MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„The WIBARAB Database (October 2021-October 2023): Data Modelling and Linguistic Analysis“

verfasst von / submitted by

Veronika Engler, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 676
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Arabische Welt: Sprache und Gesellschaft
Betreut von / Supervisor:	Lic. Ana Iriarte Díez, MA PhD

Table of Contents

Acknowledgements

First and foremost, I want to thank Ana Iriarte Díez for being such an engaged supervisor and for putting so much of her time and effort into this thesis. Without her comments, guidance, patience and support, I could not have written the thesis in this form. Similarly, I want to thank Stephan Procházka for providing me with the opportunity to be part of the WIBARAB project, for commenting on a draft of chapter 4, and for always having an open ear.

I would also like to thank Charly Mörth and Daniel Schopper for igniting the spark of digital humanities in me and for always making me feel welcome and appreciated in the database team. Special thanks go to Charly for commenting on a draft of chapter 3. Thank you to Stanley and Johanna for the fun we had with the TEI Enricher and everything else.

Thank you to Gunda, for listening to me and bearing with me at every step of the process and making me believe that this can be done. Thank you to all my friends and family for the encouragement and support. And thank you also to Urmel for keeping me fit and emotionally stable.

Finally, I want to thank you, Sebastian, for being the best partner imaginable. Thank you for taking care of me and making sure I didn't starve or dehydrate or go mad during the intensive writing phase. Thank you for proof-reading and sharing your wisdom. Generally, thank you for being my rock, for supporting me on every level and for making life beautiful.

List of Abbreviations

ACDH-CH – Austrian Centre for Digital Humanities and Cultural Heritage

APiCS – Atlas of Pidgin and Creole Language Structures

CSS – Cascading Style Sheets

DAD – Database of Arabic Dialects

DD (coordinates) – Decimal Degree

DMG (transcription) – Deutsche Morgendländische Gesellschaft

DMS (coordinates) – Degree/Minute/Second

ERC – European Research Council

HTML – HyperText Markup Language

ID – Identification

IPA – International Phonetic Alphabet

IT – Information Technology

ISO – International Organization for Standardization

L1-speaker – first language speaker

MA – Master of Arts

MSA – Modern Standard Arabic

ODD – One Document Does It All

PhD – Doctor of Philosophy

PHP – Hypertext Preprocessor

PI – principal investigator

SHAWI – research project "Die arabischen Dialekte des Shawi-Typs"

SQL – Structured Query Language

TAC – Tunisian Arabic Corpus

TEI – Text Encoding Initiative

TUNICO – research project "Linguistic Dynamics in the Greater Tunis Area"

TUNOCENT – research project "Tunisia's Linguistic *terra incognita*: an Investigation into the Arabic Varieties of Northwestern and Central Tunisia"

UTF-8 – 8-Bit UCS (Universal Coded Character Set) Transformation Format

VICAV – Vienna Corpus of Arabic Varieties

VSC – Visual Studio Code

WALS – World Atlas of Language Structures

WIBARAB – research project "What is Bedouin-Type Arabic?"

WP – work package

XML – Extensible Markup Language

XSD – XML Schema Definition

XSLT – Extensible Stylesheet Language Transformation

1. Chapter 1: Introduction

With this MA thesis, I intend to provide an example of how digital tools can enhance and facilitate research in the field of Arabic dialectology and linguistics in general. In order to do this, I documented the first two years of the development of the WIBARAB (**W**hat **I**s **B**edouin-type **ARAB**ic?) database, focusing on the data model and how it was customised according to the nature of our data on the one hand, and the requirements of our project's research questions on the other hand. The WIBARAB database will include approximately 125 features, mainly from phonology, morphology, and syntax, but also a selected set of lexical and phraseological items. It is being developed and built within the framework of the WIBARAB project. Due to the project's focus on (former) nomadic communities and their Arabic varieties, the linguistic features covered by the database are chosen for their salience in examining the Bedouin-sedentary split. Many of the varieties included in the database were selected on the same grounds, although also well-documented sedentary varieties were included for comparative purposes. Therefore, the WIBARAB database will not only be the first all-Arabic database of this size, but it will also be the first one to focus on the varieties spoken by Bedouins. The opportunity of taking part in the database's development and construction, and now of writing this thesis, was given to me through my work as part of the WIBARAB database team.

Due to its topic, this thesis is positioned on the interface of Arabic dialectology and digital humanities. Therefore, it aims to connect established research fields and modern technologies that bring new perspectives and possibilities and is intended to be a bridge between the more traditional approaches of Arabic dialectology and newer ideas and methods of the digital humanities. Hence, the thesis will not only serve as documentation of the database and as a trial run for the database, but it will also contribute to the growing body of literature on digital humanities tools in (Arabic) linguistics.

1.1. The WIBARAB Project[1]

In this section, I will first introduce the WIBARAB project, its Work Packages, the WIBARAB team members and their responsibilities in the project. Then, I will give a short overview over the historical and linguistic dimensions of the Bedouin-sedentary split (section 1.1.1.) and a general description of the WIBARAB database (section 1.1.2.).

The WIBARAB (**W**hat **I**s **B**edouin-type **ARAB**ic?) project is funded by the ERC Advanced Grant 101020127. Its duration is five years, from October 2021 to September 2026, and it is hosted by the Department of Near Eastern Studies at the University of Vienna in cooperation with the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) at the Austrian Academy of Sciences. The WIBARAB project investigates the linguistic and socio-historical realities behind the millennia old dichotomy of nomadic and sedentary communities in the Middle East and Northern Africa. WIBARAB adopts a combination of methods grounded in Arabic dialectology, general linguistics, sociolinguistics, and historical interpretation. The project comprises six Work Packages (WP):

WP-1 "New linguistic data from the field" generates new data by providing comprehensive grammatical descriptions of three Bedouin-type varieties which are hitherto under-investigated or hardly known: the Bani Ṣaxar in Jordan, the Zʕir and Šawiya in Morocco, and the Rašāyda in Saudi-Arabia, Kuwait, and Sudan. The generation of new data is primarily accomplished by the project's prae-doc researchers.

WP-2 "Sociolinguistics" explores language ideologies and attitudes towards Bedouins and how 'Bedouin speech' is perceived today. For example, one WIBARAB researcher investigates

[1] This section and especially subsections 1.1.2.1 and 1.1.2.2 are based in considerable parts on the WIBARAB website: https://wibarab.acdh.oeaw.ac.at/ and the project's proposals (Procházka, 2020 B1 & B2).

the widely believed idea that Bedouins speak "purer" forms of Arabic than sedentary communities.

WP-3 "Historical backgrounds" investigates the history of Bedouin migrations and will illustrate how history can provide proxy data for diachronic linguistic data and vice versa. An example for this is the Rašāyda tribe in Sudan whose Arabic clearly differs from other Sudanese varieties, a fact that can be explained by their history, i.e. their migration to Sudan from Najd.

WP-4 "Linguistic typology" addresses the topic of linguistic typology and aims at a comparative survey of grammatical and lexical features of Bedouin-type Arabic. For this purpose, WP-4 encompasses the construction of the WIBARAB database, which is the object of this thesis. Thus, WP-4 is described in more detail in section 1.1.2.1.

WP-5 "Language and dialect contact" is dedicated to language and dialect contacts historically and in the present by focusing on contact-induced linguistic changes which are most likely the outcome of Bedouin migrations. This means researching questions such as how intra- and interdialectal contact has influenced certain varieties, what role other languages play in language changes and why there seem to be regional differences in the way and extent contact has induced linguistic changes.

WP-6 "Historical linguistics and language change" will draw on the findings of all other work packages to present an answer to the overarching question of whether and to what degree difference in lifestyle, i.e. nomadic versus sedentary, is mirrored in the past and present of the Arabic language, and which factors – social and/or historical – have led to either the retention or the abandonment of typical linguistic features related to one group or the other.

The current[2] WIBARAB team consists of specialists in Arabic linguistics from the University of Vienna: the principal investigator (PI), one postdoctoral researcher, and four PhD candidates. In addition, two MA students work on transcription and administration and two work on the database, collecting data and feeding it into the database, and operate as a link to the digital humanities specialists from the Austrian Academy of Sciences whose knowledge and work also benefits WIBARAB. The Vienna-based team is supported by research partners in the Arab countries.

The PI, **Stephan Procházka**, holds the chair for Arabic Studies at the University of Vienna and works on comparative and syntactic aspects as well as on historical linguistics. He is mainly responsible for WP-5 and WP-6, but also acts as a supervisor of the PhD theses (WP-1 and WP-2). **Ana Iriarte Díez**, postdoctoral researcher in linguistics, works on language and dialect contact and historical linguistics and language change (WP-5, WP-6). In addition to her research on Bedouin Arabic varieties in Lebanon, especially in Karantina, she supervises the design and feeding of the WIBARAB database (WP-4) and is thus part of the database team. **Claudia Laaber**, prae-doc researcher, works on perceptions of and language ideologies towards Bedouin Arabic (WP-2**). Antonella Torzullo**, prae-doc researcher, works on the description and analysis of the Arabic dialect of the Bani Ṣaxar tribe in Jordan (WP-1). **Maria Rebecca Zarb**, prae-doc researcher, investigates the dialect of the Rašāyda Bedouins in Saudi Arabia and Kuwait (WP-1). **Terlan Djavadova**, prae-doc researcher, deals with the Bedouin dialects of the Zʕir and Šawiya in Morocco and their influence on the Moroccan Koiné (WP-1). **Sara Fatihi,** MA student in general linguistics (University of Vienna), assists Terlan in her research on the Bedouin dialects of the Zaer and Chouia (Morocco). **Philipp Marc Hessabi**,

[2] In this thesis, I document the first two years of the project, thus by "current" I refer to September 2023 when this chapter was written. For practical reasons, I do not mention former team members or international collaboration partners. However, information about them can be found on the WIBARAB website: https://wibarab.acdh.oeaw.ac.at/.

MA student, assists the project with general administrative tasks. **Stanley Kochem** and I, MA-students, work on the conceptualization and implementation of the database as well as the collection and input of data on features of Bedouin-type Arabic from already published sources (WP4). While Stanley Kochem writes his MA thesis on the Arabic of the Rašāyda of Sudan, supplementing Maria Rebecca Zarb's research, my focus lies solely on the database, initially working on its conceptualisation and now the day-to-day running of the data input process.

The second part of the team is made up of our project cooperation partners at the ACDH-CH at the Austrian Academy of Sciences. First, we work with **Karlheinz "Charly" Mörth**, who is the head of the ACDH-CH unit. He is an expert in the interface between linguistics and digital language resources doing research in the fields of virtual research environments and language-related standards. **Daniel Schopper** is a senior advisor and head of the working group "Data". He is also the MS Teams and Sharepoint administrator in our project. Finally, **Omar Siam** is a senior advisor and systems administrator at ACDH-CH. He works as a technical supervisor for the Vienna International Corpus of Arabic Varieties (VICAV) infrastructure.

While all the team members contribute invaluably to the WIBARAB project as a whole, not everyone works on the conceptualisation of the WIBARAB database. Thus, when I refer to 'the database team' in this thesis, I refer to the part of the team that was particularly involved in the process of data modelling and the implementation of the database. This core database team currently consists of Stephan Procházka, Charly Mörth, Ana Iriarte Díez, Daniel Schopper, Stanley Kochem and me.

1.1.1. The Bedouin-Sedentary Split: Historical and Linguistic Dimensions

Now, I will give a short overview of the historical and linguistic dimensions of the Bedouin-sedentary split. For millennia, nomadic and sedentary communities have lived side by side in the Arab world. The coexistence of these two groups has long been an essential factor with regard to the history, economy, social structure and (linguistic) identity of the Middle East and

North Africa. The Bedouin tribes[3] not only played an important political and economic role in the region, but they are also crucial for the history of the Arabic language.

On the one hand, they shaped the diachronic development of Arabic and its spread throughout the Middle East and North Africa as a result of several waves of migration, starting with the expansion of Islam in the 7th century. These movements of populations occurred not only as migrations on a large scale over great distances such as the 7th century movements from the Arabian Peninsula to North Africa and Spain in the West and Iran in the East, or the 11th century migrations of the Bedouin Banu Hilal tribe into North Africa (Benkato, 2019:3). Rather, they also consisted of movements on a smaller scale, for instance when Bedouin speakers from the centre of the Arabian Peninsula moved to repopulate Baghdad and Lower Iraq after the Mongols left the city in 1258, or when Bedouin tribes migrated from Najd into the Gulf littoral (Holes, 1991).

On the other hand, Bedouins and their speech had an important part in the standardisation of the Arabic language, since they were considered by the traditional Arabic grammarians as speaking the purest and most eloquent form of Arabic. McNeil (2019:31) shows that "[t]he earliest Arab grammarians used desert Bedouins as their linguistic informants, claiming that the speech of the urban population, including their own, was corrupted." This view is supported by the accounts of many scholars, e.g. Al-Farabi in his *kitab al-ḥurūf* (Al-Farabi, 1969:147), Ibn Faris (Ech-Charfi, 2018:74,) and even the famous historian Ibn Khaldun (Bagatin, 2019).

[3] I am aware that the English term 'tribe' can be highly problematic, especially in the context of research traditions that were closely connected to colonialism. However, in the case of Arabic speaking Bedouins, the English term is the correct translation of the Arabic terms *qabīla* or *ʕašīra* (Wehr, 1979: Q-B-L/ʕ-Š-R), both of which are endonyms of (former) nomadic Arab communities. Due to the fact that belonging to a certain *qabīla/ʕašīra* can be an important part of the identity for many Bedouins (and for sedentary populations in some parts of the Arab world), the concept 'tribe' is a meaningful social concept in this context. Therefore, I decided to use the term 'tribe' in accordance with the conventions of Arabic Studies.

As we can see, the (linguistic) Bedouin-sedentary split is not a Western invention considering traditional Arab grammarians' reliance on Bedouin, rather than sedentary, speakers as linguistic authority. These language ideologies, despite their remote date of origin, can still be found among Arabic L1-speakers[4] today. This notion of Bedouin Arabic as very pure and original was taken up by Western scholars who take these "anecdotes of Arab grammarians concerning the linguistic purity of Bedouin speech" (al-Sharkawi, 2010:74) as literal to a varying degree (al-Sharkawi, 2010:74). What is more, it persists in the field of Arabic dialectology to the extent that still, Bedouin dialects are often considered particularly 'archaic' and 'conservative': "Generally, it is claimed that Bedouin dialects are more conservative, sedentary dialects more innovative. This is because sedentary communities particularly urban communities are more likely to be open to new linguistic forms, to come into contact with people from other communities with whom they have to communicate, and thus avoid the more salient features of their dialect" (Watson, 2011:869).

The Bedouin-sedentary split persists both in Arab and Western dialectology. 'Bedouin' and 'sedentary' have long been considered to be the main categories for the classification of Arabic in the literature. For instance, Heath (2002:1) writes that "[t]he major classificatory division of the vernaculars spoken in the Arab world is between originally sedentary and Bedouin types". This is supported by Younes (2018:265): "Within the field of Arabic dialectology, two big types of dialects have been identified a long time ago and still serve as a basis for the classification of different Arabic vernaculars. These two groups are the sedentary dialects and the nomads' dialects" and Watson (2011:868): "the major classificatory division of dialects in the Arab world has traditionally been seen in terms of bedouin versus sedentary". Some

[4] Speakers who speak Arabic as their first language.

scholars have added a subdivision to the dichotomy by further dividing the category of ‘sedentary’ dialects into ‘rural’ and ‘urban’ (Cadora, 1992).

However, despite its long tradition, this classification is (still) based on very few linguistic facts. Prior studies which have attempted to find linguistic features that are shared by all Bedouin-type dialects – e.g. the voiced reflex of *q which is regarded to be the shibboleth of Bedouin-type Arabic – have not provided convincing results (Procházka, 2020 B1). Therefore, the WIBARAB project aims at filling a significant gap in the research by answering its main question whether the difference in lifestyle is mirrored in the language and backing up this answer with solid data provided by the WIBARAB database.

1.1.2. The WIBARAB Database

The main topic of this thesis is the WIBARAB database, which constitutes an important part of the WIBARAB project. In this section, I offer a first general introduction to the database and its aims. The database will bring together a great amount of linguistic data, including a large number of Arabic varieties that have traditionally been classified as Bedouin-type dialects in order to enable linguistic analysis by WIBARAB researchers, the wider Arabic dialectology community, and researchers in general linguistics. For the purpose of consistency and maintaining scientific standards, we chose to adopt the transcription system of Encyclopedia of Arabic Language and Linguistics (Versteegh, 2006) with a few modifications for the WIBARAB database. The database is part of WP-4[5], and its primary objective is to facilitate research on WIBARAB’s research topics and finally to help answer some of the project’s research questions. Therefore, in the following sections I will first describe WP-4 in more detail

[5] See section 1.1 for a description of WIBARAB’s Work Packages.

and then sketch out WIBARAB research topics and questions to which the database will be relevant.

1.1.2.1. Work Package 4

One of the project's six work packages (see section 1.1 for the others), WP-4 is focused on the WIBARAB database. Its main aim is modelling and constructing a database which enables users to carry out efficient comparisons across vernacular varieties of Arabic. The database will focus mainly on phonological, morphological, and syntactical features, but will also include – to a limited extent – phraseological and lexical features. The database will be developed in close collaboration with (and hosted at) the ACDH-CH and incorporated into the Vienna Corpus of Arabic Varieties (VICAV) in order to benefit from the existing infrastructure. This made adopting an XML-based approach an obvious choice.[6] The goal was to produce TEI-compatible data and metadata that allow for long-term preservation, reproducibility of research results, follow-up research, and reusability in other contexts by adhering to relevant and state of the art standards in text technology (Procházka, 2020 B2).

1.1.2.2. Research Questions and Topics Expected to Be Answered by the Database

The principal aim of the database is to help answer WIBARAB's research questions and shed light on some of its main research topics. The overarching research question is: **to what degree is the difference in lifestyle between Bedouins and sedentary people mirrored in the history and the present of the Arabic language?** To help answer this question and investigate all its aspects, several sub-questions have been posed, and research topics have been established in the project's proposal. Some of these questions and topics are of a linguistic nature and some are mainly concerned with language ideologies and attitudes, or of a historical character. The database's purpose to facilitate the project's research is limited to the linguistic questions, as the database does neither include data on how language is perceived, nor does it contain a

[6] See section 1.3 on why this is the case.

distinct diachronic dimension. In the remainder of this section, I will give a brief overview of the research topics and questions that are expected to be answered with the help of the database. For a more detailed description of the database's analytic functions that will accomplish this, see section 3.3.

1) **To what degree is the social dichotomy nomadic versus sedentary mirrored in the language?** In particular, which social and/or historical factors led to either the retention or the abandonment of typical linguistic features related to one or the other linguistic group?
2) **How conservative is Bedouin-type Arabic really?** Often, the Arabic varieties spoken by Bedouins are considered to be more conservative and closer to Classical Arabic than the varieties spoken by sedentary communities, not only by scholars but also among L1-speakers.[7] If this is really the case, a hypothesis that the project intends to prove or disprove is that the still existing, close-knit tribal structures are an important factor in social and linguistic dynamics, and thus have a considerable impact on preserving a tribe's dialect. This assumption goes hand in hand with the question whether, or to what extent, a tribe can be considered a linguistic unity (Procházka, 2020:3f B2).
3) **New linguistic data from three different regions.** The provision of new data about varieties spoken by Bedouins is especially important for mainly two reasons: on the one hand, the scarcity of grammars and description of Bedouin dialects in comparison with sedentary dialects constitutes a gap in Arabic dialectology. On the other hand, the advancing sedentarisation of Bedouin communities and the disappearance of their lifestyle around the Arab world creates an even more pressing need to document the original dialects of these communities (Procházka, 2020:4f B2).

[7] See section 1.1.1 for an introduction to some of the literature.

