification which discusses several cases of multi word units, however at the time of writing does not provide guidance on how to encode MWEs as the headword of a dictionary entry. In fact, lemma and compound are both proposed values for the @type attribute on <form>, making them mutually exclusive. Our decision to separate both pieces of information into two attributes seeks a balance between the need for homogenously encoding all headwords while maintaining the possibility for differentiating both types of entries.

```
<entry xml:lang="ar-acm-x-shawi-vicav" xml:id="DShaAr.dibis_efrengi_00000">
   <form type="lemma" subtype="compound">
       ...
       <orth>dibis ᵊfrənği</orth>
   </form>
   ...
   <sense xml:id="DShaAr.s_dibis_efrengi_00000_0">
       <cit type="translationEquivalent" xml:lang="en">
           <form><orth>tomato paste</orth></form>
       </cit>
       <cit type="translationEquivalent" xml:lang="de">
           <form><orth>Tomatenmarkpaste</orth></form>
       </cit>
       <cit type="translationEquivalent" xml:lang="tr">
           <form><orth>domates salçası</orth></form>
       </cit>
   </sense>
   ...
</entry>
```

Figure 3: Structure of a multiword unit

As stated before, usage examples are not inserted directly into the <entry> elements. They are referenced via pointer elements that contain unique identifiers (IDs) to address them. In principle, they belong into the respective sense elements.

```
<cit xml:lang="ar-acm-x-shawi-vicav" xml:id="DShaAr.al_halib_gabb" type="example">
    <usg type="geographic">
        <name type="place" ref="geo:harran_urfa">Harran-Urfa</name>
    </usg>

    <quote>al-ḥalīb gabb.</quote>
    <cit type="translation" xml:lang="en">
        <quote>The milk has spilled over.</quote>
    </cit>

    ...

</cit>
```

Figure 4: Structure of an example

2.4 Digital tools

Employing a well-defined subset of elements within concrete projects always poses a deliberate decision between a project's specific outlook and encoding needs vs. the lines of generalisation any interchange format must draw. The special case of the SHAWI Dictionary building upon Lex-0 is not an exception. Fortunately, the TEI infrastructure natively provides a structured process to derive a schema from another one, called ODD chaining (Rahtz, 2014). This method allows a project to start from a base schema—in our case TEI Lex-0—and extend or restrict it to its needs (cf. section 23.8.1 *TEI Customizations* in: TEI Consortium 2025). Like this, the schema draws on the commonly agreed semantics of a publicly available format while transparently

documenting any differences which could hamper data reuse by others.

In the case of the SHAWI Dictionary, we are following a two-step approach: given the diverse background of the VICAV dictionaries, we have identified the need to implement an encoding scheme which defines their common structural backbone (e.g. the macrostructure described above) and a baseline shared by all our lexicographic resources, which is derived from Lex-0 via ODD chaining. *ACDH generic dict schema*, as it is called, is not expected to be used directly in a dictionary project, but serves as starting point for project-specific customizations which usually introduce nuances not only in encoding but have a need for their own editorial guidelines and documentation, all of which can be generated from the ODD format. On the other hand, the *generic dict schema* should serve as an integration data layer which also eases the re-use of legacy data within this dictionary infrastructure.

The technical part of the lexicographical infrastructure is made up of two main components: a central dictionary server, implemented on top of the open-source XML database *BaseX*, and a bespoke XML editor, *<TEI>Enricher*, which reads and writes data through a REST API. The majority of the VICAV data has been created and edited using *<TEI>Enricher*, a versatile XML editor developed over many years at the ACDH. This freely available, general-purpose XML tool is designed to facilitate the creation of TEI-compliant documents and supports experimental workflows needed for our lexicographic work. It includes several built-in features that simplify the composition of large text documents and enable their visualisation via XSLT. Among its functionalities is support for working with TEI-encoded documents that include geospatial data. A dedicated map component, integrated with the *GeoNames* API, allows for the seamless incorporation of geo-coordinates. *<TEI>Enricher* has also been used extensively in the editing of the VICAV dictionaries.

Other important software components that have been used in the project are *oXygen*, which was mainly used for ODD editing, and *Github* which is being used for data management and data conversion through *Github Actions*.