4) **Comparative survey of actual (or alleged) grammatical and lexical features of Bedouin-type Arabic.** This comparative survey is intended to investigate the question what makes a dialect 'Bedouin-type'. By comparing large amounts of data, including the data collected in the project in addition to previously published data, we hope to find meaningful bundles of linguistic features that provide us with a substantial and fact-based answer to the aforementioned question. In order to find these bundles, we intend to look both at features that are traditionally considered 'Bedouin' as well as broadening the scope to include features which have previously been to a large part ignored, such as features pertaining to phraseology and syntax. In short, the project intends "to provide additional substantial criteria for better understanding the nature of bedouin-type Arabic—most probably a picture that is much more diversified than usually presented because bedouin-type Arabic is the result of long and still ongoing processes of convergence and divergence" (Procházka, 2020:7 B2).
5) **Language and dialect contacts in the past and the present.** Due to the historical expansion of the Arabic language over large areas, contact is central to understanding the development of Arabic and its many varieties. To see which role Bedouin dialects played in this, the project intends to compare several dialect areas that were in contact with Bedouin communities to a varying degree regarding intensity and duration, and to investigate how this contact influenced linguistic realities. In this investigation, some phenomena that are also discussed in general linguistics will be addressed, for instance, the evolution of linguistic areas, the question of whether contact results in greater complexity or more simplicity (Trudgill, 2009:175-181), or shared grammaticalization paths and typological changes (Heine & Kuteva, 2005) (Procházka, 2020:7f B2).

1.2. Digital Resources for Linguistic Analysis of Arabic

In this section, I will introduce some types of available digital resources in the fields of Arabic studies and in particular Arabic dialectology.

The importance of digitisation and the use of digital tools has been growing in the last decades, not least in the humanities. Not only do technical tools enhance the way research can be

conducted within traditional methodological approaches of the humanities, e.g. by making it possible to process large amounts of data much more efficiently than before, but additionally, they allow for the development of new methodologies and new fields of research (Jannidis, 2017: 14). Therefore, it is not surprising that an essential part of the WIBARAB project lies in its intersection with the field of digital humanities: the WIBARAB feature database. Ultimately, the database and the linguistic analyses which it will enable will provide answers to the project's research questions as well as substantiated empirical data to back the project's conclusions.

The growing importance of digitisation in the humanities is evident in the fields of Arabic studies in general and Arabic dialectology in particular, judging by the range of digital resources that have been developed in recent years. While it would be impossible to enumerate all these digital resources, in the following section I will give an overview of the different types of available digital resources and briefly describe some that are most relevant to Arabic dialectology,[8] i.e. most of those which deal with one or more spoken variety of Arabic.

1.2.1. Types of Available Resources

1.2.1.1. Atlases

While there are quite a few Arabic dialect atlases (e.g. Behnstedt & Woidich, 2010, Behnstedt, 2016, Behnstedt & Woidich, 1985, Behnstedt & Geva-Kleinberger, 2019, Behnstedt, 1997), they usually are not (yet) digital tools, but rather come in the format of books. However, some

[8] For a more extensive but also not comprehensive list of digital Arabic research tools see VICAV https://vicav.acdh.oeaw.ac.at/#map=[biblMarkers,_profiles_,]&4=[textQuery,vicavArabicTools,ARABIC%20TOOLS,open].

of their functions would be greatly enhanced in a digital format. This can be seen from the example of the World Atlas of Language Structures (WALS).

1.2.1.1.1. World Atlas of Language Structures (WALS)

The WALS is "a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors" (Dryer & Haspelmath, 2013:n.pag.) and its online version is hosted by the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. The WALS includes over 2,500 languages and around 200 linguistic features which are described in chapters. The query function of the WALS is connected to a map which displays the query's results in their geographic distribution. Additionally, the WALS includes a bibliography containing more than 7,000 items.

Although the WALS is unfortunately not very detailed with respect to Arabic varieties, it remains a very useful tool for general linguistics, and it served as an inspiration for the design of the WIBARAB database and its functions in the early conceptualisation phase. Thus, the WIBARAB database will not only include a map to visualise the data's places of collection, but also fully formatted chapters for each feature where a more detailed description of the feature and its values will be provided within the context of the existing literature. Therefore, we hope that, eventually the database can function as a tool complementary to the WALS by making data on a multitude of Arabic varieties accessible for general linguists as well as Arabic dialectologists.

1.2.1.2. Corpora

A subfield of linguistics that has gained importance with the rise of digitisation is corpus linguistics, since the new technologies allow for the efficient processing and analysis of large text corpora. In the following, I will describe a few of the Arabic corpora available to the date.[9]

1.2.1.2.1. ArabiCorpus

ArabiCorpus is a text corpus of mostly Modern Standard Arabic (MSA) containing around 100 million words and hosted by Brigham Young University in Utah, USA. In addition to the MSA material, it contains a small amount of data of medieval Arabic and Egyptian Arabic. A major advantage of arabiCorpus is its user-friendliness that makes it accessible to all users, regardless of whether they are trained linguists or not. This sets arabiCorpus apart from several other existing MSA corpora (Parkinson, 2019).

The corpus is mainly based on newspaper texts with small amounts of other non-fiction texts and literature. For the queries, users can use either Arabic script or a one-to-one transliteration system of Latin characters which is explained on the site. In the interface, the user can narrow down their searches by specifying for instance the parts of speech or the part of the corpus one wants to search. ArabiCorpus can be used to search for:

- specific words and forms
- morphological and syntactic structures
- distinct word senses of individual forms
- collocations and idiomatic uses of words by looking at the contexts in which those words appear
- overall frequencies and subfrequencies of words and forms

[9] For a more extensive inventory of Arabic corpora, see McNeil, 2019.

- comparisons of frequencies and usages between different countries and genres (Parkinson, 2019:28).[10]

In order to use arabiCorpus with all its functions, one needs to create a user account, which is free of charge and serves statistical purposes. As mentioned above, one of the great strengths of arabiCorpus is its accessibility which is highlighted and promoted by the detailed instructions given on the web interface as soon as one logs in. These instructions not only give general information about the corpus and about the searches that can be done, but they also contain a tutorial and explanations of how to analyse the results of a search. However, for technically more advanced users, arabiCorpus also allows for more complex searches with regular expressions. So if, for instance, a user wants to do more detailed searches than the filters which are provided on the web interface allow, they can create their own search filters.

While arabiCorpus is an important tool for Arabic linguistics and corpus linguistics in particular, it does have limitations. Apart from the fact that, compared to corpora of other languages, it is relatively small, one of its biggest downsides is that it is not lemmatised or part-of-speech-tagged, unlike, e.g. many English language corpora. This prevents a number of useful searches. The other major drawback is the 'unbalanced' nature of the corpus, i.e. the vast majority of the texts being newspaper texts. Therefore, only certain topics are covered and there is not enough representation of other genres, topics, or, for instance, spoken text.

1.2.1.2.2. Tunisian Arabic Corpus (TAC)

The TAC currently consists of almost 3,000 texts, comprising around one million words. The project's goal is to reach a total of 4 million words and to eventually use the corpus to create a Tunisian – English dictionary (McNeil, 2019:37). The texts come from different materials: 1)

[10] For more detailed information on the functions of arabiCorpus and how it is programmed, see Parkinson, 2019 and https://arabicorpus.byu.edu/.

traditional written sources, e.g. folklore, collections of proverbs, or screenplays; 2) new written sources, e.g. blogs or postings on the internet; and 3) transcriptions of audio sources, e.g. radio broadcasts or podcasts (McNeil, 2019:38f). The categories covered in the TAC include blogs, literature, television (drama), internet forums, folk tales, and more (McNeil & Faiza 2010-).

In the search mask, the user can select the category(s) in which to search, but other than that, the search functions of the TAC are not as elaborate as the functions of arabiCorpus, for example. In fact, they are restricted to three different types: 'exact', 'stem' and 'RegEx' (regular expression). 'Exact' means that the search returns only the exact word the user enters into the search box, without any variation. For example, my search for the word يمشي returned 361 results in 160 texts, showing only the exact word that was searched. Using the search-type 'stem', the user has to enter the root radicals of a word in order to get all different forms that are formed with those radicals. For example, my search for مشي returned 2078 results in 544 texts, including in its results forms like يمشي, تمشي or مشيت. While for 'exact' and 'stem' searches, the word has to be typed in in Arabic script, the regular expression searches only work with Latin characters according to a specific transliteration system that is given in the instructions of the corpus. These regular expression searches allow for more sophisticated searches – as long as the user knows how to use regular expressions. For instance, the search for [sS]Hb will return both S-Ḥ-B and Ṣ-Ḥ-B. Generally, the results are always given in a sentence to make the context in which the word appeared visible.

The limitations of the TAC are similar to those of arabiCorpus, if not more pronounced. The corpus is still quite small, and the topics covered are by their very nature restricted to the few areas where the Tunisian vernacular is written, resulting in a not entirely balanced corpus. Additionally, it is, like arabiCorpus, not lemmatised and part-of-speech-tagged.

1.2.1.2.3. TUNICO

The TUNICO corpus is part of the TUNICO project which was conducted jointly by the University of Vienna and the Austrian Academy of Sciences. The website is hosted by the ACDH-CH at the Academy of Sciences. In contrast to arabiCorpus and the TAC, the TUNICO corpus consists not of written, but spoken texts which were transcribed for the corpus. The texts were initially more than 30 hours of recordings transcribed into 24 digital documents. The

speakers who were recorded are residents of the Greater Tunis area, who were born and raised there and were 35 years old or younger at the time of recording. They come from different social backgrounds, and the numbers of female and male speakers is balanced. The corpus is closely annotated with over 90% of the words tagged and tokenised, it contains over 140,000 tokens. All the text is transcribed in Latin characters, using for the most part the DMG transcription system. The special characters are provided next to the search boxes.

TUNICO has two basic functions: the corpus and a dictionary. Both the corpus and the dictionary sections contain the subsections 'overview', 'search', and 'help.' The latter option provides basic information on how the searches work. For example, it explains Wildcards and regular expressions that can be used in the search. In the corpus the searched item is returned in its immediate context with the possibility of opening the whole text both in TEI/XML format and readable format. When moving the cursor over the text, each word brings up its dictionary entry, providing the part of speech, the root, and translations into English, German and French. Additionally, the original audio recordings of the texts are provided. The dictionary includes not only data from the corpus, but also from additional materials like complementary interviews and some publications. While the dictionary is part of the TUNICO website, it is also accessible through the VICAV interface (see section 1.2.1.5.1). The dictionary can be searched by roots in addition to words and regular expressions.

The same reasons that on the one hand make the TUNICO corpus stand out among the described corpora – that its items are tokenised and tagged, and that the texts are originally spoken Arabic which were then transcribed – are on the other hand the cause for its main limitation: it is by far the smallest of the three corpora and can only be representative for a very specific group of speakers. Another fact that can be seen as both an advantage and a limitation is the precision of the transcription system. While it allows the user to know the exact pronunciation of the words, it can somewhat impede the searches. For example, when one searches the corpus for the word عربي with the transcription *ʕarabi*, there are no returns. Only when one considers the exact pronunciation and searches for *ʕaṛabi*, the search engine is able to find the word. This makes the search function of the corpus quite a challenge for the unexperienced user.

1.2.1.3. Dictionaries

While there are several online dictionaries for MSA, the dictionary I describe in this section is one of only a very limited number of online dictionaries that deal with spoken varieties of Arabic.

1.2.1.3.1. Lughatuna: The Living Arabic Project

Lughatuna, which translates to "our language" in English, is a website that contains several dictionaries: Classical Arabic/MSA-EN, Levantine AR-EN, Levantine AR-AR, Egyptian AR-EN, Maghrebi AR-EN, Gulf AR-EN, Iraqi AR-EN, Sudanese AR-EN, and Yemeni AR-EN. While there are a number of websites which combine some MSA and Classical Arabic dictionaries, Lughatuna is one of the very few that also include several dialect dictionaries. The Lughatuna dictionaries include a query function where the user can select the dictionary(s) that should be searched as well as the categories that should be returned (root, word, meaning, examples). The fact that one can search an item in all these varieties simultaneously provides an efficient way to make quick cross-dialectal comparisons on the lexical level.

In addition to the dictionaries, Lughatuna also provides various learning materials. Generally, the website seems to be targeted to Arabic learners rather than linguists, for unlike all the other resources presented here, Lughatuna was not created in an academic framework but – as the Tools & Technology section on VICAV put it – by "an enthusiastic hobby lexicographer" (VICAV, n.d.). Although this is somewhat noticeable in the general tone of the website, the dictionaries are still a useful resource, especially for learners of Arabic. However, the website's biggest limitation as a scientific resource is also connected to this and the resulting lack of transparency regarding the sources of the data, since there are no clear references to where the author of the website found the data used in the dictionaries. Although there is the possibility for the user to rate translations which may go some way to assure quality, this hardly compensates for the fact that there is no way to verify the author's sources.

1.2.1.4. Databases

Databases come in all forms and shapes. Very generally, they all store and make available logically connected data.[11] In the following, I introduce Alexander Magidow's Database of Arabic Dialects.

1.2.1.4.1. Database of Arabic Dialects (DAD)

The Database of Arabic Dialects is "a project to collect the vast amount of published and unpublished data on Arabic dialects in a searchable electronic format" (Magidow, 2015:n.pag.). It was created by Alexander Magidow, who was an associate professor of Arabic at the University of Rhode Island. This database includes data on a few morphological features, namely demonstratives, interrogatives, pronouns and verb affixes, for around 500 varieties. The database is targeted at researchers and students of Arabic. The web interface has four main search functions that visualise the data which is entered in characters of the International Phonetic Alphabet (IPA). The 'map' and 'list' views allow the user to view all publicly available data in two different formats. Additionally, the user can select dialects and see full paradigms in the 'paradigm view'. Finally, the 'cross-search' view offers the possibility to look for several forms at the same time and, for instance, to check whether dialects that have form x also have form y. Furthermore, in the section 'info lists,' the database includes lists of the tags and the references used in the database in addition to the dialects that are mentioned and the contributors who inputted data. Generally, the project is open to contributions of either data or code.

While Magidow's Database of Arabic Dialects definitely represents a contribution to the field by making information on Arabic dialects easily accessible and comparable, its main limitation is the relatively small amount of data it includes. Moreover, the database does not seem to get updated anymore, i.e. the amount of data will probably not increase over time. Additionally,

[11] See section 1.3 for a more in-depth explanation of databases.

although some basic instructions as to how to use the database are available, including a video tutorial on YouTube, it is still not very user-friendly, e.g. because queries have to be typed in in IPA without the website offering the symbols that are missing from a standard keyboard, while the majority of resources discussed in this section do in fact offer ways of entering necessary non-standard symbols.

1.2.1.5. Combined Resources

The following is a resource that features more than one purpose or function.

1.2.1.5.1. Vienna Corpus of Arabic Varieties (VICAV)

The Vienna Corpus of Arabic Varieties is a joint project of the University of Vienna and the Austrian Academy of Sciences. The VICAV infrastructure is hosted at the ACDH-CH. While VICAV's name rightly suggests that it is a corpus, this is not its only function. It is described on the website as "an international endeavour aiming at the collection of digital language resources documenting varieties of spoken Arabic. While in principle being open to any type of text, we currently focus on bibliographical data, on so-called language profiles, standardised feature lists and lexical data" (VICAV, n.d.). This means that the VICAV infrastructure has several functions, the most important of them being its dictionaries, its bibliography and its language profiles. VICAV contains five dictionaries – Tunis (TUNICO), Damascus, Cairo, Baghdad and MSA – which can be queried in different ways. In addition to searching for a specific word in either Arabic or English, it is also possible to search the dictionaries for parts of speech or specific verb forms (e.g. all form IV verbs) or combine the queries (e.g. search for the root 'k-t-b' in combination with 'verb' in order to find all verbs containing this root). It is also possible to search for simple regular expressions. The bibliography is quite large and data entry is an ongoing effort. The entries are tagged so they are connected to the map and additionally the bibliography can be searched. The language

profiles are descriptions of Arabic varieties that contain general information and a short research history.[12]

Besides being a resource for the Arabic language, VICAV also "has a strong methodological component focusing on the development of digital data and tools. A large part of these efforts is concentrated on modelling data and providing documentation for developmental and pedagogical purposes" (VICAV, n.d.). Thus, through its use, users can not only learn about Arabic but also about text technology and digital standards. Furthermore, it is open to contributions from anyone who wishes to cooperate, and all the technologies and data used are open source and open access, meaning that anyone can reuse both the data and the code.

1.2.2. Limitations of Available Resources: The Need for the WIBARAB Database

As we can see, although the number of digital linguistic resources for Arabic is growing, there are still large gaps to be filled. While each resource provides a valuable and important effort towards the availability of enough data and ways to linguistically analyse a language as rich and diverse as Arabic, it should be clear from the discussion in the previous section that each resource has its limitations. The most common limitations found in the examined resources are their narrow scope, i.e. small amount of data, their reduced scope in terms of Arabic varieties or of grammatical features. In order to fill the aforementioned gaps, the WIBARAB database aims to include a large number of both varieties and linguistic features, thus broadening the range of research questions that could potentially be answered through it.

[12] For more detailed information on the language profiles see section 3.3.2.6.

A limitation which is closely connected to the issue of a narrow scope, and which arose several times among the above-mentioned resources, was the fact that many of the resources contain data that is in some way 'unbalanced', sometimes pertaining to the topics and genres covered and sometimes to the written or spoken nature of the language. Although the WIBARAB database will not primarily be a corpus where this issue mostly occurs, during the data conceptualisation the WIBARAB team aimed to create as 'balanced' a resource as possible. On this line, we intend to (1) include and make use of both fieldwork data and data from academic sources, (2) include a vast yet geographically balanced number of Arabic varieties, and (3) include linguistic features from different areas like phonology, morphology, syntax, and to a lesser extent, lexicon.

Another limitation prevalent in currently available resources lies in the fact that most of them have an independent nature and do not connect to other resources. While the VICAV infrastructure incorporating the TUNICO dictionary represents an example of interconnection between resources, the WIBARAB database aims to take this further by building on some VICAV materials and finally being integrated into the VICAV infrastructure too. This will be possible thanks to the TEI-based structure of the database.[13]

Finally, and most importantly, what becomes especially clear after reviewing the resources, is that the vast majority of the Arabic dialects included are generally dialects of big cities, which in most cases correspond to traditionally sedentary populations. In order to research the linguistic implications of the Bedouin-sedentary split, it is necessary to be able to access a larger amount of data on varieties that are traditionally classified as Bedouin-type. While this is not necessarily a limitation of the existing resources for many other research questions, it is

[13] For more on this, see chapter 3, especially section 3.3.

crucial for WIBARAB's aim of investigating the Bedouin-sedentary split. This makes the necessity of the WIBARAB database all the more evident.

1.3. Technologies and Mark-up Languages Used for Linguistic Databases

After the discussion of several available types of resources, their advantages and their limitations in the previous section, I will now introduce the technical specificities behind some of them, with a special focus on text encoding and mark-up languages, which are especially relevant for the WIBARAB database.

When looking at the different digital resources for Arabic, the technologies used to build them are almost as diverse as the purposes for which they can be used. A good example for this is arabiCorpus, which uses the programming language Perl for processing its data and query matching, PHP for its web interface, and MySQL for saving immediate search results to increase efficiency when users want to access their previous searches (Parkinson, 2019:22). While the TAC and DAD use Python/Django, both TUNICO and VICAV are built from XML based TEI documents. Most of the above mentioned digital resources can be defined as databases following Jannidis's definition of databases as "a logically connected set of data which is physically stored" (2017:109, my translation). Generally, there are several types of databases, two of which I will briefly describe in this section to give the reader an idea of the diversity of technologies used for (linguistic) databases and digital tools.

The most commonly known databases are relational databases. An example from the field of Arabic dialectology which I have already introduced above is Alexander Magidow's (2015) Django-driven relational database. Jannidis (2017:111) defines relational databases as data fields and tables that correlate to each other and can be assessed using commands. Relational databases offer high speed processing and relatively flexible data structures. Most systems for relational databases are based on the programming language SQL (Structured Query Language) which is used to manage and query the data (Jannidis, 2017:117).

The second database type I will introduce here is the XML based database. Since we opted for an XML based database (more precisely a TEI database, see section 1.3.1.1.) for the WIBARAB project, this is the most relevant database type for the purpose of this thesis. The advantage of an XML database is that it is very flexible and can describe complex objects in a highly nested structure which is not possible in, e. g., a relational database (Jannidis, 2017:127). This kind of database falls under the category of semi-structured data and is often used in the humanities to encode and store large text corpora and digital text collections. Pierazzo defines text encoding as "the practice by which explicit codes (or tags) are added to a text in order to make some features of the text itself explicit, with the goal of making them processable by some computerized application" (Pierazzo, 2015:307). For the encoding of texts, markup languages are used. Usually, markup languages consist of a syntax, i.e. rules which specify how to encode, and vocabulary, i.e. tags which are assigned a specific meaning, in order for the markup to be understandable to others.

In the proposal for the WIBARB project, Stephan Procházka stipulates that "[t]he implementation of the database will be accomplished adhering to relevant and up-to-date standards in text technology. Data creation will be performed in a manner that will ease reproducibility of research results, follow-up research, and reusability in other contexts. The project will build on existing infrastructures, reusing tools of the VICAV system – which makes adopting an XML-oriented approach an obvious choice. The goal is to produce TEI compatible data and metadata for exchange and long-term preservation" (2020:11 B2). Since the documentation of the WIBARAB database contains a fair amount of technical terms concerning XML/TEI (mostly concentrated in chapter 3 of this thesis), the following subsections provide the reader with a brief introduction into these meta- and mark-up languages.

1.3.1. XML

The eXtensible Markup Language (XML) is a way to mark-up texts and encode information. It is widely used to store and process electronic texts. XML was developed in order to mark-up texts with additional information so that they can be processed by a computer. In contrast

to other technologies, XML uses the same characters for both the text and the encoding of information, according to today's standards mostly Unicode UTF-8. This makes the encoding readable for both machines and humans, which is often seen as an advantage by researchers in the humanities because it allows them to deal with the markup language similarly to a natural language (Jannidis, 2017:132). Thus, in order to distinguish between text and code, XML has established certain rules. Essentially, what sets the encodings apart from the text is the fact that they are put in between angle brackets, i.e. < and >. These encodings are called *tags* and they usually come in pairs, a start tag that opens the element and an end tag that closes it.[14]

The basic structure of XML documents consists of elements that are nested into each other. A fundamental rule of XML is that elements cannot overlap, i.e. when a so-called *parent* element contains a so-called *child* element, the child element has to be closed before the parent element can be closed. Each document needs to have a root element that contains all other elements. Generally, XML documents show both a tree-structure – i.e. hierarchically nested elements – and a sequential order – i.e. elements precede or follow other elements (Jannidis, 2017:131f).

XML is universally used due to several factors. Not only is it a cross-platform standard that was established by the World Wide Web Consortium in 1998, but it is also simple to use, human-readable, and it can be applied to many different fields. A benefit of XML that makes it especially useful for the digital humanities is its strength in dealing with complex and loosely structured data and its aptness for flexible modelling of mixed content materials (Jannidis, 2017:133f).

Strictly speaking, XML is more of a meta-language than a markup language, since it consists of only some simple basic syntax rules and no specified, limited vocabulary as the 'extensible' in XML means that every user can extend it and create their own vocabulary. Due to this, XML

[14] For a more detailed description of XML/TEI elements see section 1.3.1.1.2

became the basis of innumerable markup languages and standards that are used in countless fields and applications. One of these XML-based markup languages is the Text Encoding Initiative (TEI) (Jannidis, 2017:133).

1.3.1.1. TEI (Text Encoding Initiative)

TEI follows the XML syntax but has its own specialised vocabulary. The *Guidelines of the Text Encoding Initiative*[15] propose a very flexible and yet intuitive markup language that allows for the annotation of a wide range of types of digital text. This means that the TEI Guidelines establish what meaning is attached to specific elements, how and in which context they can be used and which attributes they can contain. At the same time, TEI is also the institution which develops and maintains this mark-up language (Jannidis, 2017:107). The *Guidelines* limit what and how elements can be used, thus providing TEI with what is called a 'controlled vocabulary.' While the TEI vocabulary is controlled, it still is – like XML – extensible, which means that users can create their own elements. The only customised element we use for our data model is the *feature value observation* (for more information see section 3.3.1.3.). The WIBARAB database follows the P5 Guidelines for TEI, which is the current version.

1.3.1.1.1. The Development of the Text Encoding Initiative (TEI)

The Text Encoding Initiative was founded in 1987 to develop and promote a standardised system of encoding humanities data in a digital form. By the 1980s, more and more humanities projects, libraries and archives attempted to make use of the newly evolving technologies. However, they faced serious obstacles in the form of poorly developed standards and a lack of exchange within the scientific community that made it nearly impossible to create long living and re-usable tools. Thus, in order to address and solve this problem, scholars from institutions

[15] The TEI Guidelines can be found under: https://tei-c.org/guidelines/p5/.

in Europe, North America and Asia held a meeting at Vassar College that led to the foundation of the TEI in 1987 (TEI, n.d.).

The first draft of the TEI Guidelines known as "P1" was created in 1990 and subsequently revised and further developed several times, until the Guidelines were adapted to conform to XML-standards for the release of the P4 version in 2002. The current version of the TEI guidelines to which the encoding of our WIBARAB database adheres is the version P5 released in 2007 (TEI, n.d.).

The TEI had an enormous impact on digital humanities scholarship in creating a free, open-source and easily accessible tool for processing and storing data. It has been endorsed by many agencies and associations around the world, including the European Union's Expert Advisory Group for Language Engineering Standards, the US National Endowment for the Humanities, the UK's Arts and Humanities Research Board and the Modern Language Association (TEI, n.d.). As a result, it has become a commonly used mark-up language for all kinds of humanities research projects. On one hand, it contains exactly the kind of vocabulary needed for encoding humanities texts and data. On the other hand, it permits considerable flexibility that, in turn, allows for the customisation to the exact needs of a highly specialised and comprehensive project such as WIBARAB.

1.3.1.1.2. TEI Elements

The TEI Guidelines determine how a TEI document should be structured and what elements it must and can contain. Elements can be defined as building blocks that contain text, other elements, media objects, and attributes. Attributes define the properties of the element, i.e. they give precise information about the element. The TEI Guidelines list all the commonly used elements with descriptions of the purposes for which they are conventionally used. These descriptions also specify which attributes can be used in which elements.

TEI elements – like XML elements in general – usually consist of a start tag, content, and an end tag. The tags have to be enclosed by angle brackets; the end tag is additionally marked with a slash: <tag>content</tag>. Furthermore, an element can contain attributes: <tag type="start">. The attribute consists of a name (type) and a value (start). An element can

contain several attributes. An element can also contain other elements. In fact, each document must have one root element which contains all other elements. Since TEI is based on XML, it needs to comply with the XML rule that elements cannot overlap.

1.4. Outline of the Thesis

After providing an introduction to the WIBARAB project and its database, to the digital resources that are available for linguistic analysis of Arabic, and to the technologies and markup languages used in linguistic databases, specifically XML and TEI, I will now outline the next chapters of the thesis.

In chapter 2, I introduce the research objectives of this thesis, and the methodology I applied in order to fulfil these objectives. Furthermore, chapter 2 contains a description of the different types of data on which the thesis is based, as well as brief descriptions of the technical tools used for the work on the database and the thesis.

Chapter 3 constitutes the main body of this thesis. It contains a detailed documentation of the WIBARAB database. In the first section of chapter 3, I describe the different types of data used in the database. In section 3.2., I discuss the methodological considerations and challenges that arose during the development of the WIBARAB database. In section 3.3., I give a technical description of the heart of the WIBARAB database, i.e. the data model. Section 3.4. is dedicated to the documentation of the data encoding workflow. Finally, the last two sections of chapter 3 provide an outlook into the database's future: in section 3.5., I give an overview of the analytic functions the WIBARAB database will have when it is finished and how they can contribute to answering the project's research questions, and section 3.6. contains an outline of the next steps in the construction of the database.

In chapter 4, I provide an analysis of the linguistic feature *qāf* in order to demonstrate some of the analytic functions the database will ultimately have. After providing general information about the feature, I describe the geographical distribution of the different reflexes of *qāf,* a

comparison of varieties spoken in cities and varieties spoken by tribes, and a comparison of three countries.

Finally, in chapter 5 I will summarise the thesis's conclusions and its main contributions to the field within which it is framed.

2. Chapter 2: Methodology

For the purpose of this thesis, it seems pertinent to differentiate between two dimensions of methodological considerations. The first one would encompass the **methodological considerations of the database**, how the data and sources are dealt with and how the database is structured; the second would include the **methodology underlying the thesis itself**. The former, i.e. the methodological considerations underlying the database and the data model, constitutes one of the main topics of this thesis and is further discussed and analysed in chapter 3. In the following sections, therefore, I will describe the methodological approach adopted and applied in the present thesis after defining its research objectives.

2.1. Research Objectives of the Thesis

The overall goal of this thesis is to illustrate how digital tools, and particularly databases, can enhance linguistic analysis. This illustration is realised by making a case study of the WIBARAB database. To achieve this, I set three main research objectives for the thesis.

1) **Documentation of the Database and its Construction:** This thesis aims at providing a general documentation of the WIBARAB database and its construction by thoroughly describing the process of creation, design, feeding and further fine-tuning of the WIBARAB database. This documentation of the database will not only serve as a part of the general documentation of the whole WIBARAB project, but it also aims at contributing to the research field and provide guidance for potential future databases within Arabic dialectology.
2) **Illustration of the Specificities of a TEI-Based Database:** From the technical side, the thesis aims at illustrating the specificities of TEI-based databases for linguistic analysis. Here, I particularly want to demonstrate the advantages of TEI's flexible structure, which allowed us to customise the WIBARAB database to address our needs, especially in terms of granularity and specific challenges that arose from the data and the complexity of the project's research questions.

3) **Demonstration of the Potential of the Database's Analytic Functions:** From the linguistic perspective, the thesis aims at demonstrating the potential of the analytic functions of the WIBARAB database within the fields of Arabic linguistics, Arabic sociolinguistics and Arabic dialectology. More concretely, I want to show how this database represents a useful tool both to answer both the project's main research questions as well as a myriad of potential questions that other researchers within these fields may pose in the future.

2.2. Methodological Approaches

According to the three main objectives outlined above, I will describe the methods applied in this thesis in order to fulfil these objectives. Due to the versatile nature of my research objectives, I applied several different methods which I describe briefly in the following:

- **Practical Experience with the Database:** Being a member of the WIBARAB team, and specifically the database team from the beginning of the project in October 2021, gave me the opportunity to deeply immerse myself into the topic of linguistic databases. My work in the database team resulted in countless discussions with other team members and (external) digital humanities specialists as well as a number of new skills I acquired, beginning with the use of technical tools and insights into the development of a data model. This experience of actively taking part in developing the data model and in encoding and validating the data provides me with the expertise needed to write this thesis.
- **Participant observation and descriptive approach:** In order to document the expertise gathered through participant observation in the course of my work on the database and in the database team, and to document the WIBARAB database and its creation, I chose a descriptive approach for this thesis. This descriptive approach aims at providing the reader not only with the technical aspects of the database but also with detailed explanations of the processes behind it. The structure of the thesis and its chapters is substantially based on this descriptive approach.

- **Literature research:** For the purpose of strengthening the factual basis of the thesis and mainly embedding it into the academic context of both digital humanities and Arabic linguistics, I conducted careful literature research. Due to the interdisciplinary nature of the WIBARAB database and thus of the thesis, the literature research encompassed works and resources on Arabic dialectology and linguistics, information technology, and digital humanities. The diverse content of the reviewed literature was also mirrored by the diversity of its formats, including 'traditional' formats like books and journal articles as well as digital resources – a fact that becomes evident when looking at the many websites listed in the bibliography.
- **Applying technical tools:** This thesis could not have been written without the use of technical tools, namely the TEI Enricher, GitHub, Visual Studio Code (see section 2.4.1.) and the WIBARAB database (see section 4.2.2.) itself. While not all of these tools were directly used to produce the analysis presented in the thesis, I used them all during the writing process to work with the data on which the thesis is based.

Before specifying exactly how I applied these methods to fulfil the three research objectives of my thesis, I want to mention the fundamental skills which I acquired through my experience working on the database, and which underly the whole thesis and thus contribute to fulfil all three of the research objectives. Due to my non-technical background, the focus in the beginning was learning on the one hand to use the technologies, but on the other hand also to get familiar with (linguistic) databases in general and TEI based databases in particular. This included a basic understanding of XML and rudimentary knowledge of XSLT, the ability to encode data in TEI and the ability to work with the different programmes and tools we use. Furthermore, I gained an understanding of how data can be modelled in order to fulfil specific purposes and provide a database with specific functions. In addition, I developed an understanding of the methodological considerations and challenges that can arise in the data modelling process, particularly through participant observation.

1) **Documentation of the Database and its Construction**: In order to fulfil the objective of providing a thorough documentation of the database and its construction, I applied all of the above-mentioned methods. As a first step I acquired the technical and

theoretical knowledge as a member of the database team. Then, after gaining all the necessary expertise in two years of participant observation, I started to describe what I had learned about databases in general and the WIBARAB database in particular. For the process of the description, I used notes that I had taken during discussions, and also relied on information stored in the background files of the database (see section 3.3.2.), especially the ODD, which is the material result of the many hours of discussions with and information received from the ACDH-CH part of the team. In order to provide the descriptions with some theoretical background, I used literature research, both for the technical as well as the linguistic side of the database. At this point, I want to stress that I deliberately chose a less theory-oriented and more hands-on approach to the documentation, due to the possibilities this offers; in particular for the thesis to act as a guide to scholars who want to build their own database. For this reason, and to ease the process of understanding certain (technical) concepts for readers with limited or no technical experience, I provided a substantial number of figures and visual examples. This was done by using the technical tools named above.

2) **Illustration of the Specificities of a TEI-Based Database:** To fulfil this objective, I mainly had to deepen my understanding of digital humanities and XML based technologies in general and TEI in particular. This was achieved by mainly three factors: discussion and courses with experts, practical work on TEI, and literature research. My work on the database team provided me with the first two. In addition to the weekly database meetings with the digital humanities experts on the WIBARAB team, I was able to take part in a workshop with TEI expert Laurent Romary in May 2022,[16] and a Tool Gallery organised by the ACDH-CH on 'XSLT for digital

[16] This workshop was organised by the WIBARAB project and held at the University of Vienna, in order to discuss the progress of the database which we had made at that point. In December 2023 Laurent Romary will visit us for a follow-up workshop.

humanists.' Furthermore, my daily work on the database creating new feature files, inputting data and validating the finished files provided me with ample opportunity to practise my TEI skills. Both the theoretical knowledge acquired in discussions and courses and the practical experience with TEI allowed me to give a technical description of the WIBARAB database. Therefore, with the third factor, literature research, I intended specifically to close gaps and connect the dots of the technical side of the thesis, and to illustrate how flexibly TEI can be used.

3) **Demonstration of the Potential of the Database's Analytic Functions:** While this objective is dealt with most extensively in the analysis in chapter 4, it is present throughout the whole thesis, especially the aim to show how the database with its analytic functions will contribute to answering WIBARAB's research questions. For this specific aim, I relied heavily on the ERC proposals for the WIBARAB project to demonstrate how the data model and the analytic functions are aligned to these questions so that they can most effectively contribute to answering them. In order to fulfil the general objective of demonstrating the potential of the database's analytic functions, I used mostly literature research, and then the WIBARAB database itself. For the literature research, in this case, I relied mostly on sources from Arabic dialectology, many of which are also included in the database. In addition to that, I used the current version of the database, which is mainly TEI Enricher-based, to give an example of how an analysis done with the database could look like.

Finally, I would like to point out that the WIBARAB database is still under construction at this very moment of writing. Therefore, the present thesis will cover the first two years of the database's development, namely the creation of the data model and the data encoding workflow. It is imperative to highlight here that the description of the database contained in this thesis describes the database in its current state. This means that while the main structures of the data model, the encoding workflow and even the future analytic functions of the database have been decided on, it is still possible – and even probable – that some details will change between now and until the database's completion.

2.3. Data

The foundation of this thesis consists of a large amount and significant variety of data that can be divided into four different categories: (1) sources of the database; (2) the files that make up the database and their content; (3) the expertise and know-how that are the outcome of discussions and participant observation; and (4) the published resources which do not fall into category (1). It is important to note that all the data enumerated below is based on the database's state of development by mid-October 2023.

The first category of data, i.e. sources of the database, consists of publications, fieldwork campaigns, and personal communications. In writing this thesis, I was able to draw on these sources which are currently made up of approximately 300 publications mainly in the field of Arabic dialectology and linguistics, 145 personal communications and 8 fieldwork campaigns.

The second category of data, the files that make up the database and their content, consist of feature-files, background files, feature value observations, geo data, and references.[17] At the moment, the database consists of 77 feature files which contain 18.975 *feature value observations* for 318 Arabic varieties. Moreover, it includes six background files, containing among other information geo-data on 572 places, and a total of 5.192 bibliographical items. In addition, the database incorporates 418 language profiles. Although it is important to mention that not all of the feature files and language profiles are completed yet, this still adds up to a considerable amount of data even at the current level of completion.

The third category of data, the skills and knowledge resulting from discussions with and work alongside experts of both Arabic dialectology and digital humanities, is more difficult to quantify, although it plays an essential role for both the WIBARAB database and this thesis.

[17] See chapter 3.3 for detailed discussions of this category.

On the one hand this knowledge includes technical skills I acquired by working on the database such as a basic understanding of XML and rudimentary knowledge of XSLT, the ability to encode data in TEI and the ability to work with the different programmes and tools (TEI Enricher, Visual Studio Code, GitHub). On the other hand, this category of information entails the results of all discussion and considerations which underly the WIBARAB database and its data model and which are in some part documented in the ODD and primarily in this thesis.

The fourth and final category of data includes published resources which do not feed into the database. These are mostly made up of publications on digital humanities and information technology. Like all publications used to write this thesis, they are listed in the bibliography at the end.

2.4. Technical Tools

The WIBARAB team uses several technical tools for the construction of the database. In the following sections, I will briefly introduce the technical tools used for the construction of the database and for the analysis. Since the WIBARAB database itself was essential for the analysis, and will be a technical tool for further analyses by scholars in the future, it is mentioned here alongside the preexisting tools which are also mainly used for the work on the database.

2.4.1. Preexisting Tools

2.4.1.1. TEI Enricher

The TEI Enricher is an experimental XML editor designed for the easy production of TEI documents. It contains several built-in functions that allow the user to compose comparatively large text documents, create and edit standoff annotations, and chunk and tokenise texts. It can be used to work with a range of basic text technologies: XML, XSLT, HTML, CSS. The TEI Enricher was developed at the ACDH-CH by Charly Mörth and Omar Siam, with the original goal of creating digital teaching resources in TEI-format. At its current stage, the TEI Enricher

contains a few functions that are specifically useful for working on language-related projects. For instance, the TEI Enricher is capable to deal with documents containing text passages written in alphabets that run right-to-left, which is of special interest to scholars working with Arabic. The main components of the editor are organised in different forms: The main form, the editor form, preview form, browser form, query form, structure form, and macros form. The TEI enricher is arguably our most essential tool for the WIBARAB database. All the data encoding workflow (data input, revision, validation; for details see section **Fehler! V erweisquelle konnte nicht gefunden werden.**) is done with the TEI Enricher. Additionally, the analysis provided in chapter 4 is executed in the TEI enricher due to the current lack of a database user interface. It is extremely useful for us that the TEI Enricher is being developed by members of the ACDH-CH section of the WIBARAB team, since this means that some of its functions pertaining to the data input are customised especially to suit our needs.[18] Since the start of the WIBARAB project, more than 30 new versions of the TEI Enricher have been developed. Currently, we are working with version 1.3.9. To sum up, the TEI enricher is a practical tool which is used for both the work on the database and the analysis presented in this thesis.

[18] For instance, when we noticed that many items in a feature file had double IDs, which is a problem for the validity of the document, an easy way to check for those IDs was built into the updated version of TEI Enricher.