3. Encoding particularities

The SHAWI Dictionary has been under development for quite some time now, and while there remain open issues, it is in a much more mature state at the time of writing this paper than ever before. In this chapter we will try to furnish additional evidence for the practical usability of TEI Lex-0 for the particular task of encoding dialectological lexical data of Semitic languages and discuss some of its features in the light of linguistic varieties for which it has not been used before. Before addressing issues in encoding grammatical phenomena typical of Semitic languages in general and Arabic in particular, this section focusses on modelling decisions concerning the encoding of diatopic information as well as the integration of heterogeneous sources as this dialectological dictionary also contains information from previous linguistic

descriptions of the dialects covered, requiring bibliographic references to indicate the source of certain blocks of information.

3.1 Encoding diatopic information (vs./and social dimension)

A dialectological dictionary dealing with more than one variety needs mechanisms to indicate where particular lexical items are used. In accordance with TEI Lex-0, the SHAWI schema employs <usg type="geographic"> for this purpose. Quite unsurprisingly, wordforms and variants can vary from place to place, thus this construct mainly occurs within <form> and <cit type="example"> in the SHAWI Dictionary; however, in some cases it can be equally important to document geographic distribution of other aspects of a lexicographic description. Thus, the schema also allows to embed <usg> within <sense> to express the scope of a particular semantic nuance, and even <gramGrp> to cater for cases where distinct grammatical properties are attested for a place.

The linguistic reality of the varieties under research is that certain wordforms are not solely limited in their geographical distribution but within this also to social restrictions. In the case of SHAWI, this is particularly frequent since specific lexicographic features are attested for specific tribes at a given place. Typically, such constraints are expressed in Lex-0 with a <usg> element of @type="socioCultural", defined as a "marker which identifies the use of a given lexical unit by particular social groups […]" (Section *12.1.42. <form>* in: Tasovac et alt., 2018ff.). However, complications arise when it comes to encoding several combinations of tribe and place for a single wordform, as the following example shows: the headword *ŝrib* is documented to be used by the tribes of the *Idin* and the *Abu Id* for the place of Bekaa, while the same word is documented to be used in Harran-Urfa without any (known) social restrictions. Simply encoding this series of data in a flat sequence of <usg> elements would hide this kind of relation amongst them:

```
<entry xml:lang="ar-acm-x-shawi-vicav" xml:id="DShaAr.shrib_00000">
    <form source="#d1e202928" type="lemma">
        <usg type="geographic">Harran-Urfa</usg>
        <usg type="geographic">Bekaa</usg>
        <usg type="socioCultural">Idin</usg>
        <usg type="socioCultural">Abu Id</usg>
        <orth>šrib</orth>
        ...
    </form>
    ...
</entry>
```

Figure 5: Diatopic information as a sequence of <usg> elements

While it would have been technically feasible to associate <usg> elements with each other by general-purpose pointer mechanisms provided by the TEI *Guidelines* (e.g. @corresp), this seemed rather impractical both from the perspective of data entry and retrieval. Another theoretic alternative would have been to repeat <form> elements for each distinct place- and tribe-combination. This, however, would have undermined one of the fundamental principles of VICAV dictionaries which states that common

morphological forms should be encoded in one and the same <form> element. In search of a wrapper element to group corresponding information we have also considered extending the TEI Lex-0 schema, e.g. by allowing <usg> to nest or even by introducing an entirely new element like <usgGrp>. All these approaches seemed rather weighty to solve a relatively minor encoding issue. At the end, we have chosen a pragmatic path: Since the information about tribal affiliation is always bound to a given place, we embed it within <usg type="geographic"> and use <name> elements to accommodate place names (which are mandatory in our schema) and tribe names alike. This way, we are able to keep the structure flexible and efficient while at the same time retaining human readability.

```
...
<form type="lemma">
    <usg type="geographic">
        <name type="place" ref="geo:harran_urfa">Harran-Urfa</name>
    </usg>

    <orth>šbiʕ</orth>
    <form type="variant">
        <usg type="geographic">
            <name type="place" ref="geo:bekaa">Bekaa</name>
            <name type="tribe" ref="pgr:idin">Idin</name>
            <name type="tribe" ref="pgr:abu_id">Abu Id</name>
        </usg>
        <orth>šbaʕ</orth>
    </form>
    ...
</form>
...
```