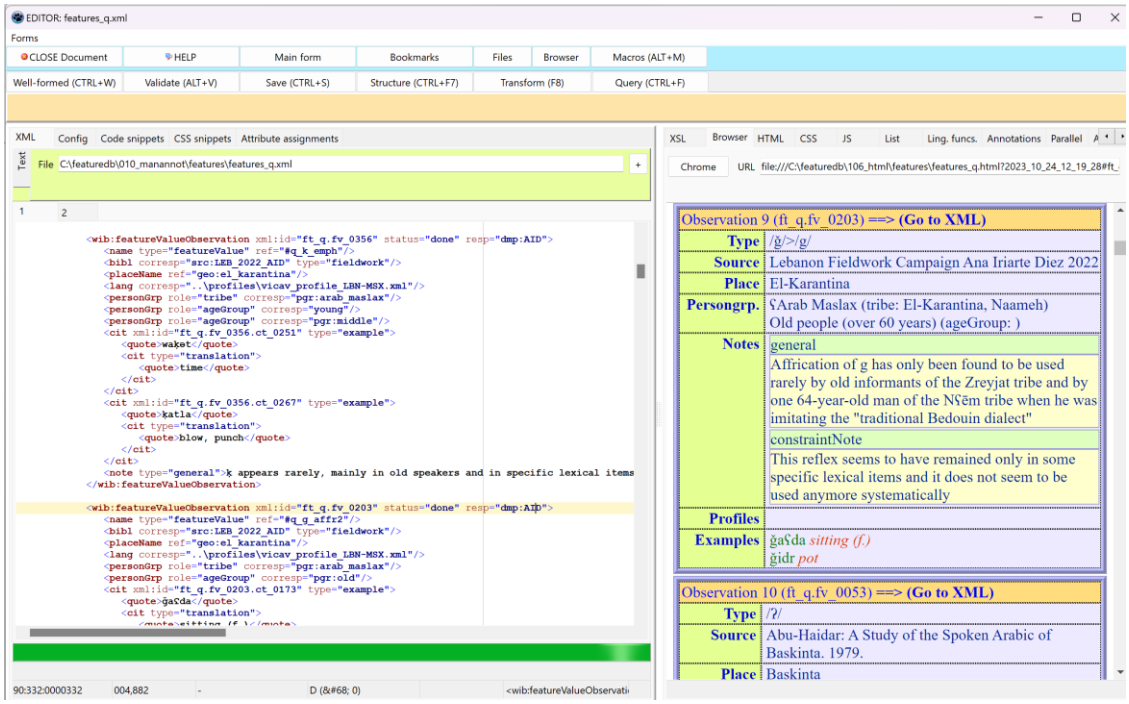

Figure 1 - Interface of the TEI Enricher; the data in TEI format on the left side and in transformation on the right side.

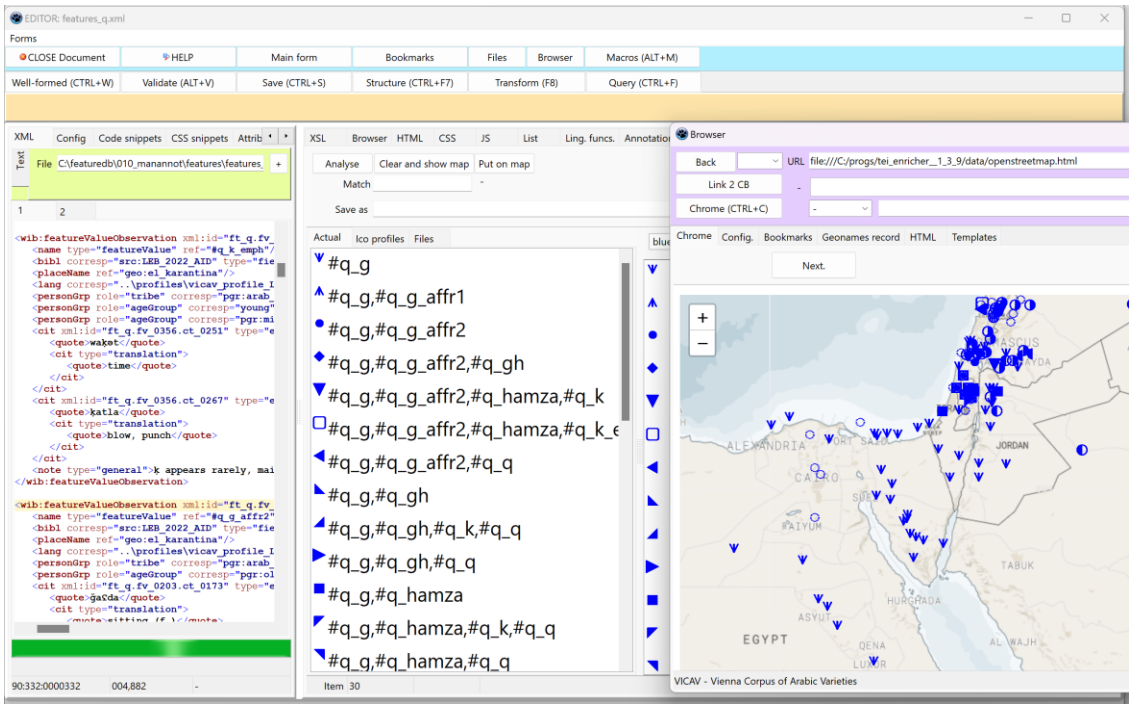

Figure 2 - Interface of the TEI Enricher; from left to right: data in TEI format, analysis showing values of /q/ in blue icons, map of the different values of /q/.

2.4.1.1. GitHub

GitHub is a platform that allows its users to collaborate on projects and store them. It is a cloud-based service that is mostly used for software development and version control through its connection to Git. GitHub provides unlimited repositories and a powerful open-source community in addition to flexible project management tools such as GitHub Issues and GitHub Projects. All this allows teams to co-work very efficiently (GitHub, n.d.)

For the WIBARAB database, we use GitHub both to store the database's files and for project management. The former means that all the files that make up the WIBARAB database are stored in a public repository on GitHub. Therefore, the WIBARAB team can collectively work on them while interested parties can see the repository and its content. As for project management, we mostly use GitHub Issues to raise issues that we encountered in the work process and assign team members certain tasks in order to solve those issues. In other words, GitHub is an important tool for the construction of the WIBARAB database, but it is not directly used for the analysis presented in chapter 4.

2.4.1.2. Visual Studio Code (VSC)

VSC is a source-code editor that runs on the desktop and supports different programming languages. It is a free tool that was built on open source by Microsoft. However, it is not only compatible with Windows, but also with Linux and MacOS. Its main features contain debugging, task running and version control (Visual Studio Code, n.d.).

We do not use it to edit the code, but mainly for version control, i.e. keeping track of changes in the TEI files, which is especially important when several people work on the files simultaneously. The second function we use, is the 'push' and 'pull' function that is available in VSC due to its connection to Git. This means that we use VSC to 'push', i.e. upload, any data we added to the files of the database during the data encoding workflow to the GitHub repository where it is stored. We also use VSC to 'pull', i.e. download, the modifications other team members made to the files in order to be able to work on the newest version of the database files. In sum, we make use of VSC mostly as a connector between the TEI Enricher (where we work on the files) and GitHub (where we store them). The main reason behind our choice to

use VSC is its user-friendly interface which makes it accessible also for team members without technical training, an important point in a project where the majority of the core team members do not have a background in IT. To summarise, while VSC is not really used for the analysis part of this thesis, it is still an important means for the workflow, especially the data input process.

2.4.2. WIBARAB Database

The analysis of *qāf*, i.e. chapter 4, is almost solely based on the WIBARAB database. Although the queries and visualisations of the maps were carried out via the TEI Enricher due to the current lack of a user interface, all the data needed to produce those query results and customised maps comes from the files that make up the database. This makes it necessary to list the WIBARAB database as one of the tools used for the analysis. A detailed description of the database is provided in chapter 3. Chapter 4 contains a practical example of how the database can be used.

3. Chapter 3: The Database

In the third and main chapter of my thesis I will describe in detail the WIBARAB database and its development.

3.1. The Data and the Sources

Generally, the WIBARAB database aims to enable linguistic analyses of spoken Arabic by making data accessible and comparable on as many Arabic varieties as possible. The basis for the construction of the WIBARAB database was the aim of the database as set out in the project's proposal, outlining the database's contribution to the project's central goal: "The database will enable us to analyse significantly more data than before. This will allow us to find bundles of related features which characterize bedouin-type dialects and thus provide the possibility of distinguishing them from sedentary-type dialects on sound linguistic criteria" (Procházka 2020:5 B1). In support of this goal, we are building a "database allowing for efficient cross-dialectal comparisons, particularly with regard to phonological, morphological, syntactical, and, to a limited extent, phraseological and lexical features" (Procházka 2020:11 B2).

For this purpose, we had to examine the diverse nature of the different types of sources that were to be included in this database in order to define best how to envision the data modelling process. One possible approach would have been the construction of a corpus-based database that would mainly consist of texts from recordings obtained during fieldwork campaigns by members of the WIBARAB team. However, this approach would revolve around the new data from the relatively scarce number of varieties that is being gathered in the frame of the project. These varieties are: The Arabic of the Bani Ṣaxar tribe in Jordan, of the Šawiya and Zʕir tribes in Morocco, of the Arab Xalde and Arab Maslax in Lebanon, of the Rašayda tribe in Saudi Arabia, Kuwait, and Sudan and of Arabic speaking communities in Turkey. This focus would, however, have significantly narrowed the scope of the database and thus prevented us from satisfactorily achieving its goal. Considering how wide-ranging the project's research questions are, it was evident that the broader and the more varied the data, the better the chances to find

satisfying answers to those questions. Therefore, the decision was to adhere to an approach that focuses on including data from as many Arabic varieties as possible, paying, of course, special attention to those traditionally considered to be Bedouin-type varieties. The Database therefore includes 318 varieties, most of which were described in previous academic studies, in addition to the aforementioned new data gathered in fieldwork campaigns within the framework of the project. The academic studies range from full grammatical descriptions to short articles focusing on specific grammatical features, and also include dialect atlases and dictionaries.

The major part of the data will be extracted from publications, although the fieldwork data will provide much more granular information on its varieties, which generates a certain quantity versus quality situation. This diversity of the data brings about the need for a more complex data model than a more homogeneous data set would require. When available, we also include data collected via personal communications with experts or native speakers of certain varieties. In the following sections, I will provide a more detailed description of the sources feeding the WIBARAB database. How we dealt with these different types of sources on a technical level is described in section3.2.

3.1.1. Publications

The main bulk of the data compiled in our database comes from academic publications. These publications vary significantly in terms of format, length, and topicality. The written sources can be roughly divided into highly specialised sources, comprehensive grammatical descriptions, and dialect atlases. Works from each of these subgroups have their own advantages and limitations. A difficulty that we encountered throughout many publications regardless of their format is that sometimes the transcription systems are not entirely clear, for example, when characters that are not conventionally used occur and are not properly explained. This is in many cases connected with either the rather remote publication date of the work or with different transcription systems related to conventions of the author's first language or educational tradition. Another limitation that is quite common throughout different types of publications is the fact that they often do not contain much information on the speakers, which makes it difficult to base studies with a sociolinguistic aspect on these works.

The highly specialised sources include “brief” articles that describe one or two linguistic features in great detail, e.g. Asiri's (2008) article *Relative Clauses in the Dialect of Rijal Alma’* as well as papers containing preliminary remarks on a certain variety that are often based on a few texts, sometimes on a single text, e.g. Holes' (2013) *An Arabic Text of Ṣūr, Oman*. Although these studies allow us to include varieties in our database which are studied and described to a lesser extent and which we would otherwise not be able to include, they come with obvious limitations. The most crucial limitation is their limited scope, meaning that they only provide information on a few features for the varieties in question. Similarly, monographs dedicated to a certain subarea of linguistics, e.g. Watson's (1993) book on the syntax of Ṣanʕānī Arabic or Shaaban's (1977) study on the phonology of Omani Arabic, while providing detailed information on their specific topic, do not cover other areas that are relevant for our database.

The second subgroup of publications are full grammatical descriptions like Woidich's (2006) grammar of Cairene Arabic, or grammars that comprise a dialectal area, i.e. encompass several related varieties, e.g. De Jong's (2000) description of the Bedouin dialects of northern Sinai. A possible limitation of multi-dialect grammars is that, when describing some features, they may fail to specify which variety, among those described, they refer to, which sometimes leaves data gathering open to the deductive skills of the reader. For full grammatical descriptions, the only apparent limitation is their scarcity; if every variety were described in a comprehensive grammatical study, the data collection for databases like ours would be much easier. However, even the most comprehensive grammars might omit explicitly mentioning some of the features we intend to include in the WIBARAB database, again forcing us to use our deductive skills for data collection.

Another important source for our database is language or dialect atlases, e.g. Behnstedt’s (2016) *Dialect Atlas of North Yemen and Adjacent Areas*, the dialect atlas of Egypt by Behnstedt & Woidich (1985), the dialect atlas of Galilee by Behnstedt & Geva-Kleinberger (2019), Behnstedt’s (1997) dialect atlas of Syria, and the *Wortatlas der Arabischen Dialekte* by Behnstedt & Woidich (2010). These atlases present large amounts of data in an organised and systematic way. They provide maps which show how a feature is realised in a particular location, just as our database will do. Nevertheless, the atlases also often present considerable

limitations for WIBARAB purposes. On the one hand, they often do not provide sufficiently specific information on the group of speakers, the method, time, and place of data collection. As a result, in some regions it is very difficult to know which of the varieties depicted in the maps are spoken by nomadic and which by sedentary communities. On the other hand, and similarly to what happens with other types of sources, atlases do not always include the totality of features that are covered in the database. For instance, in Behnstedt & Woidich's (1985) dialect atlas of Egypt as well as in Behnstedt & Geva-Kleinberger's (2019) dialect atlas of the Galilee the bound pronoun for the 1st person plural is not accounted for in any of the maps, which poses an evident challenge for the data collection and input of specific features into the database.

In addition to the differences in the length and degree of comprehensiveness of the studies, our sources provide data that was collected throughout a period of around 130 years. On the one hand, we include publications that date back to the late 19th century, e.g. Reinhardt's (1894) book on the dialect of Rustāq in Oman. On the other hand, we include recently published studies such as Torzullo's (2022) article on the Bani ʕAbbād in Jordan, and a myriad of sources in between these extremes. This gives our database a certain diachronic dimension that we had to consider in the data modelling process.

3.1.2. Fieldwork Data

The fieldwork sources consist of new data collected by WIBARAB team members in campaigns within the framework of the project. The data was collected in several countries and consists of recordings of different traditionally or previously nomadic and/or semi-nomadic communities. Generally, the fieldwork data consists of transcribed recordings of both free speech and linguistic elicitation via a grammatical questionnaire and a variety of visual stimuli.

Table 1 lists all fieldwork campaigns that were conducted in the framework of the WIBARAB project. Additionally, we have data from previous fieldwork from Arabic speaking communities in Turkey collected by Stephan Procházka, data from the Bani Ṣaxar tribe in

Jordan collected by Antonella Torzullo and data from the Rašayda of Sudan collected by Stefan Reichmuth in the 1970ies.

Table 1 - Fieldwork campaigns that were conducted in the framework of the WIBARAB project.

CAMPAIGN	Person in charge	Place of the campaign	Dates of the campaign
LEB_2022_AID	Ana Iriarte Díez	Lebanon	February-March 2022
LEB_2022_CL	Claudia Laaber	Lebanon	February-March 2022
JOR_2022_CL	Claudia Laaber	Jordan	May 2022
SAU_2022_GK	Gunda Kinzl	Saudi Arabia	March-April 2022
KWT_2022_GK	Gunda Kinzl	Kuwait	May 2022
MOR_2022_TD	Terlan Djavadova	Morocco	March 2022
TUR_2002-2022_SP	Stephan Procházka	Turkey	August 2002-November 2022
SDN_2023_SK	Stanley Kochem	Sudan	February 2023-March 2023
LEB_2023_CL	Claudia Laaber	Lebanon	May 2023
LEB_2023_AID	Ana Iriarte Díez	Lebanon	May 2023
MOR_2023_TD	Terlan Djavadova	Morocco	February 2023-March 2023
JOR_2023_AT	Antonella Torzullo	Jordan	September 2023

What makes data collected via fieldwork so important and interesting not only for the database, is that linguistic data gathered directly through fieldwork often comes hand in hand with detailed sociolinguistic information and observations. In these cases, the fieldwork data provides us with a substantial amount of information about variation, social class, accommodation, gender and age differences, etc. This variety of information and the need to reflect all these specificities was essential for the conceptualisation of our data model.

3.1.3. Personal Communications

Personal communications consist of linguistic data that we elicit from experts or L1-speakers regarding specific features. Personal communications may therefore come in different shapes and forms, e.g. as e-mails, conversations, or team members inputting data based on their ability as L1-speakers. These sources are different from fieldwork insofar as they are not collected

systematically in campaigns, but rather serve to fill in lacunae in the literature which we notice during data input. They are restricted to the expertise and native speakers available to the team. The most important limitation of personal communications is that they are not publicly available, so in order to independently verify the information, the database's users would need to contact the person who is the source of the information, and whose name is duly provided.

3.2. The Development of the Data Model: Challenges and Solutions

The objective of the first step in the construction of the database was collecting and modelling (published) data in order to encode it. In the process of developing the data model we encountered many challenges and questions. The solutions we reached after long discussions and careful consideration constitute the central methodological decisions concerning our data model. Necessarily, the collection and modelling of data entail interpreting the data to a certain degree. The way in which we interpret data determines how it is presented and subsequently perceived by the users of the database. Deciding how to deal with this issue was one of the most essential difficulties in the whole data modelling process. From the beginning, we attempted to adhere to as descriptive an approach as possible, displaying the data the way we find it in the sources as faithfully as possible. Thus, we aim to build an objective database. However, a certain degree of interpretation is intrinsic to the creation of any analytical framework as we will see, e.g. in section 3.3.1.3.12 on the feature value list. Therefore, in this section I intend to provide an overview of the most important questions that have arisen during the data modelling process until now and disclose the considerations and reasons that shaped our decisions and thus the data model of our database.

Generally, especially among sedentary speakers, one variety has traditionally been associated to one location, and to one "homogeneous" group of speakers. Consequently, scholars who studied a particular variety usually named it after the community who speaks it or after the place where it is or was spoken, e.g. the dialect of Damascus, or the dialect of the Baḥārna in Bahrain. Traditionally, it was a one to one, ideology-driven correspondence.

This approach has been challenged recently by some contemporary studies within the fields of Arabic linguistics and dialectology, where, more and more, scholars find that variation is, and has probably always been, the norm. Increased mobility fosters a more diverse and dynamic relation between different speech communities and their "places of residence". In Bedouin communities, this has often been, in fact, the norm. Some tribes are known to be present nowadays in different locations, and in some of these locations, they may coexist with other populations. Hence, we also find single locations with diverse groups of speakers. In essence, this means that in our data we deal with a diversity of locations for one tribe or speech community on the one hand, and a diversity of speech communities in one location on the other hand.

This complexity created two of the main challenges we had to face in building the database. First, our data model had to handle the aforementioned complexity of the relation between diverse speech communities and their places of residence, and second, it needed to remain an objective source of (socio)linguistic data that allows its users to interpret the data. Although we had initially placed the concept of the 'Arabic dialect' at the centre of our data model, we recognised early in the process that it was quite challenging to come up with an exact and objective definition of the term. At first, we intended to structure the data in a TEI-feature-structure-element,[19] which would connect a certain feature realisation with a dialect as classified in the literature. However, it became evident that most of the sources used as basis for our data had different and/or undefined or ideologically charged notions of what constitutes a dialect, so replicating this would have hindered us from achieving the database's aim of objectivity. This became especially apparent when we invited the TEI specialist Laurent Romary for a workshop in May 2022 to give us feedback on our progress on the database. Therefore, after careful consideration, we decided against putting the often ideologically

[19] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html.

charged concept of ‘dialect’ at the centre of our database and avoided the term altogether instead. Taking a step back, we decided to rethink our approach and build a database that purely records the occurrences of independent specific linguistic realisations in specific places, by specific speakers, rather than one that perpetuates the ideological grouping of these data points into clear-cut, well defined ‘dialects.’ Hence, the WIBARAB database is a database revolving around the occurrence of linguistic features, rather than a database revolving around the notion of ‘dialect.’

3.2.1. Objectivity

In order to preserve the database’s objectivity and place the realisation of a linguistic feature in a specific location as reported by a specific source in the centre of our data model, we created a customised TEI element, the *feature value observation* element, to fit our exact needs. The *feature value observation* brings together all the information about the occurrence of a given feature. It contains elements that indicate how the feature is realised – the feature value –, the source of all the information, the place where it occurs, the variety it is typically ascribed to in the literature, the rough date of data collection, the group of speakers who reportedly make use of this value, and all additional information that is found in the source, both about the value itself and about the speech community. For more detailed descriptions of the TEI elements used and the type of information stored in them see section **Fehler! Verweisquelle konnte n icht gefunden werden.**.

Figure 3 shows an example of a *feature value observation* for the feature ‘realisations of *qāf*’ and the information it can contain. The first line shows that *qāf* is realised as g, hence the feature value /g/. The next lines provide the source of this information, Ana Iriarte Díez’s 2022 fieldwork campaign to Lebanon, and the place this feature realisation was recorded, El-Karantina. After this, more information is provided about the speakers’ tribal affiliation and age group. Finally, this *feature value observation* contains additional information in the form of two notes and provides examples of the feature.

Observation 3 (fv_0071) ==> **(Go to XML)**	
Type	/g/
Source	Lebanon Fieldwork Campaign Ana Iriarte Diez 2022
Place	El-Karantina
Persongrp.	ʕArab Maslax (tribe: El-Karantina, Naameh) Middle aged people (30-59 years) (ageGroup:) Old people (over 60 years) (ageGroup:)
Notes	general q > g seems to be used consistently by the old generations of this community. Most middle-age speakers alternate between ʔ and g, and also use k and q with specific lexicon constraintNote Young speakers do not seem to use this reflex and prefer ʔ instead
Profiles	
Examples	grayyib *near* miglāy *frying pan*

Figure 3 - Example of a feature value observation in transformed format.

Although we decided against putting the concept of dialect in the centre of the data model, we acknowledge that the associations between speech communities, linguistic features and places made by authors may still be of scholarly interest. For this reason, we included this information as "language/dialect profile" linked to every *feature value observation*. Here, the database benefitted from its connection to the existing VICAV infrastructure. VICAV contains around 85 dialect profiles, each of which describes a specific variety of Arabic. The dialect profiles are pinned to the VICAV map. A profile usually consists of some basic information on the variety and its classification to a certain dialect group, on the location in which it is recorded, and on the history of research done on that variety. The flexibility of the *feature value observation* element allowed us to link our database to these profiles, that is, to link each *feature value observation* to a profile hosted in VICAV for the variety in question. For the varieties covered in our database which do not possess a VICAV profile yet, profile stubs were created which will be continually fed with data during the duration of the WIBARAB project and hopefully thereafter. Linking *feature value observations* to the VICAV profiles gave us the

opportunity of providing the database users with additional information, while maintaining the objectivity of the linguistic data displayed.

3.2.2. Complexity of Bedouin Tribes and Locations

Furthermore, the complex situation of the Bedouin tribes and their location(s) was partly solved in the data model by allowing for a *feature value observation* to be connected to more than one place and to distinguish between locations and regions at the level of encoding the data and thus making this information clearly available to the database users. Additionally, we linked the tribes to their multiple locations and vice versa in the background files to ascertain the interconnectedness of this information.

A good example for this is the Bani Ṣaxar tribe in Jordan. As we can see in Figure 4, members of this tribe were encountered in six locations, all of which have been added both to all *feature value observations* pertaining to the Bani Ṣaxar and to the entry of the tribe in the persongroup-file. Additionally, notes explaining the connection to the Bani Ṣaxar tribe have been added to all the locations in their entries in the geodata-file.

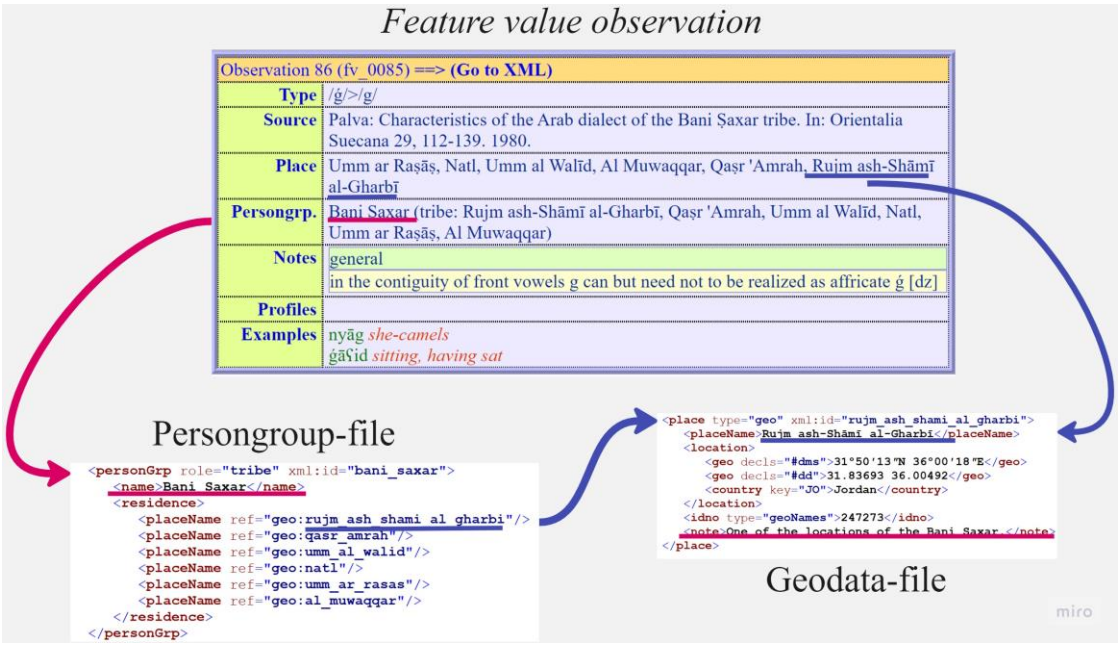

Figure 4 - Example of how the tribes are connected to locations in the data model.

3.2.3. Diversity of Data

The third main challenge we faced during developing the data model was the diversity of the data. As mentioned in section 3.1, the WIBARAB database relies on a diverse array of sources. This diversity in the sources results in the diversity of our data. On the one hand, we have our new fieldwork data that comes mostly in text corpora and is rich in sociolinguistic information, and on the other hand, we have pre-existing data from written sources that is feature-based and covers a large number of varieties. Even when only looking at the information from the written sources, the data is still varied mostly because it is often very unevenly distributed. The flexible structure of the *feature value observation* makes it possible to store and encode all the sociolinguistic information available for varieties for which we have fieldwork data just as well as the strictly necessary grammatical information. This enables us to deal with the unevenly distributed data by leaving it open to the team members to follow the agreed upon guidelines for data input (see section **Fehler! Verweisquelle konnte nicht gefunden werden.**.) and i nclude either only the most basic information about feature realisation, location, and source or to add supplementary sociolinguistic data, depending on its availability in the source.

Furthermore, the *feature value observation* and the TEI elements it contains also allowed us to solve most of the challenges posed by the sources. As mentioned before, we did this by remaining faithful to the sources and including this data as objectively as possible, without replicating (questionable) concepts or without necessarily assuming responsibility for the data's soundness or degree of certainty. This applies, for example, to old publications or studies that were based on questionable methodology, when measured against the scientific standards of today. By studies with questionable methodology, I mean, for instance, those whose authors based a whole dialect description on data they obtained from one single informant. In many cases, these publications were the only available source on several varieties that were important to be covered in the database, hence the necessity to include them. Since the WIBARAB database aims at being a tool providing objective information and not propagating subjective interpretations we chose to trust in the users' ability to put these sources into context and leave

any evaluation of these sources to the users, especially as the database clearly targets specialized users.

Concerning old sources and the resulting diachrony, we strive to give the database users as much information as possible and thus enable them to put the given information into context themselves. To achieve this and make the users aware of the remote origin date of the data, the decade of data collection will be added to every publication and made available to the database users. Whenever the date of data collection is not specified in a publication, we estimate the decade of data collection according to the date of publication and indicate in a remark that the period was estimated and is therefore not of high certainty. For instance, in T. Prochazka's (1988) *Saudi Arabian Dialects*, it is specified that the data on which the study was based was collected between 1973-1978. So, in this case we add the information that the decade of data collection was the 1970s and the certainty of this information is high since it was explicitly mentioned and not an estimate. By contrast, Lombezzi's (2019) article on the dialect of ʕIbrī, Oman does not mention the date of data collection, hence we need to make an estimate. Considering the date of publication being 2019, we assume that the data was collected in the 2010s, but the certainty of that is low. With regard to the historic varieties of Andalusia, we are confident that the target audience of the database knows enough about Arabic, the area it is spoken in, and its history to be able to recognise the historical nature of these varieties and the diachronic dimension they add to the database.

Another common challenge connected to the diversity of the data is the diversity of transcription systems used in the publications. In some cases, especially in older texts, the variation in the transcription makes it inconvenient to read in the best, and hard to understand in detail which sounds exactly are meant by the author in the worst case. To solve this, we came up with a way to keep both the standardised transcription system we use throughout the database and the original transcription of the source in the *feature value observation* and thus in our data model. Hence, we use the conventional transcription system for the feature value which facilitates the intelligibility and comparability of the data within the database, but we also add a particular element to preserve the spelling of the source. This element is called *source representation*, and it is used when the transcription of the original differs significantly

from the transcription system used in the database, or sometimes to preserve a specific feature realisation when several feature values have been grouped into one value.

For example, in the feature 'independent pronoun 3rd person singular masculine' the feature value *huwwa* incorporates the following realisations: *huwwa*, *hūwa*, *huwa*, *huwwä*, *hōwwa*, *hóuwạ*, and *huwah*. The source for the NʕēmTribe in Jordan (Cantineau, 1936) reports the realisation to be *hūẉa*, to preserve this specific realisation, it is added in a *source representation element* as can be seen in Figure 5.

Observation 81 (fv_0141) ==> **(Go to XML)**	
Type	huwwa / hūwa / huwa / huwwä / hōwwa / hóuwạ / huwah
Source	Cantineau: Etudes sur quelques parles de nomades arabes d'Orient 1. In: Annales de l'Institut d'Études Orientales 2, 1-118. 1936.
Or. src.	hūẉa
Place	Irbid
Persongrp.	Nʕēm Jordan (tribe: Irbid)
Profiles	

Figure 5 - Example of a feature value observation containing a source representation element.

To sum up, the development of the data model was shaped and influenced by mainly three crucial factors: (1) the overall wish to remain objective and avoid replicating interpretations, (2) the complex relation between the spoken varieties, locations and the speech communities, and (3) the diversity of the sources and the resulting diversity of the data. Considerations of the first and second factors provided the initial idea of the current data model with the *Feature Value Observation* at its heart. However, it was mainly the third factor, that shaped to a greater extent the details of the current data model. As we will see in the next section, the essential purpose of the data model is to provide a 'template' or a 'skeleton' that is flexible and dynamic enough to be customised in order to faithfully reflect the complexity of the methodological and (socio)linguistic variation we found in the sources. A detailed description of the data model and an explanation how this is implemented on a technical level is provided in the following section.

3.3. The WIBARAB Database's Data Model[2021]

The term *database* in the case of the WIBARAB database must be understood in a broad sense as an organized collection of electronically stored and accessible semi-structured data. The entire textual source of the WIBARAB feature database is encoded in accordance with the P5 Guidelines of the TEI.

The database is made up of a series of XML documents which fall into two categories: the feature files, i.e. linguistic descriptions, and the background files, i.e. documents which store data that is referenced in the linguistic descriptions, such as information on the sources or locations.

The first category of documents contains descriptions of a series of salient linguistic *features.* The criteria used for the selection of these features rely on typological considerations and on high levels of sociolinguistic salience, i.e. features typically labelled as "Bedouin" or reported to be perceived as "typical" for traditionally nomadic or sedentary communities. The feature descriptions are organised in feature files which bring together all data pertaining to particular linguistic phenomena, e.g. all observations concerning the different realisations of the letter *qāf.* This information makes up the largest part of the feature files and is encoded in *feature value observations.*

---

[20] This section, and subsection 3.3.1.3, is partly based on the ODD document for our database which was co-authored by Stephan Procházka, Charly Mörth, Daniel Schopper, Ana Iriarte Díez, Johanna Doppelbauer and me. The file is found in our public repository on GitHub (https://github.com/wibarab/featuredb/blob/main/802_tei_odd/featuredb.odd).

[21] I use the angle brackets < > to mark elements and the at sign @ to mark attributes according to the conventions in the TEI community, e.g.: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Every *feature value observation* specifies that one particular *feature value* occurs in a specific place and is used by a specific speaker or group of speakers. When information about the speakers and their communities is available, the sociolinguistic profile of the speaker or the community of speakers is indicated via the *person group* element, which may specify values regarding the speakers' tribal affiliation, religious affiliation, age group, gender, and first language. When specific remarks that are not codifiable via the aforementioned elements are necessary, they are indicated within a note element. As discussed in section 3.2, many of our sources tend to group combinations of realisations of feature values and characterise them as a linguistic variety or dialect. These linguistic varieties are generally labelled after the approximate location in which they are spoken, and/or the community of speakers that use them, e.g., the Arabic of the ʕAtīğ in Wādi Xāled, the Arabic of Damascus, the Arabic of the Šukriyya. In order not to replicate the ideological systems leading to this labelling, but to still provide the user with this information, such previous “groupings” and “labels” are indicated within the *feature value observation* via *language profiles.* Language profiles provide additional historical and linguistic information that is available regarding a specific variety according to one or more sources and fall under the second category of documents in the database.

The background files, our second category of documents, contain all the data surrounding the linguistic features, such as information about the locations and sources vertebrating the *feature value observations,* or the aforementioned language profiles. Although the feature files with the *feature value observations* constitute the heart of our data model, the information stored in the background files is just as important for the database and makes up its skeleton, so to speak. In the following subsections I will first give a detailed description of the feature files and *feature value observation* and all the elements it can contain. Secondly, I will explain in more detail each one of the background files, along with the kind of data they contain, and the nature of their connection to the feature files.

3.3.1. Feature Files

The feature files consist of three sections. The first of these sections is the TEI header which contains technical information and is compulsory for every TEI document. Since these headers are of a very similar structure in all TEI documents, and detailed descriptions for them can be found in the TEI Guidelines P5, I do not provide one here. The second section contains a chapter that gives a general description of the feature and includes a list of the feature values. Every linguistic feature has this set of *feature values* that indicates the different attested realisations of a specific feature in the studied varieties. The documentation of an occurrence of (a) specific feature value(s) in a specific location is accomplished via *feature value observations*. These feature value observations make up the third and main section of each feature file. The following sections provide descriptions of the chapter, the list of feature values, and the *feature value observations*.

3.3.1.1. Chapter

The chapter aims at providing the database users with general information on the linguistic feature described in the feature file. It contains general information and in-depth discussions about the feature. A chapter is, therefore, usually structured into the following sections: general remarks, including loanwords, assimilations and sociolinguistic considerations, values, and geographical distribution. Additionally, the chapter will contain information on how the concerned feature was dealt with in the database. For instance, when values had to be grouped, the information will be provided for the user in the chapter. Another case that can be included in the chapter are exceptions that occur with regularity as opposed to exceptions that are particular to single locations, which are included in the form of notes. An example of this is shown in Figure 6, where the 'loanwords' section specifies, that regardless of the local realisation of /q/, loanwords from Standard Arabic usually keep the reflex *q*. Generally, in Figure 6, we can see a part of the chapter for the feature 'realisations of *qāf*' both in TEI format and transformation. It is important to note that the formatting of the transformation is preliminary, i.e. the chapter will look differently when the database is completed.

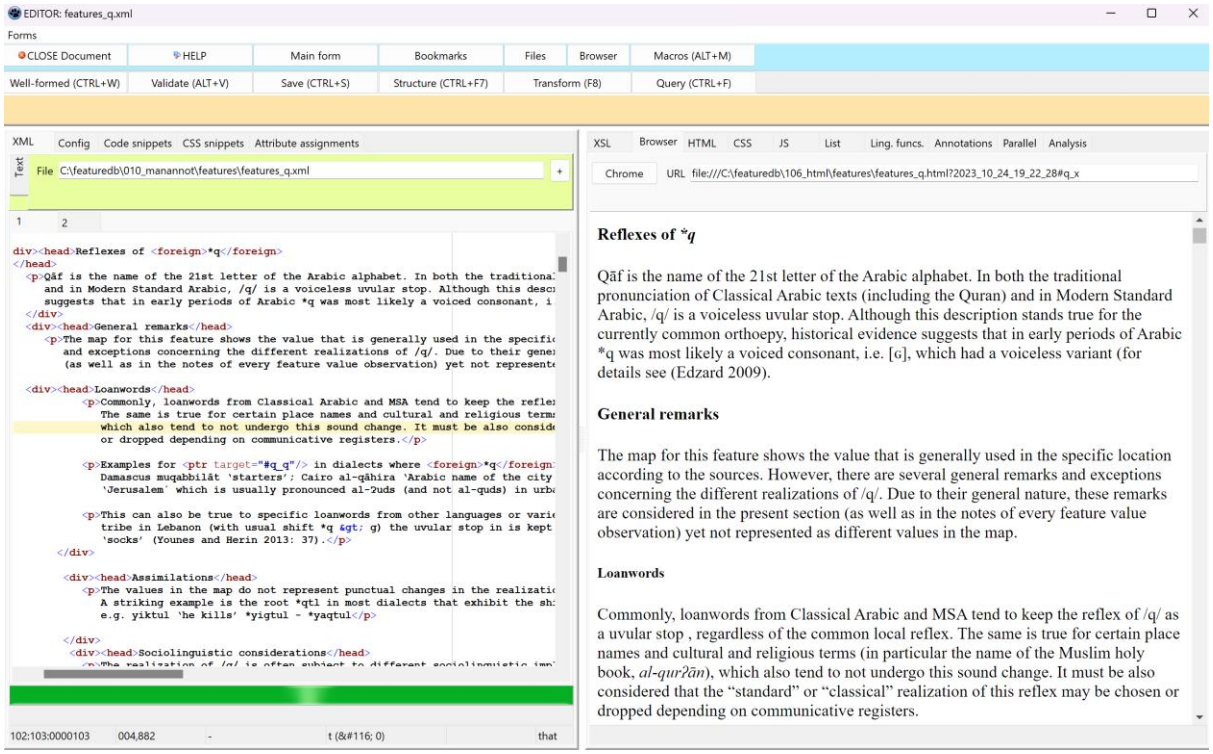

Figure 6 - The chapter of 'feature qāf' opened in the TEI Enricher; on the left side in TEI format, on the right side in transformation.

3.3.1.2. Feature Value List

At their beginning, each feature file contains a list of feature values that allow to neatly group the observed linguistic phenomena under clearly identifiable labels. For many features, e.g. the independent pronouns, we found a great number of different realisations, often due to the different transcription systems used by the authors of our written sources. Therefore, after careful consideration, we decided to group some of the variants together in order to make the data comparable. Although this decision to group some of the feature realisations might be considered a substantial interpretation of the data, it is a necessary one: to keep such a great number of different feature values would have rendered meaningful linguistic comparison very difficult if not impossible. However, to give the database users the opportunity to determine exactly which realisations are grouped together, this information is kept in each item in the *feature value list* in the <label> and <desc> (description) elements. If a certain feature file contains a significant amount of grouping of its values, this will also be mentioned in the

chapter. Moreover, to ensure that no viable data is lost and that it remains easily accessible to the users, the particular feature realisation as originally stated by the source can be included into the *feature value observation* in the *source representation* element (see section **Fehler! Verweisquelle konnte nicht gefunden werden.**).