Figure 6: Encoding social affiliation

Moreover, adding <name> to the content model of <usg> allows us to use the @ref attribute to link to local reference resources which are being developed as part of the WIBARAB project: One is the WIBARAB gazetteer which helps aggregating various language resources in the dataset which are related to a common geographic unit. Given that the fieldwork in many VICAV projects is exploring areas and small villages which are not covered by general-purpose reference resources like *GeoNames* or *Wikidata*, the need for such a resource was evident early on.

```
<place type="reg" xml:id="harran_urfa">
   <placeName>Harran-Urfa</placeName>
   <location>
      <geo decls="#dms">36°51'36"N 39°01'53"E</geo>
      <geo decls="#dd">36.860000 39.031389</geo>
      <country key="TR">Turkey</country>
   </location>
   <idno type="geoNames">312531</idno>
</place>
```

Figure 7: VICAV gazetteer example

For the social dimension, the same mechanism allows us to reference a list of tribes which is also being established in the WIBARAB project in absence to a suitable authoritative machine-readable resource.

```
<listPerson xml:id="tribes">
    <head>List of tribes in <name type="project">WIBARAB</name></head>

    <personGrp role="tribe" xml:id="abu_eid">
        <name>Abu ʕĪd</name>
        <residence>
            <placeName ref="geo:haouch_el_harime"/>
            <note type="general">fieldwork 2021 - Beqaa close to Zahle</note>
            <placeName ref="geo:haouch_el_nabi"/>
            <note type="general">Younes</note>
            <placeName ref="geo:rayak"/>
        </residence>
    </personGrp>
</listPerson>
```

Figure 8: WIBARAB tribe list

While data on each tribe is still sparse, the list is already a pivotal point for the WIBARAB data architecture and provides a framework for future extension and refinement by other projects.

3.2 Encoding of bibliographic references

At the core of the VICAV collection lies a substantial and continuously growing digital bibliography of scholarly works on spoken Arabic varieties, compiled from a wide range of sources since the project's inception. The data is maintained in a Zotero group library (https://www.zotero.org/groups/2165756/vicav), where bibliographic records are enriched with a custom-designed keyword system, developed to allow the integration with other textual components of the database. This system enhances discoverability and interoperability of the records within the broader VICAV infrastructure. The bibliography encompasses not only academic research articles but also a variety of other language-related materials, including dictionaries, grammars, and textbooks. As of mid-2025, the database comprises more than 5,300 bibliographic records which makes it the largest digital bibliographic resource in this research field worldwide. The map-based display in the VICAV web application allows users to explore the data in terms of its regional distribution.

Given that no digital resources exist for the Shawi varieties so far, the project has to create its own corpus which is being annotated in parallel to the dictionary writing process. During this phase, the connection between the corpus and the dictionary is unidirectional, meaning that each token in the corpus is linked to its respective dictionary entry. Towards the end of the project, this will allow for the efficient semi-automatic generation of dictionary examples to be embedded again into the dictionary.

While the corpus provides the basis of the dictionary, it is also complemented by

information from research literature. To identify these sources, the dictionary provides links to the above mentioned Zotero library. Generally, the TEI Lex-0 Specification (cf. section 2.2.1. *Source description*) defines how the sources of various kinds should be documented in the <sourceDesc> of the dictionary as a whole. However, since various components of an entry in the Shawi dictionary may originate from different sources, we found it necessary to adopt a more granular method to encode attribution. Thus, in our current encoding scheme, every <entry> containing information from a publication is extended by a standard TEI <listBibl> element after the lexicographic information proper. The <listBibl> element contains placeholder bibliographic records which serve as proxies to the full entries from the VICAV library and can be referenced through the @source attribute wherever needed in the entry. Like this, it's possible to avoid the multiplication of bibliographic records being still able to quote exact page numbers of the relevant publications and add additional commentar to single occurrencessingle occurrences, if needed.

Figure 9: Bibliographic references

Context-sensitive display of possible values is ensured by the dictionary editor <*TEI*>*Enricher* which also helps to create the insertion of the <listBibl> elements and the linking to the Zotero library.