From a technical perspective, the list includes several <item> elements that each contain an ID, a <label> element where all the realisations subsumed under one particular value are mentioned in transcription, and a short description of the value in question, containing for example the transcription of the value in the International Phonetic Alphabet (IPA). Figure 7 shows excerpts of the feature value lists of the phonological feature *reflexes of qāf* and the morphological feature *independent pronoun 3. Person singular masculine.* In the latter, the <label> elements of the first three items indicate the grouping of realisations into certain feature values.

Figure 7 – Excerpts of the list of feature values for qāf and the independent pronoun 3. person singular masculine.

3.3.1.3. The Feature Value Observation

In this section, I describe the *feature value observation* which constitutes the very core of our database's data model. The first and very important step for building a database is the conceptualisation of its basic functions and the structuring and modelling of the data that

follows from this conceptualisation. In our initial experiments, our team made use of a highly abstracted system consisting of typified <fs> (feature structure) elements. However, after extensive discussions and consultations with internal and external experts, we remodelled the entire structure attempting to make use of more expressive existing TEI elements. To come up with a solution fitting our particular goals and needs, we customised the TEI Schema attempting to reuse as many TEI elements as possible. So far, we managed to do so with only one additional new element, <featureValueObservation>, which had emerged from our discussions and appeared to serve our particular purpose best. This element is the centre piece of our customisation. All other elements exist in the TEI Guidelines and are applicable without excessively twisting their originally intended semantics.

The newly introduced <featureValueObservation> element contains different pieces of information, three of which are mandatory: (1) an indication of the linguistic phenomenon being described, (2) information on the location of the phenomenon, and (3) the source where the information comes from. Additionally, it can include information on the group of speakers, the variety it is usually ascribed to, and/or include notes and examples. The specificities of the respective elements are described below.

Each *feature value observation* is based on the occurrence of a linguistic feature in a specific way, i.e., a *feature realisation*. These realisations are grouped into a set of *feature values*. Hence, the *feature value observation* shows how a source attests that a given feature is realised as a certain feature value by a specific speaker's community at a specific location. The *feature value observation* can additionally include information on the *person group*, i.e. the community of speakers for which this particular *feature realisation* is attested. Furthermore, it is linked to the language profile of the variety to which it is customarily assigned.

Each <featureValueObservation> element can be identified with an xml:id that is unique in the document and the @resp attribute shows the name of the responsible person, i.e., the team member who is responsible for adding data to this very *feature value observation*. Finally, the @status attribute states whether the data input for this *feature value observation* is completed (status=done) or not (status=draft).

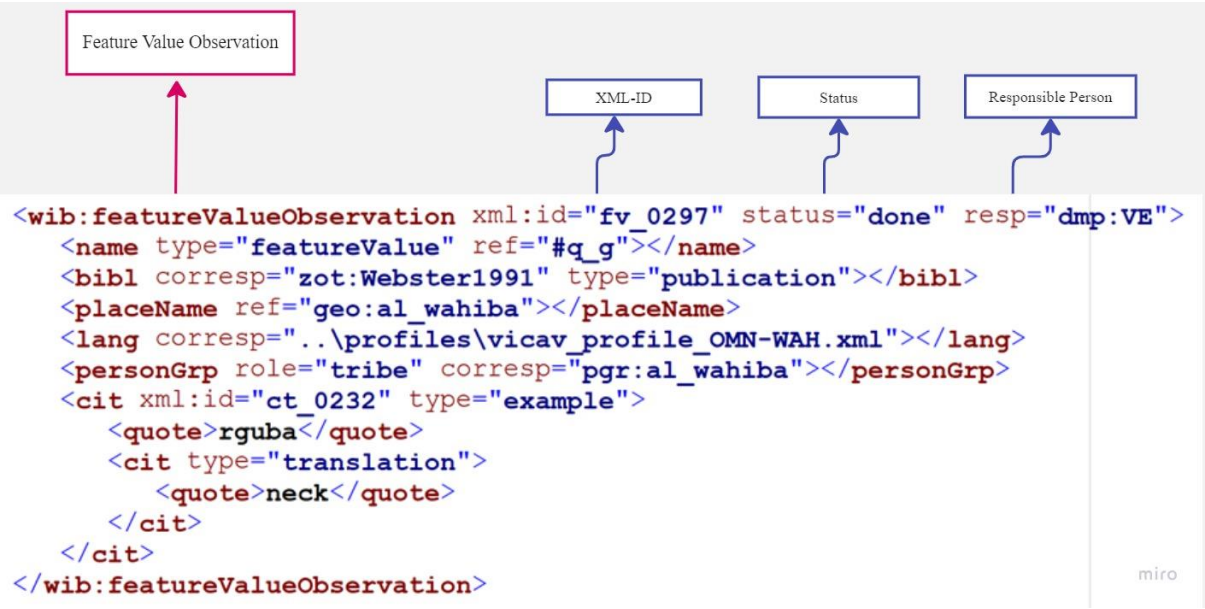

Figure 8 – Example of a feature value observation in TEI format.

To sum up, every *feature value observation* must contain three core elements: *feature value, source* and *place.* Additional to the mandatory elements, a *feature value observation* may enclose several optional elements containing more detailed information, mostly of a sociolinguistic nature.

3.3.1.3.1. Feature Value

The information stored in the feature value list (see section 3.3.1.2) is referred to in each *feature value observation* in a name element with a @ref attribute pointing to particular items of the list of feature values. Each *feature value observation* must contain exactly one feature value. If more than one value was documented by the source, a new *feature value observation* must be created to contain the additional value.

```
<name type="featureValue" ref="#q_g"></name>
```

Figure 9 – Name element containing the feature value in a feature value observation.

3.3.1.3.2. Source

Each *feature value observation* element must contain exactly one source indicating where the information has been collected. The information linking to the source is encoded in a *bibliographic citation* element <bibl>. As we have different types of sources, the <bibl> element contains @type attributes indicating the type of source they refer to, i.e. whether it is a publication, a fieldwork campaign, or a personal communication. The information about the sources is stored in different documents (see section 3.3.2 on the background files) and referenced in the *feature value observation* by the <bibl> element via a @corresp attribute. The most common case is a bibl[type="publication"] element which refers to the bibliography (Bibliography-file, see section 3.3.2.2.) as shown in Figure 10. The bibliography contains written and mostly published academic works.

```
<bibl corresp="zot:Webster1991" type="publication"></bibl>
```

Figure 10 – Example of a <bibl> element containing a reference to the bibliography.

However, as mentioned above the <bibl> element can also contain a reference to a fieldwork campaign or a personal communication, as is illustrated in Figures 11 and 12. In these cases, the @corresp attribute points to the Sources-file in which the metadata on the fieldwork campaigns and personal communications is stored (see section 3.3.2.3.).

```
<bibl corresp="src:LEB_2022_AID" type="fieldwork"></bibl>
```

Figure 11 – Example of a <bibl> element containing a reference to a fieldwork campaign.

```
<bibl corresp="src:PC_SB" type="personalCommunication"></bibl>
```

Figure 12 – Example of a <bibl> elements containing a reference to a personal communication.

3.3.1.3.3. Place

Every *feature value observation* must contain at least one location that references the place where a feature occurs as documented in the source. Due to the complexity of the data pertaining to Bedouin tribes, a *feature value observation* is not limited to one place element. Rather, it can contain a number of places, indicating the multiple settlements of a tribe, due to previous or present migratory movements. Furthermore, the same location can be referenced in more than one feature value observation, e.g. when more than one speech community resides in one place. On a technical level, the locations are indicated via <placeName> elements. These elements each contain a @ref attribute which points to the VICAV Geodata-file, a background file where all information on the locations is stored (see section 3.3.2.4).

```
<placeName ref="geo:al_wahiba"></placeName>
```

Figure 13 - <placeName> element in a feature value observation.

3.3.1.3.4. Person Group

The concept of person group describes any specific discernible group of speakers who use a particular *realisation of the feature* according to the source. According to what we found in our data, we defined the groups relevant to our database as tribe, religion, age group, gender, and first language. Each of these groups has subgroups. For instance, age contains the subgroups "young" (0-29years), "middle-aged" (30-59) and "old" (60+). One *feature value observation* can contain as many person group elements as needed to store all available information regarding the speakers. This kind of sociolinguistic information can facilitate research on different topics, e.g. variation or language change.

To encode this kind of information, we make use of <personGrp> (person group) elements. These are furnished with @role attributes that establish to which of the main groups – tribe, religion, age, gender, and first language – the data applies. Finally, the content of the second attribute, the @corresp attribute, specifies the subgroup within the main group. This @corresp attribute refers to the Persongroup-file, which stores all information concerning the five main person groups (tribe, religion, age group, gender, and first language) and their subgroups.

```
<personGrp role="tribe" corresp="pgr:al_wahiba"></personGrp>
```

Figure 14 – <personGrp> element for tribe, referencing the Āl Wahība.

```
<personGrp role="ageGroup" corresp="pgr:old"></personGrp>
```

Figure 15 – <personGrp> element for age, referencing the oldest age group of 60+.

```
<personGrp role="religion" corresp="pgr:druzes"></personGrp>
```

Figure 16 – <personGrp> element for religion, referencing the druzes.

3.3.1.3.5. Variety

Information on a particular linguistic variety as labelled by the source can be specified via a reference to the corresponding VICAV language profile. The encoding of the reference to the profile in the *feature value observation* makes use of a *language name* <lang> element. This <lang> element contains a @corresp attribute that links to a VICAV profile.

The *language profile* is a file which contains any extra information collected about the variety and the place where it is spoken. These profiles are named in accordance with our internal naming conventions with a three-letter country code, dash, variety code, e.g., OMN-WAH, standing for Oman, Āl Wahība. The profiles themselves are described in section 3.3.2.6. This way of encoding allows us to make the information the source gives about the classification of a dialect available to the database users without making it the centre of our data model.

```
<lang corresp="..\profiles\vicav_profile_OMN-WAH.xml"></lang>
```

Figure 17 - Example of <lang> element pointing to a VICAV profile.

3.3.1.3.6. Source Representation

The source representation element is the element that contains a feature realisation exactly as documented by the source. In contrast to the normalised transcriptions of the feature values, *source representations* reflect the literal information from the source quoted in its original form, i.e., in the source's transcription system. In addition to helping us overcome the issue of diverse transcription systems, this element proves to be especially useful in those features where the vast diversity of realisations found obliged us to heavily group values under a less extensive set of values. For instance, in the third person singular masculine independent pronoun, the feature value '*huwwa'* encompasses the *representations huwwa, hūwa, huwa, huwwä, hōwwa, hóuwạ,* and *huwah*.[22] First, the s*ource representation* element allows the user to access the original representation of the value. Second, it allows us to enable future grouping processes, but also to track the grouping of representations we did for the sake of clarity in the data display. This information is encoded making use of cit[@type="example"]/quote constructs which is in line with how similar phenomena are dealt with in the TEI dictionary module and also in TEI Lex-0[23]. This means they consist of a *cited quotation* <cit> element containing a @type attribute and a <quote> element which holds the content. In the case of the source representation, the <cit> element is provided with a @type attribute that specifies it as such.

```
<cit type="sourceRepresentation"><quote>húwwa</quote></cit>
```

Figure 18 - Source representation example.

[22] Oftentimes these numerous representations are themselves the product of the great number of different transcription systems that have been used in sources during the last century.

[23] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (accessed on 2023-08-09).

3.3.1.3.7. Examples

Examples may be used to illustrate a feature value and/or show it in context. They are taken from the original sources and provided to the database users in normalised transcription. Additionally, they contain an English translation, regardless of the language to which they were translated in the source. They are encoded – similarly to the source representation – with cit[@type="example"]/quote constructs which have been chosen in analogy to common lexicographic procedures. Here, the @type attribute specifies these elements as examples. Each example has an @xml:id that is unique in the document.

```
<cit xml:id="ct_0232" type="example">
   <quote>rguba</quote>
   <cit type="translation">
      <quote>neck</quote>
   </cit>
</cit>
```

Figure 19 - Example of an example.

Depending on the nature of the feature or the nature of the source, there can either be no example at all or several of them in one *feature value observation.* Some features, e.g. the independent pronouns, never include examples, since the values established for them are serving as examples of the pronouns themselves. I.e., a value of the independent pronoun 3rd person singular is '*huwwa*', another '*huw*'. For other features, e.g. most phonological features, there should always be an example. However, sometimes the sources provide information about a feature without providing an example which makes it impossible for us to include one in the database.

3.3.1.3.8. Notes

Notes can contain any kind of information that does not fit into the other elements of the data model with a more narrowly defined scope. They can be used to clarify specificities of the use, sociolinguistic tendencies, development, or typological nature of a particular linguistic phenomenon. This element is frequently used to indicate exceptions and constraints concerning the use of a feature value. The information contained in the notes draws also on the source of its *feature value observation.*

We decided to distinguish between three types of notes: general notes, constraint notes, and exception notes in order to place the content of the note in context for the database users. The notes are encoded by simple <note> elements which contain a @type attribute. This @type attribute allows us to differentiate between the different types of notes.

```
<note type="general">less frequently used than /g/, except for MSA</note>

<note type="constraintNote">q_q for informants affected by LA</note>

<note type="exceptionNote">In very few cases, q might have been realised as g.</note>
```

Figure 20 - Different types of notes in our database.

It is important to highlight that most of the general exceptions and constraints that are known to be common to the feature itself, are dealt with in greater detail in the chapter (see 3.3.1.1) of the feature, and not systematically in the independent *feature value observations*.

3.3.2. Background Files

As mentioned above, the database contains several files where we store fundamental background information which the feature files or respectively the *feature value observations* point to. These are: the WIBARAB data management plan, the Bibliography-file, the Geodata-file, the Sources-file, the Persongroup-file, and to some extent the VICAV profiles. Many of the background files are not only linked to the feature files, but also contain references to interconnect them. The following subsections describe these files in greater detail.

3.3.2.1. WIBARAB Data Management Plan (wibarab_dmp.xml)

The WIBARAB data management plan contains general information about the WIBARAB project, including a list of all the team members that contribute to the database by inputting data. Every *feature value observation* and the data input into it are assigned to a responsible person among the team. To document this step of the knowledge transfer from the source to the database, the responsible person is referenced in each *feature value observation*. This is

accomplished by the @resp attribute in *the feature value observation* that points to the data management plan as illustrated in Figure 8 (section 3.3.1.3.)

3.3.2.2. Bibliography-File (vicav_biblio_tei_zotero.xml)

The Bibliography-file contains crucial information, namely a TEI list of all the published sources we use in the WIBARAB database as well as further published sources that are not included. For this purpose, the WIBARAB project uses the VICAV Zotero library. The creation of large parts of the library predates the project by several years. This library has been built up mainly by students and researchers of the Department of Near Eastern Studies of the University of Vienna and has been used as part of the VICAV digital infrastructure by several projects (TUNICO, SHAWI, TUNOCENT, WIBARAB). It brings together research publications in the fields of Arabic dialectology and Arabic linguistics focusing on spoken varieties. All bibliographical items have been annotated with additional information. They contain at least the region dealt with in the publication. However, many items also hold further information about content and type of the publication (e.g. dictionary, grammatical description, textbook, etc.). Currently, the VICAV library contains more than 5,000 titles.

The name 'VICAV Zotero library' refers to the software that has been used to collect the data. Zotero is a freely available widely used bibliographical tool which allows to collaboratively compile bibliographies. It also provides a number of export functions, most conveniently also one producing TEI-conforming bibliographical records.

To address the issue of old sources that came up in our data modelling process, all bibliographical items used in the WIBARAB database will additionally include information on their respective decade of data collection. This will enable the database users to put the information provided by each source into their diachronic context. The entries of the Bibliography-file are linked to *feature value observations* via the <bibl> element.

3.3.2.3. Sources-File (wibarab_sources.xml)

The sources-file contains all sources that are not from the VICAV bibliography, i.e. fieldwork campaigns and personal communications. Each entry includes information on the date of the event and the participants. In the case of a fieldwork campaign, the only participant's name mentioned in the Source-file is the scholar conducting the fieldwork. The informants' names

cannot be shown due to data protection regulations. For this very reason, no sensitive metadata concerning the fieldwork campaigns is published in our repository on GitHub, but it is instead stored according to ERC guidelines. The entries concerning fieldwork campaigns include the start and end date of the campaign in ISO conformant format. They also include the country where the campaign was conducted and the person who conducted it, as illustrated in Figure 21.

```
<event type="campaign" from-iso="2022-02-22" to-iso="2022-03-18" xml:id="LEB_2022_AID">
   <head>Lebanon Fieldwork Campaign Ana Iriarte Diez 2022</head>
   <desc>
      <listPerson type="participants">
         <person>
            <persName>Ana Iriarte Diez</persName>
         </person>
      </listPerson>
      <placeName ref="Lebanon"/>
   </desc>
   <note></note>
</event>
```

Figure 21 - Example of a fieldwork campaign in the Sources-file.

In the entries concerning personal communications, the informant's name can be shown in the title, if the person agreed to it. This is due to the fact that many of the personal communications were conducted with experts or native speakers who are at the same time team members or scholars at other institutions. If consent to the publication of their name was not given, a pseudonym is used. In entries for personal communications, the variety the informant speaks is indicated via a pointer to the VICAV profile, as we can see in Figure 22, where the person who gave the personal communication, Shada Bokir (SB), is a native speaker of the Arabic variety spoken in Wadi Hadramawt, Yemen.

```
<event type="personal_communication" subtype="native" from-iso="2022-10-01" to-iso="2023-01-31" xml:id="PC_SB">
   <head>Exchange with Shada Bokir</head>
   <desc>
      <listPerson type="participants">
         <person>
            <persName>Veronika Engler</persName>
         </person>
         <person xml:id="SB" role="native">
            <idno>SB</idno>
            <langKnowledge>
               <langKnown tag="YEM-HAD" corresp="profiles/vicav_profile_YEM-HAD.xml"/>
            </langKnowledge>
         </person>
      </listPerson>
   </desc>
   <note></note>
</event>
```

Figure 22 - Example of a personal communication in the Source-file.

3.3.2.4. Geodata-File (vicav_geodata.xml)

The geodata-file contains a complete list of all the places referenced in the *feature value observation*s (and more). The data included in the entries consists of the name of the location, coordinates (in both the degrees/minutes/seconds (DMS) and the decimal degree (DD) systems), ISO-conformant country key, and ID-number. For the sake of coherence, these have been taken from the locations' respective entries in the GeoNames geographical database (GeoNames, n.d.). As visible in Figure 23, the entries can additionally include notes, many of which pertain to a Bedouin tribe which resides in this particular location.

```
<place type="geo" xml:id="al_wahiba">
   <placeName>Al Wahiba</placeName>
   <location>
      <geo decls="#dms">21°45′00″N 58°40′00″E</geo>
      <geo decls="#dd">21.75 58.66667</geo>
      <country key="OM">Oman</country>
   </location>
   <idno type="geoNames">289096</idno>
   <note type="general">Āl Wahība tribe</note>
</place>
```

Figure 23 - Example of an entry in the Geodata-file.

3.3.2.5. Persongroup-File (wibarab_PersonGroup.xml)

The Persongroup-file contains all information on the five main person groups relevant for the database: tribe, religion, age, gender and first language. The file is made up of a list for each of these groups with every list containing the relevant subgroups. Depending on the person group, the number of subgroups contained in the lists differs significantly. Each item in a list constitutes one subgroup and is encoded via a <personGrp> element with a @role attribute and an @xml:id.

```
<personGrp role="tribe" xml:id="bani_saxar">
   <name>Bani Ṣaxar</name>
   <residence>
      <placeName ref="geo:rujm_ash_shami_al_gharbi"/>
      <placeName ref="geo:qasr_amrah"/>
      <placeName ref="geo:umm_al_walid"/>
      <placeName ref="geo:natl"/>
      <placeName ref="geo:umm_ar_rasas"/>
      <placeName ref="geo:al_muwaqqar"/>
   </residence>
</personGrp>
```

Figure 24 - Example of a tribe in the Persongroup-file including several locations.

The tribe list contains 125 entries. Each of the tribes is connected to at least one location from the Geodata-file. The interconnectedness of the files mirrors the high complexity of the data and the reality where, on the one hand, one tribe can be found in many locations, but on the other hand, several different tribes or communities can be found in a single location.

```
<personGrp role="religion" xml:id="sunnis">
   <name>Sunnis</name>
</personGrp>
```

Figure 25 - Example of a religious group in the Persongroup-file.

```
<personGrp role="age" xml:id="young">
   <name>Young people (up to 29 years)</name>
</personGrp>
```

Figure 26 - Example of an age group in the Persongroup-file.