3.3 Linguistic particularities: additional *gram/@types*

One of the most obvious differences between TEI P5—that has been used in the older VICAV dictionaries—and Lex-0 is the way how grammatical descriptions are encoded. While in traditional TEI a number of specific elements such as <pos> (=part of speech, word class), <case>, <gen> (=gender), <mood>, <number>, <tns> (=tense) and <per> (=person) exist, Lex-0 has adopted a more flexible approach in which the different categories are encoded as @type attributes of <gram> elements. The Lex-0 approach makes the system much more flexible as additional @type values can easily be added. For instance, TEI P5 did not have an element for *aspect*, which in Lex-0 can elegantly be encoded as <gram type="aspect">.

```
...
<form type="inflected">
   <gramGrp>
      <gram type="aspect">imperfective</gram>
      <gram type="number">singular</gram>
      <gram type="person">3</gram>
   </gramGrp>
   <orth>yimši</orth>
</form>
...
```

Figure 10: Encoding aspect

The current Lex-0 specification (July 2025) lists 13 values for @type. The next two subchapters address lacunae in the existing system and will give some insight into which values have been added to the SHAWI schema.

```
<gram type="pos">n.</gram>
<gram type="case">acc.</gram>
<gram type="gender">f.</gram>
<gram type="inflectionType">7</gram>
<gram type="mood">indic.</gram>
<gram type="number">sg.</gram>
<gram type="person">3rd</gram>
<gram type="tense">aorist</gram>
<gram type="colloc">de</gram>
<gram type="aspect">imperf.</gram>
<gram type="valency">intr.</gram>
<gram type="government">[+conj.]</gram>
<gram type="degree">comp.</gram>
```

Figure 11: Lex-0 @type values on <gram>

3.3.1 About roots and stems: encoding morphological information

Lemmatisation in the VICAV dictionaries deviates from that of traditional Arabic dictionaries. To be more specific, it is necessary to say a quick word about roots and patterns in Semitic languages like Arabic, Aramaic or Hebrew. These languages are characterized by a root-based morphology, a root typically consisting of three, sometimes also of two or four consonants and roots carrying the basic semantics of derived words and wordforms. By inserting vowels and affixes into roots and thus forming specific patterns, words like Classical Arabic *kataba* 'he wrote', *kitāb* 'book', *kātib* 'writer' and *maktaba* 'library' are formed which all share the same root and an obvious common semantic base. In Arabic dictionaries, words are often grouped on the basis of these roots.

In the VICAV dictionaries, each of the above-mentioned words is a lemma in its own right, which also holds true of derived verbal stems. The roots are thus not used to group lemmata pertaining to one particular root by using a <superEntry> element (which does not exist in Lex-0 anyway) or nesting <entry> elements (a construct which would be theoretically allowed in Lex-0). Rather, all entries are furnished with information about the roots which allows users to use them as a search criterium. As

we are talking about grammatical data, the respective information is consequently encoded in <gram> elements with @type="root".

The SHAWI schema introduces an additional differentiation into synchronic and diachronic roots. The introduction of these categories has become necessary to provide a *tertium comparationis* when cross-searching dictionaries of different Arabic dialects with the aim of comparing lexical features. The reference to the equivalent root in Classical Arabic (CA) is used to detect cognate lemmas which in many cases are different from the realisation of similar forms in other contemporary varieties and in CA. This is why all lemmas are provided with information regarding these two types of roots. The noun *čalib* 'dog', used in the Harran-Urfa region is indirectly related to the CA word *kalb* 'dog', by indicating both the synchronic and diachronic root.

```
...
<form type="lemma">
   ...
   <orth>čalib</orth>
</form>

...

<gramGrp>
   <gram type="pos">noun</gram>
   <gram type="synRoot" xml:lang="ar-acm-x-shawi-vicav">člb</gram>
   <gram type="diaRoot" xml:lang="ar">klb</gram>
</gramGrp>
...
```

Figure 12: Encoding Semitic roots

Another addition to the proposed set of type-values is morphPattern (=morphological pattern), which is applied to all word classes but particles and proper names. The numbers 1, 2 and 3 (v. Figure 13) represent the root consonants of the forms. This information allows researchers to query for all wordforms with a particular morphological pattern and can co-occur with any word class.

```
...
<form type="lemma">
   ...
   <orth>čalib</orth>
   <gramGrp>
      <gram type="morphPattern">1a2i3</gram>
   </gramGrp>
</form>

...
```