The list of religious groups contains 10 different denominations (Christians, Muslims, Jews, Malikites, Sunnis, Shiites, Druzes, Ibadis, Alawis, and Samaritans). The age group consists of three subgroups (young: up to 29 years, middle aged: 30-59 years, old: over 60 years) and the

gender group also consists of three subgroups (female, male, other). The person group ‘first language’ slightly differs from the other groups, since it does not include pre-established subgroups. This means that during data input the responsible person can enter any relevant language into the *feature value observation* in question. This person group mostly pertains to speakers in East and Sub-Saharan Africa where the first language of the speaker has a significant influence on the realisation of some features. For instance, in the Dahlak Archipelago in Eritrea, most speakers realise *qāf* as *ġ* but L1-speakers of Dahalik and Afar realise it as either *q*, *k*, or *g*. Since the number of languages that might thus influence local Arabic varieties was too large, we decided not to limit the options of the persons responsible for data input.

3.3.2.6. Profiles

The VICAV profiles differ slightly from the other background files, since some of them were created in previous projects. They contain prose descriptions of particular varieties and their research history. For varieties that do not have a profile yet, stubs have been created to be subsequently filled during the project. While many VICAV profiles can already be accessed on the VICAV map, the profile stubs that were newly created are not yet displayed.

In Figure 27, we can see how the profiles are displayed on the VICAV interface. While this illustration is a viable example of a VICAV profile content wise, the design might still change significantly. This is due to the fact that the whole VICAV infrastructure is currently undergoing reconstruction and the interface of the WIBARAB database is still in an early phase of its development.

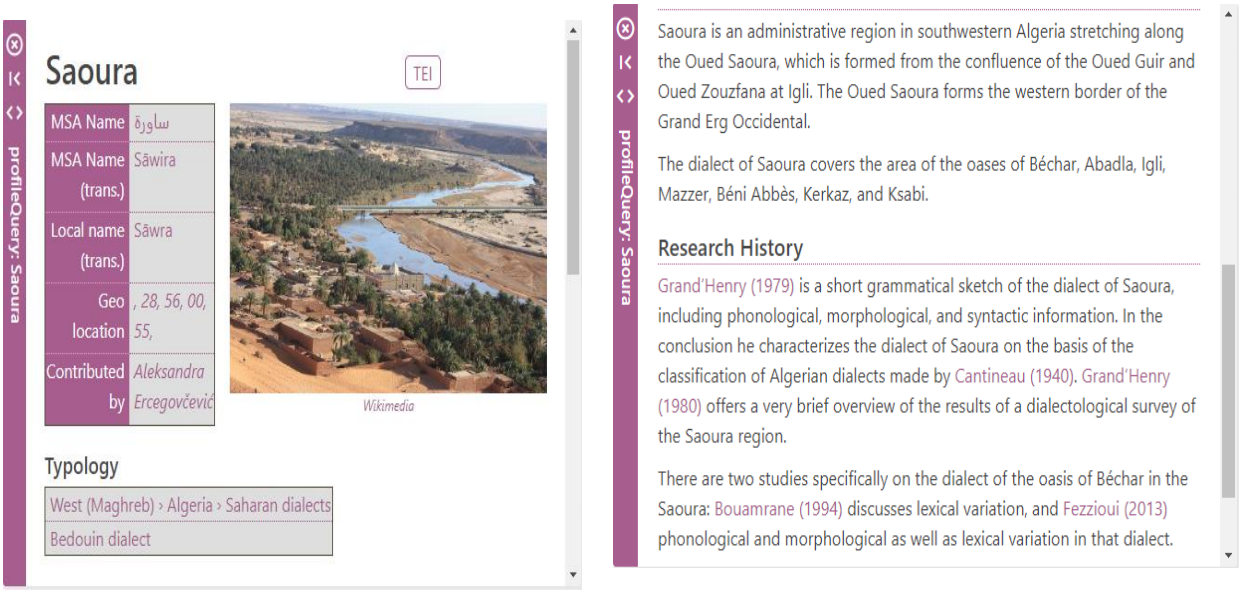

Figure 27 - Example of a profile on the VICAV interface.

3.3.2.7. ODD / Schema

In order to make the data model and both the methodological considerations underlying it and its technical properties intelligible to others, it has to be documented. This documentation is achieved on a technical level by an ODD document and a schema which is derived from the ODD. ODD stands for 'One Document Does it All' and is "a TEI XML format used to express schema fragments, prose documentation, and reference documentation" (TEI Consortium, n.d.). An ODD document is a TEI document which contains components from the module of documentation elements (tagdocs). The WIBARAB database's ODD contains both prose sections describing the database, the data model and the TEI elements used in it, and sections in code that specify the exact use of the elements and attributes, e.g. which attributes and attribute values elements can contain, how often one element can be contained in another element, and in what order specific elements have to follow each other. In order to implement these rules in all the files of the database, the ODD is converted into a schema which all the database's files have to follow. Schemata are used to make sure that data is coherent by defining for the computer how it is encoded and determining the structure of the encoding (Jannidis, 2017:135). For the WIBARAB database we use a schema in the schema language XSD (**X**ML

Schema **D**efinition). XSD is a very expressive schema language that is formulated in XML and was developed by the World Wide Web Consortium (Jannidis, 2017:136). When a TEI document is encoded according to the schema, it is 'valid'. Checking the validity of the database's documents is part of the data encoding workflow which I describe in the next sections.

3.4. Data Encoding Workflow

The Data Encoding Workflow is divided into two processes: first, data input, and second, revision and validation. These will be described in the following sections.

3.4.1. Data Input

After a rough outline of the data model was established and encoded, we moved on to the next phase of the database construction: the phase of data input. Although it was necessary to collect some preliminary data in order to develop a rudimentary draft of the data model, it was only after we started the actual phase of data input that many nuances of the data became apparent. As crucial as it was to put down ground rules on how the data must be encoded by each team member prior to the start of the data input phase, it was only possible to make most of the specific decisions concerning the data input protocol once data collection had started, and questions were arising from the nature of data itself. This section documents this important phase of the database's construction.

In preparation for data collection and data input we first had to select the varieties included in the database. This selection was made through an extensive review of the existing literature by the team under the careful supervision and expertise of the project's PI, Stephan Procházka. Each team member was assigned specific regions according to their expertise and research topic wherever this was possible. The team members then studied the regions for which they were responsible to make sure to include the relevant varieties. Relevance was determined by several factors. First, owing to the research questions and goal of the WIBARAB project, we

included as many traditionally Bedouin varieties as possible, while also including well-documented varieties of sedentary communities for comparative purposes. Secondly, it was important to make sure that all areas of the Arabic speaking world were covered. The third factor was of a rather practical and yet substantial nature: the availability of data. This selection of the varieties covered by the database was, naturally, heavily restricted to those varieties that have been previously documented and studied, or to those undocumented varieties that are the object of study of our team members. After selecting the relevant varieties, they were distributed among the team, so that each team member is responsible for an average of 35 varieties.

As soon as the skeleton of the data model was established, the team started to input data into the brand new TEI feature-files. In the beginning, a few feature-files were used for trial runs, in order to see what challenges would come up and to test possible ways to enhance the workflow. After this refining phase, an efficient and thorough workflow was established. In order to be able to duly discuss and study every feature included in the database, the database team picks a small number of features for each data input cycle. The data input cycle is the course of encoding a feature, starting with the creation of its TEI document and ending with the completed input of all data relevant to this feature.

In the first step of each input cycle, after choosing a feature and holding preliminary discussions on it, the database team creates a feature-file, and the team members subsequently start the data input, which means they extract all relevant information from their sources and fill them into the *feature value observations* of the varieties for which they are responsible. When questions concerning theoretical aspects of the feature, or doubts on the data encoding arise, they are discussed in a designated team meeting, where a common solution is reached. The solutions are then implemented by all members of the team to ensure a unified and coherent modus operandi. As we will see in further sections, once data input is completed and all available data has been encoded in a feature-file, the status on the file is changed to “done” and the document is then validated against the database’s schema to be prepared for publishing. As soon as data input on one group of feature-files is finished, the next cycle of data input begins with several

new features. This process will continue until the data on all planned features is duly encoded and validated.

As for technical tools, the process of data input makes use of three main programs: the TEI Enricher, Visual Studio Code and GitHub.[24] The process itself consists of the following steps: first, the most recent version of the file is pulled with Visual Studio Code from the GitHub repository on the server to the local repository on the team member's personal computer. Then the feature file is opened with the TEI Enricher, and the data is put in. After working on a file, this file has to be first saved to the local repository and subsequently pushed to the GitHub repository with the help of Visual Studio Code. All the files are currently being stored in the GitHub repository, which is publicly available. This workflow was chosen for the data input both for collaboration and safety purposes, since it allows for several team members to work on the same document simultaneously, but it also keeps a backup copy of every document after any 'push' is done by any member of the group. In order to maintain an overview of all the team members' work and the different versions of our files, we use the Visual Studio Code extension Git Graph which allows us to track all new contributions to the database as well as the merging points of two versions of the same documents and possible merge conflicts.

[24] For more in-depth descriptions of these tools see section 2.4.1.

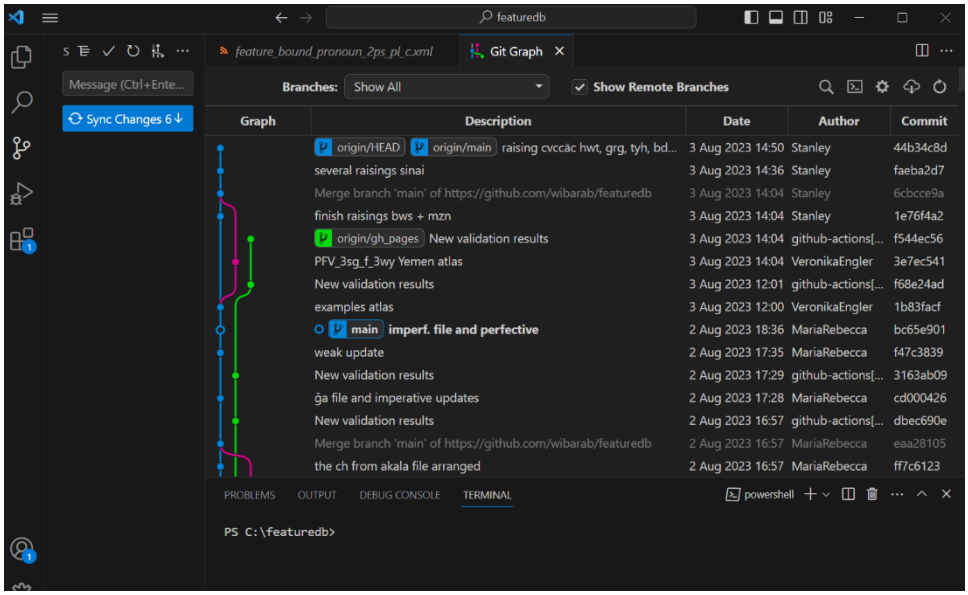

Figure 28 - Git Graph on Visual Studio Code tracking the changes that were made by different team members.

The importance of continuous deliberations on both different strategies for data input and the features themselves becomes apparent when looking at the manifold questions and challenges that can emerge during the data input process. In the following paragraphs, I summarize the most commonly encountered challenges in this respect.

1) **Lack of explicit mentions**: Sometimes, a source does not explicitly mention any value for one of the selected features. To solve this challenge, we agreed to check whether a value can be deduced from an example, and if so, we specify both the exact example, and a note stating that this information was extracted from an example and not from an explicit grammatical description.

   Example: We are interested in the distribution of the different reflexes of *ġayn* in the dialect of the Āl Wahība Bedouins of Oman. Its designated source (Webster, 1991) does not mention it explicitly but includes the word ġaḍaf (leaves of the date plam), which is featuring *ġayn*. We then include the value ġ → ġ, insert the example ‘ġaḍaf’, and mark in a note that this information is deduced from an example.

   If, however, there is no mention of the feature at all, it is impossible to deduce from examples, and there is no expert available to us who we can consult in a personal

communication, we have no information to include in the database and therefore exclude this particular datapoint by deleting the *feature value observation* for the variety in question.

2) **Mention of inexistent feature:** By contrast, when a source specifically mentions that a particular feature does not exist in the described variety, this information is included in the form of a feature value and referenced in the *feature value observation* accordingly. However, this is rather rare, since it is only done in features where we first establish that 'exists' and 'does not exist' are relevant categories because it is of interest whether a feature was retained or not, for instance the derived form IV in verbs.

3) **Mention of more than one value:** Often, a source provides multiple realisations of the same feature for the same location. In order to faithfully document this variation in our database, we need to include all given values. However, since each *feature value observation* can contain only one feature value, the solution for encoding several feature values is creating a new *feature value observation* element for each additional value. If the variation is not conditioned by any sociolinguistic factors – e.g. age, gender, religious or tribal affiliation – the original *feature value observation* can be duplicated and a new XML-ID generated. If, however, the variation is conditioned by sociolinguistic factors, this information must be encoded accordingly, making use of person group elements and in some cases note elements.
   Example: According to Cotter, 2022, speakers in Gaza use two different reflexes of *qāf*. This variation is conditioned by gender, i.e. the realisation of *q* as ʔ is used predominantly by women and the realisation as *g* predominantly by men. This is encoded by two *feature value observations*, each containing one feature value and a person group element which specifies to which group of speakers this feature value is attributed.

4) **General trends:** Alternatively, sometimes we observe some general trends with respect to a feature during data input, e.g. many sources reporting similar exceptions, such as the letter *qāf* being pronounced *q* in loanwords from Modern Standard Arabic in a wide

range of different varieties. In our deliberations, we decided to include remarks on these general observations in the chapter of the feature instead of repeating them systematically within each *feature value observation*.

Often, communication among the team members during the process of data input is not only important to solve more substantial issues, but it is also useful to discuss and clear up occasional doubts or ambiguities connected to the content of some sources. On the whole, this continuous exchange and meticulous discussion of all arising questions ensures that the high standards of the data input are met, and as a result, guarantees the quality of the data. An aspect that generated ample and numerous discussions was the process behind determining a meaningful, yet limited, set of feature values. I discuss the process in the next section.

3.4.1.1. Determining Sets of Feature Values

Determining a meaningful set of feature values for each feature is essential for our database since they reflect our analytic approach to each feature and ultimately define how information about a feature is shown to the database users. Therefore, it is important to consider all factors that play a role in establishing the feature values for each feature.

On the one hand, the inherently varied nature of the features, especially given that they belong to different linguistic subareas like phonology, morphology, and syntax, generates the need for different types of values. For example, while the values of phonological features usually are the reflexes of a specific sound, for morphological features such as the imperfective endings of sound verbs, they can be inflectional suffixes. On the other hand, it is essentially our research questions on a given feature that determine on which aspect of the feature we want to focus and which consequently shape the values we include.

For instance, the imperative form of weak verbs for the 2nd person singular masculine can be analysed in different ways, depending on the focus of one's research. While in a phonological study it would be of interest to focus on the quality of the initial vowel or on the length of the final vowel, when establishing a set of feature values, we focused on what we considered relevant to answer the project's research questions. In this case, it is the overall pattern, mainly

to differentiate trends concerning the use of vCvC versus CCv structures, since the first has been identified as typically Bedouin by the literature.

Thus, it is important to highlight that the process of establishing feature values, i.e. choosing a theoretical framework to analyse a feature, is in itself a subjective interpretation of the data. A similar subjective interpretation of the data inevitably occurs when deciding how in-depth to analyse a certain linguistic feature. The mere choice of creating either one or more files per feature depends on the degree of granularity of data we need, which is proportional to the degree to which this feature and its investigation might shed light onto our research question. For instance the feature 'raising of a' was divided into six sub-features, each having its own file: 1) 'raising of a in a pretonic closed syllable before ā - CaCCāC', 2) 'raising of a in a pretonic open syllable before ā - CaCāC', 3) 'raising of a in a pretonic closed syllable before ī - taCCīC', 4) 'raising of a in a pretonic open syllable before ī - CaCīC', 5) 'raising of a in a stressed open syllable before a - CaCaC', and 6) 'raising of a in a stressed closed syllable before a – maCCaC'. This division allows us to document and examine the phenomenon 'raising of a' in a meticulous and granular manner. In contrast, the conjugation of the imperfective in sound verbs is limited to two features: imperfective verbs 3rd person plural masculine and 2nd person singular feminine.

Needless to say, establishing a fixed set of feature values was not always possible before conducting a preliminary exploration of the data. Thus, we developed different approaches, depending on the nature of the feature and the feasibility of determining the values beforehand. The first and simpler approach consists of determining the feature values before the beginning of data input. It is used for all features where this is possible. In the case of features for which it is not possible to establish the feature values before data input, e.g. the personal pronouns, the second approach is used. This means that there are no predetermined feature values, and each team member creates new preliminary values according to what they find in the sources. Here the role of the source representation element is key, for it allows us to track the original value in the face of possible value set creation or restructuring. After data input is completed on these features, the values are reviewed and if any of the preliminary values show some overlap, they are grouped together into one value. Oftentimes, a blend of these two approaches

is the most pragmatic option, i.e. we determine a preliminary set of values before data input and then while inputting the data, we add as many values as needed before we then review and group them.

3.4.2. Revision and Validation

After the data input phase is completed, i.e. all the information on a given feature is encoded in the TEI feature-file, and the status of the file is changed to "done", the next step consists of a brief revision of the data and the validation of the document on a technical level. These two steps are done simultaneously by a member of the database team to find and rectify any technical errors that might have occurred during data input. Although all WIBARAB team members work on the TEI files as carefully and precisely as possible, it is inevitable in such a large endeavour that some technical errors occur. The errors can be detected through the validation function of the TEI Enricher which validates the feature-file against the XSD schema, which was generated from the ODD (see section 3.3.2.7.).

The revision of the data mostly consists of checking that no information is missing, i.e. none of the mandatory elements (containing feature value, source and location) are left empty and that all the links from the feature-file to the background files are checked to make sure that they work. In addition to this content related revision, the document must also be checked for errors on the level of the TEI standards and the 'rules' we established in our ODD. This includes checking for any double xml:ids of *feature value observations* and of examples, which are impermissible as each ID that is assigned to an item must be unique within the document. If a double ID is found in an element, a new ID must be generated. Another commonly occurring error is that the elements within a *feature value observation* were not put in the correct order as specified in the ODD. This can be rectified by re-sorting the elements in question. The rest of the errors are made up mostly of elements which were accidentally left empty or attribute values that are impermissible according to the ODD, for example when the @type attribute of the <personGrp> element contains "religiousAffiliation" instead of "religion," as is prescribed in the ODD. When, after correcting all errors, the document is valid, its status can be changed from "done" to "validated," signifying that the document is ready for publication.

3.5. Functions of the Database

As discussed earlier, the general aim of our database is to facilitate and enhance linguistic analysis and to make a wide range of data easily accessible and quickly comparable for scholars and anyone interested in variation within spoken Arabic. This database should thus not only enable the WIBARAB team to answer the project's research questions and confirm or refute the existence of common linguistic phenomena among Bedouin communities across the Arab World, but also enable other researchers to answer other major questions in Arabic dialectology, e.g. the Maghreb-Mashreq divide, religion as a linguistic factor, etc. This will hopefully be achieved through the database's functions which I will describe in this section.

3.5.1. Query Function

The database's main function will be the query function that interacts with a map and allows for both simple and combined queries, the results of which will be visualised on a map for an overview. In order to cater to the needs of both broad and more particular questions, it will be possible to query for several features at the same time, but also for specific feature values from one or more features simultaneously. The query function will enable users to generate paradigms and search for feature bundles. In addition to the "big picture" provided by the visualisation of the query results on the map, it will be possible to look at the particular information of a given location which is provided by the *feature value observation* in question.