Figure 13: Morphological pattern of the wordform *čalib* 'dog'

@type="degree" has been integrated into the Lex-0 specification only recently (June 2025) and is used in combination with adjectives. While possible content of this construct will be the terms *comparative* or *superlative* for Indo-European languages, the term *elative* can be applied for Semitic languages.

```
...
<form type="lemma">
   ...
   <orth>aḥsan</orth>
</form>

<gramGrp>
   <gram type="pos">adjective</gram>
   <gram type="degree">elative</gram>
   ...
</gramGrp>
...
```

Figure 14: Elative *aḥsan* 'better, best'

Another innovation of the Shawi schema is the <gram type="polarity"> construct. Arabic dialects often encode negation directly on pronouns: *māni* 'I am not'.

```
...
<form type="inflected">
   ...
   <orth>māni</orth>
   <gramGrp>
      <gram type="person">1</gram>
      <gram type="number">singular</gram>
      <gram type="polarity">negative</gram>
   </gramGrp>
</form>
...
```

Figure 15: Encoding polarity

An important feature of Semitic and therefore also Arabic verbal morphology is the system of what traditionally has been called verbal "stems" or "forms". They are used to specify semantic derivations such as causative, intensive, reciprocal, passive, reflexive etc. They are differentiated through particular morphological patterns and in Arabic studies usually identified by Roman numbers, e.g. *nizil* 'he went down' is "form I" or "class I". The causative thereof is *nazzal* 'he made (somebody/something) go down' which belongs into class II and is identified by reduplication of the second root consonant. In the VICAV dictionaries, these derivational classes are not grouped together according to their underlying roots, but each derived verb class is treated as a lemma in its own right. All verbs have a <gram> element with a @type="derivedVerbClass" attribute.

```
...
<form type="lemma">
   ...
   <orth>xtifa</orth>
</form>

...

<gramGrp>
   <gram type="pos">verb</gram>
   <gram type="derivedVerbClass">VIII</gram>
   <gram type="synRoot" xml:lang="ar-acm-x-shawi-vicav">xfy</gram>
   <gram type="diaRoot" xml:lang="ar">xfy</gram>
</gramGrp>
...
```

Figure 16: Encoding verbal 'stems'

Subclasses of nouns are specified by <gram type="subc">. The values used in the SHAWI Dictionary are collectiveNoun, diminutiveNoun, fraction, ordinal, pluralNoun, toponym and unitNoun.

3.3.2 Word class information

A problem in modelling non-European linguistic varieties is often that grammatical concepts have not been considered in standardisation efforts. This is even true of a rudimentary category such as word classes. The VICAV taxonomy in this regard is still a hotchpotch awaiting harmonisation. In dealing with these categories, we have been considering in particular two sources: the CLARIN Concept Registry and labels being used in Universal Dependencies. However, academic communities often have very specialised nomenclature. The dialectologists involved in our dictionary introduced a number of terms which in their specificity are not represented in existing lists. Examples of @type="pos" values used in the SHAWI Dictionary include:

- genitiveParticle: *hnīt* 'of, belonging to'
- existential: *bī* 'there is'
- pseudoVerb: *bidd-* 'want'
- presentativeParticle: *hāḏiyān-* 'here it is!'

Dedicated digital vocabularies of relevant terminology would be a major desideratum, particularly in the linguistics of lesser-resourced languages and smaller research communities such as ours. However, there remains much to be done in this respect.

4. Summary

In addition to introducing the SHAWI Dictionary, which is scheduled to go online as a beta version by the end of 2025, our paper was mainly meant to engage a broader lexicographic audience in the discussion of the issues at hand. Our attempt at modelling an Arabic dialect dictionary using the TEI Lex-0 framework has demonstrated that the system is both applicable to and efficient for our particular purposes, requiring minimal

overhead in its adaptation. The process also benefited greatly from close collaboration with the Lex-0 community, whose input helped align the project with broader community standards. TEI Lex-0's ongoing effort of establishing a baseline encoding and a target format to enhance the interoperability of heterogeneously encoded lexical resources has made it particularly suitable for the application to our data and will be valuable in its application to the older VICAV dictionaries which have not yet been converted. TEI Lex-0 promotes centralised standardisation ensuring an increased degree of interoperability in combination with a high degree of extensibility via ODD chaining, thus ensuring that the framework remains adaptable to the evolving needs of diverse lexicographic projects, especially for those dealing with lesser-resourced non-Indo-European languages.

5. Acknowledgements

This article was written as part of the project *The Shawi-type Arabic dialects* (P 33574-G), financed by the Austrian Science Fund FWF (2021–2027), and within the context of the WIBARAB project, which is funded by the ERC Advanced Grant 101020127 (2021-2026). WIBARAB stands for 'What is Bedouin-type Arabic?'.

6. References

Aldoukhi, R., Procházka, S. &. Telič, A. (2014-2015). Lehrbuch des Syrisch-Arabischen: Praxisnaher Einstieg in den Dialekt von Damaskus. Wiesbaden: Harrassowitz.

Budin, G., Majewski, S. & Moerth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative (jTEI)* 3. Available at https://doi.org/10.4000/jtei.522.

Frontini, F., Fahad A. Kh. & Romary, L. (2023). ISO LMF 24613-6: A Revised Syntax Semantics Module for the Lexical Markup Framework. Language, Data and Knowledge (LDK) 2023, pp. 316-321. Available at https://aclanthology.org/2023.ldk-1.31.pdf.

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

Moerth, K. (2024). Refining the Structure: Applying TEI Lex-0 to the Concise isiZulu-English Internet Dictionary. In P. Stöckle & S. Wahl (eds.) *Lexicography and Language Variation, Wiener Arbeiten Zur Linguistik.* Wien, Vienna University Press, pp.117-138.

Moerth, K, Procházka, S., Siam, O. & Declerck, T. (2013). Spiralling towards perfection: an incremental approach for mutual lexicon-tagger improvement. In *eLex 2013,* pp.225-242. Available at https://elex.link/elex2013/proceedings-2013.

Moerth, K. & Schopper, D. (2021). VICAV 3.0: Zooming in on Lexical Resources. In C. Katsikadeli, M. Sellner & M. Gassner (eds.) *Digital Lexis, and Beyond. Selected*

Papers from the Workshop „Digital Lexis, and Beyond" 45th Austrian Linguistics Conference 2019. Wien: Verlag der ÖAW.

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

Procházka, S. (2024). How solid is the linguistic basis for the Bedouin sedentary split used in the classification of Arabic dialects? In C. B. Ramos, J. Guerrero & M. B. Fernández (eds.) *AIDA Granada: A pomegranate of Arabic varieties*. Zaragoza: Prensas de la Universidad de Zaragoza, pp. 359–370. Available at https://phaidra.univie.ac.at/o:2108301. (19 July 2025)

Procházka, S. & Moerth, K. (2015). The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects. In M. Al-Hamad, R. Ahmed and H. Aloui (eds.) *Lisan Al-Arab: Studies in Contemporary Arabic Dialects, Proceedings of the 10th International Conference of AIDA*. Qatar University 2013. Vienna: LIT Verlag, pp. 209-218.

Rahtz, S. (2014). Advanced topics in ODD. *TEI Conference Workshop: An Introduction to TEI's ODD: One Document Does it all.* Oct 2014, Evanston, United States. 2014. https://inria.hal.science/hal-01767683.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In *Electronic lexicography in the 21st century. Proceedings of the Electronic lexicography in the 21st century (eLex 2019),* pp. 417-433. Available at https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf.

Tasovac, T., Romary, L., Banski, P., Bowers, J., Does, J. de, Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A. & Witt, A. (2018ff.). TEI Lex-0: A baseline encoding for lexicographic data. DARIAH Working Group on Lexical Resources. Available at https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Tasovac, T., Salgado, A., Costa, R. (2020). Encoding polylexical units with TEI Lex-0: A case study. In *Slovenščina 2.0. 8(2), pp.* 28–57. Available at DOI: https://doi.org/10.4312/slo2.0.2020.2.28-57.

TEI Consortium (2025). TEI P5: Guidelines for Electronic Text Encoding and Interchange, P5 Version 4.9.0. Available at www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.