3.5.1.1. Simple Queries

A simple query, in contrast to a combined query, is a query that is based on one component of the database, i.e. one feature, one tribe, or one location etc. The results of simple queries can be varied and used in different ways, depending on the basis of the query. For example, when one searches for a location, tribe, or any other group of speakers, the result will provide information on all features available for this location or community, generating a basis for a grammatical description.

In addition to the possibility of querying for a location or a group of speakers, the simple query allows for detailed searches within a feature. This means that it is possible to query for different values of a feature. For instance, within the feature '*reflexes of qāf*' it is possible to search for values with affrication versus values without affrication as illustrated in Figure 29. To improve the visualisation of the map and to help getting a clear, comprehensible search result, the icons that represent the feature values in the map are customisable. This means that database users can assign every value the specific shape and colour they find most appropriate, depending on the nature of the research question behind the query. For example, in Figure 29, all affricated values of *q are represented in pink and all the non-affricated values in blue, in order to obtain an immediate first overview of the geographical distribution of affrication of *q. The possibility to search for single feature values or specific combinations of values within a feature provides highly granular information as results.

This function both enables the exhaustive study of the features, and also provides detailed data on the varieties covered by the database. This can give insights into the classification of Arabic dialects, dialect groups and the Bedouin-sedentary split, especially by looking into the details of features that have been traditionally used to distinguish between dialect groups, e.g. the reflexes of /q/.

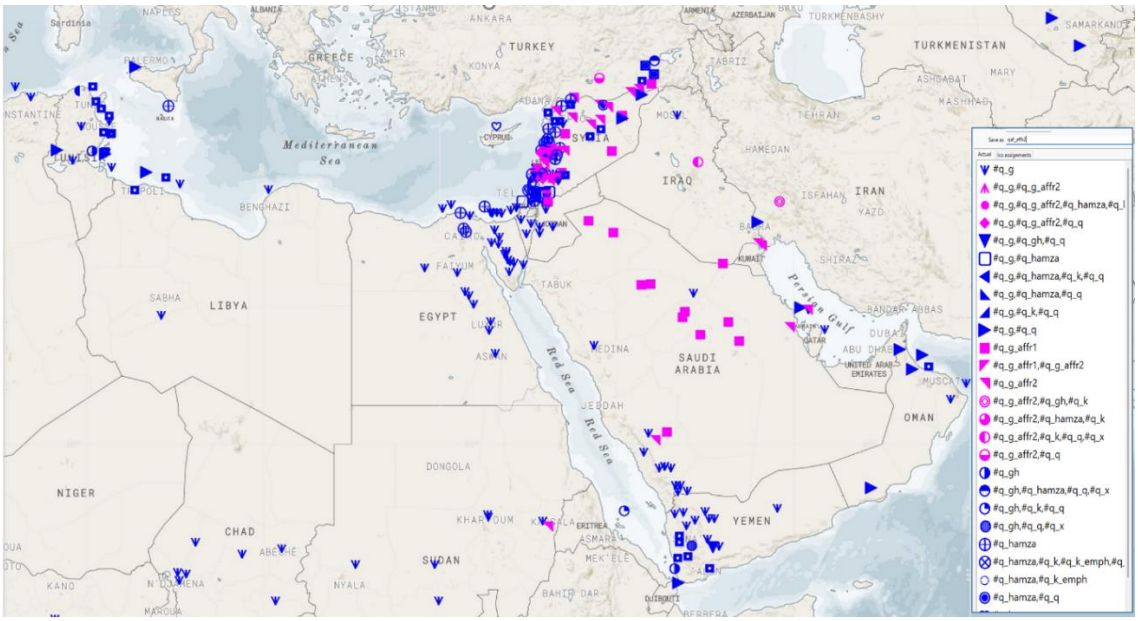

Figure 29 - Query for the feature values of qāf, blue are values without affrication, pink are values with affrication.

3.5.1.2. Combined Queries

In addition to simple queries, our database will be able to process combined queries. A combined query is a search of two or more components of the database. Components that can be queried through the combined query function include among others: features and specific feature values, locations, entries connected to a specific reference, specific person groups (tribe, religion, gender, age group, first language), or entries connected to the same language profile. These components can be combined in any constellation the user wishes. In the following, I will provide a few examples to illustrate the potential of the combined query function.

For instance, since the sources are provided with the decade of data collection, one could search for all entries of a feature connected to sources where data collection took place in a specific period. By querying for the feature plus two different data collection periods, one could compare the feature values recorded in these two periods and examine the correlation, if there is any, between the time of data collection and the occurrence of specific feature values.

Other examples of combined queries contain searches for a group of features in combination with specific person groups or locations. For example, one could compare the feature values attested for women/men, Sunnis/Shiites, young/old people, etc. Or one could compare the feature values attested for different traditional 'dialect areas.'

A further obvious example of a combined query would be a query of specific values from different features, e.g. a query for affrication in both *qāf* and *kāf* to look for correspondences. Since the combined query function allows the search for more than two features at once, this will also enable the user to look at and determine relevant bundles of features that can benefit a great variety of research questions.

Since the database still lacks a user interface, most of the combined queries cannot be executed yet, and while the TEI Enricher is in theory able to draw up maps that show the values of two features, in practise the visualisation is still very flawed: the feature values of the second feature are put over the feature values of the first feature in a way that the former often fully cover the latter, thus making them invisible. Therefore, I unfortunately cannot provide visual examples

of combined queries in this thesis. However, in the long run, this will be the most useful function of the database for a wide variety of comparative research questions.

3.5.2. Supplementary Information Provided by the Database

In addition to the data that can be accessed through the query function, the database offers several other types of information that can be helpful for the users. As mentioned in section 3.3.1.1., each linguistic feature contains a chapter where a literature review on the feature is offered along with specific information on the treatment of this feature on the database. Therefore, the chapters form a kind of "encyclopaedia" of features.

A resource of the database that makes use of its close connection to the VICAV infrastructure are the language profiles. The profiles will be directly accessible through the map, by zooming in to a particular location and following a link to the profile that describes the variety spoken in this location, as stated in the source. The profiles contain a classification of the dialect, information on the place (e.g. al-Khaburah in Oman)[25] or the community of speakers (e.g. the Khawētna tribe in Syria)[26] and the research history of the dialect. For both the profiles and the chapters, we will gratefully accept contributions from other experts of Arabic dialectology, thus a call for collaborations was announced at the first WIBARAB Symposium.

---

25 https://vicav.acdh.oeaw.ac.at/#map=[biblMarkers,_profiles_,]&1=[textQuery,vicavMission,MISSION,closed]&2=[textQuery,vicavNews,NEWS,closed]&7=[profileQuery,profile_khabura_01,al-Khaburah,open], accessed 23 October 2023.

26 https://vicav.acdh.oeaw.ac.at/#map=[biblMarkers,_profiles_,]&1=[textQuery,vicavMission,MISSION,closed]&2=[textQuery,vicavNews,NEWS,closed]&6=[profileQuery,profile_khawetna_01,Khaw%C4%93tna%20(tribe),open], accessed 23 October 2023.

What provides the database with yet more versatility is the possibility provided by the database for users to access TEI marked corpora and speech samples. However, this is only available for the varieties on which we have our own fieldwork data and corpora transcribed and translated by members of the WIBARAB team. Finally, the database will provide a bibliography section listing all the used references. Additionally, it will be possible to access VICAV's bibliographies which, in total, contain more than 5,000 bibliographic items.[27] With these functions, the WIBARAB database will serve both as a specialized archive with data from a wide range of Arabic varieties, and as arguably the most comprehensive and user-friendly digital analytical tool available to date for the investigation of Arabic dialects.

3.5.3. How the Database's Functions Will Contribute to Answering WIBARAB's Research Questions

The WIBARAB database's analytical functions, most importantly the combined query function, will substantially contribute to answering the project's manyfold research questions. Therefore, the following section contains a few examples and ideas of how the database's functions will help answer WIBARAB's research questions as outlined in section 1.1.2.2.

Naturally, examples of combined queries that will prove especially useful for answering the project's research questions are queries concerning the Bedouin-sedentary split, e.g. a query for a linguistic feature combined with the traditional classification of the variety, in order to see if the data is actually consistent with the typological descriptions of Bedouin Arabic. For example, combined queries of 'a feature + nomadic populations (i.e. tribes)' will provide data that will serve as a basis for further investigation which will help to shed light onto our question

[27] For a more detailed description, see section **Fehler! Verweisquelle konnte nicht gefunden werden.**.

to what degree the social dichotomy 'nomadic versus sedentary' is mirrored in the language (see section 1.1.2.2. point 1). Similarly, a comparison of the data resulting from queries of both 'linguistic feature + nomadic population' and 'linguistic feature + traditional classification' will provide a basis for the planned **comparative survey of actual and/or alleged grammatical and lexical features of Bedouin-type Arabic** (see section 1.1.2.2. point 4).

Moreover, the fact that the combined query enables determining and researching feature bundles can help answer the **question how conservative Bedouin-type Arabic really is** (see section 1.1.2.2. point 2). In his presentation at the 1st WIBARAB Symposium in June 2023, Stephan Procházka proposed to look at a bundle of 29 features as a preliminary "conservativeness index" (Procházka, 2023). One of the ideas is to create possible bundles of features for conservatism and for feature values of different features that have been identified as typical for "traditional dialects"[28] and carry out searches accordingly. This will not only provide data for the question how conservative Bedouin-type Arabic really is, but in addition to the topic of conservativeness, we also expect to find feature bundles that show linguistic tendencies concerning the Bedouin-sedentary split.

More generally, the database's analytic functions will allow for regional comparisons, e.g. a comparison of several dialect areas that were in contact with Bedouin communities to a varying degree. These comparisons will contribute to the examination of how this contact influenced linguistic realities in order to answer questions about **language and dialect contacts in the past and the present** (see section 1.1.2.2. point 5). Additionally, the database will provide the **new linguistic data from different regions** (see section 1.1.2.2. point 3) which were collected by the WIBARAB team.

[28] For a definition and discussion on "traditional dialects" see Herin, 2019.

However, although our database was mainly developed to assist in solving WIBARAB's research questions and backing up our findings with data, it will undoubtedly facilitate comparative analyses on various other topics in Arabic dialectology and linguistics in the future, as I have already mentioned before.

3.6. Next Steps in the Construction of the WIBARAB Database

As mentioned in section 2.2., this study can only cover the first two years of the WIBARAB database's development. At this point in time, we have succeeded in developing a flexible and methodologically sound data model and are in the middle of the data encoding process. The next steps, after finishing the data input and validation, will entail the development of the database's user interface and, finally, the online publication and launch of our database. While the latter is yet in the future, the former is already being worked on by our partners at the ACDH-CH.

One of the most significant challenges regarding visualisation of the map is posed by the distinction between precise locations and regions. This distinction is necessary sometimes, because some of the database's sources treat regional dialects without specifying locations. For example, Qafisheh (1977) talks about 'Gulf Arabic.' While it is simple to visualise a precise location on a map with an exact pin, a region such as 'The Gulf' is more difficult to depict. This is due to the often-unclear borders of regions and due to the fact that the GeoNames database from which we obtain our geographical data (i.e. coordinates), for each region only provides coordinates for one specific location within this region. VICAV somewhat circumvents this problem by displaying pins for precise locations and circles or clouds "somewhere inside the region we want to describe" (VICAV, n.d.: Bibliography Details). Whether we will adopt this approach or find a maybe more elegant solution for our database will depend on the outcome of more substantial deliberation.

Another issue that is developing at this stage of the database's construction is the visualisation of the features on the map. Many of our features have a considerable number of values, and at this point, we intend to represent each value by a specific symbol. This plan poses a

considerable challenge, especially considering that queries will often not be simple but combined (see section **Fehler! Verweisquelle konnte nicht gefunden werden.**) and c onsidering that two or more feature values can and do often occur in one single location (e.g. six different values of *reflexes of q* have been found in Karantina). These factors multiply the potential combinations that need to be shown in the map and can render the illustration rather chaotic. A possible solution to this problem is demonstrated by the Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al., 2013), which uses multi-coloured icons, like mini pie charts, to display multiple values, an approach the WIBARAB team might choose to adopt.

To sum up, now that we are running the first trial queries with the map, it becomes clear to us that a solution on how to best visualise our data will no doubt require much more discussion and thought. For this, we not only plan to consult more specialists from the Austrian Centre for Digital Humanities and Cultural Heritage, but we also plan another consultation with the TEI specialist Laurent Romary in December 2023.

4. Chapter 4: Linguistic Analysis

In this chapter I provide an example of the linguistic analyses that could be potentially carried out by means of the WIBARAB database. Due to the fact that the database is at this point still under construction, its analytic functions cannot yet unfold their full potential. At the moment, not all query functions are available yet. In some cases, this is due to a lack of data resulting from the unfinished phase of data input and in (most) other cases for technical reasons as a consequence of the absence of a user interface. The latter is the reason why I have executed all analyses presented in this chapter in the TEI Enricher. Therefore, I want to highlight here that the design and visualisation of the map is a preliminary one and might differ somewhat from the final user interface of the finished database. An example for this is the fact that, in the current map, the icons that represent the feature values in the locations where they were documented work like the heads of pins that do not vary in their size. This means that when zooming out, the icon appears not directly on the location, but next to it, resulting in a certain inaccuracy in the visualisation. For example, in the case of Malta, its icon seems when zoomed out to be placed in the Mediterranean Sea instead of on the island.

Figure 30 - On the left, a pin on the island of Malta; on the right, the same pin visualised in the Mediterranean Sea when zooming out.

There are some other minor issues with the current map, for example in a few rare cases, when we have two varieties which are linked to two different language profiles, but which are linked

to the same location, the current map sometimes has difficulties in displaying this, by not knowing which value/combination of values to display. Although this might in a few cases lead to somewhat blurry details, it does in no case change the general picture. Therefore, the usefulness of working with the map even now in its preliminary state far outweighs the small issues that come with this preliminary state. However, it is important to be aware of this limitation and ask the reader to keep in mind that the database is still a work in progress,

To sum up, since the database is not yet equipped with a user interface, I present here the results of a preliminary analysis which was executed in the TEI Enricher (see section 2.4.1.1). Therefore, it is only somewhat representative of the analytic functions the database will contain as soon as it has a fully operational user interface. Nevertheless, I decided to include such an analysis into the thesis mainly for two reasons. First, I want to illustrate that even with only the basic analytic functions the database possesses in its current state, it already is a tremendously useful tool for linguistic analysis as it already provides access to a relatively large set of data, it allows to search this data for specific purposes, and to visualise the results on a map. Second, this thesis and in particular the analysis in this chapter serve as a trial run for the WIBARAB database, since this is the first time an analysis of this scale has been attempted with it.

4.1. Linguistic Analysis of 'The Reflexes of *Qāf*' as a Case Study.

The Arabic letter *qāf* and its voiced pronunciation has long been considered the shibboleth of Bedouin-type Arabic. In the 15th century, the historian Ǧamāl Ad-Dīn Yūsuf Ibn Taġrībirdī wrote about the different pronunciations of *qāf* and the severe consequences these differences had on the lives of some speakers:

"They placed about 10,000 men in the middle and there was no one of them whose property they had not taken and whose wives they had not captured. And if one of them claimed to be a villager (ḥaḍarī), it was said to him, "Say 'flour'!" And if he said dagīg with gāf like the Bedouins, he was killed and if he said daqīq with the normal qāf, he was released" (Ibn Taġrībirdī, 1963:153; translation by Stephan Procházka).

This longstanding importance of the feature qāf in distinguishing Bedouin-type varieties from sedentary varieties makes it a natural choice for a case study like this. In the following, I will first present general information on the feature as presented in the chapter of said feature (see section 3.3.1.1) and secondly use the map to analyse the data in order to shed some light on the following research questions: 1) How are the different realisations of *qāf* distributed over the Arabic speaking world? Are there areas that are more or less diverse than others? 2) How are the realisations of *qāf* q→g and q→g + affrication, traditionally labelled as Bedouin, geographically distributed? Does the distribution of this values allow us to establish correlations between different lifestyles and/or social groups? 3) Comparing different countries, what tendencies can we see? Do these local tendencies align with the tendencies of the whole Arabic speaking world?

4.1.1. The Chapter: A General Description

In order to provide some background information on the feature of *qāf*, and to give an example of how the chapters in the database look like, this section contains the database's chapter on *qāf*, written by Stephan Procházka and Ana Iriarte Díez:

> **"Reflexes of **q***
>
> Qāf is the name of the 21st letter of the Arabic alphabet. In both the traditional pronunciation of Classical Arabic texts (including the Quran) and in Modern Standard Arabic, /q/ is a voiceless uvular stop. Although this description stands true for the currently common orthoepy, historical evidence suggests that in early periods of Arabic *q was most likely a voiced consonant, i.e. [ɢ], which had a voiceless variant (for details see (Edzard 2009).
>
> **General remarks**
>
> The map for this feature shows the value that is generally used in the specific location according to the sources. However, there are several general remarks and exceptions concerning the different realizations of /q/. Due to their general nature, these remarks are considered in the present section […] yet not represented as different values in the map.

Loanwords

Commonly, loanwords from Classical Arabic and MSA tend to keep the reflex of /q/ as a uvular stop , regardless of the common local reflex. The same is true for certain place names and cultural and religious terms (in particular the name of the Muslim holy book, *al-qurʔān*), which also tend to not undergo this sound change. It must be also considered that the "standard" or "classical" realization of this reflex may be chosen or dropped depending on communicative registers.

Examples for in dialects where **q* is usually reflected as : Damascus muqabbilāt 'starters'; Cairo al-qāhira 'Arabic name of the city of Cairo'; however, there are also counter-examples such as the Arabic name of 'Jerusalem´ which is usually pronounced al-ʔuds (and not al-quds) in urban Palestinian dialects.

This can also be true to specific loanwords from other languages or varieties. For example, in the speech of the ʕAtīğ tribe in Lebanon (with usual shift *q > g) the uvular stop is kept in loanwords from Turkish such as in qāwūn 'melon' and qalāšīn 'socks' (Younes and Herin 2013: 37).

Assimilations

The values in the map do not represent punctual changes in the realization of /q/ that are due to assimilation processes in certain lexemes or roots. A striking example is the root *qtl in most dialects that exhibit the shift *q > g: due to the following voiceless stop, /q/ is realized as /k/ in this root, e.g. yiktul 'he kills' *yigtul - *yaqtul.

Sociolinguistic considerations

The realization of /q/ is often subject to different sociolinguistic implications implying that multiple values coexist in the same location. This is especially important for the feature /q/, whose different realizations hold an important degree of both overt and covert sociolinguistic values among different communities of speakers. However, generational, socio-economic and gender differences in the realization of /q/ will only be made explicit as a different value on the map when enough sociolinguistic studies back up this information. Otherwise, it will be mentioned in the notes section of every specific value.

For example, in the Jordanian capital Amman /q/ is stereotypically realized as by Jordanian men, whereas it is pronounced as by speakers of Palestinian origin and some Jordanian women (particularly educated young women)" (Procházka & Iriarte Díez, 2023).

4.1.2. The Values and Their Geographical Distribution

The wide array of variation within the pronunciations of /q/ in different Arabic varieties has been well documented and seems to date back in history (Fischer & Jastrow, 1980:52). In the varieties documented in our database we found nine different values.

Table 2 - All feature values found in the database.

q_g	/g/	/g/ IPA[g] only
q_g_affr1	/ǵ/>/g/	/g/ IPA[g] + affricate ǵ IPA[ʤ] in the environment of anterior vowel
q_g_affr2	/ǧ/>/g/	/g/ IPA[g] + affricate ǧ IPA[ʤ] in the environment of anterior vowel
q_hamza	/ʔ/	/ʔ/ IPA[ʔ] only
q_gh	/ġ/	/ġ/ IPA[ɣ]
q_q	/q/	/q/ IPA[q]
q_k_emph	/ḳ/	/ḳ/ IPA[kˠ]
q_k	/k/	/k/ IPA[k]
q_x	/x/	/x/ IPA[x]

However, in many locations, no single value was recorded but a combination of multiple values. This is due to the fact that variation is a common phenomenon and can have different reasons, namely communities from different socioeconomic or geographic backgrounds living in one place, language changes that are manifested by different realisations of one feature by different age groups, or for sociolinguistic reasons such as the connection of different realisations with prestige or affiliation to a social group.

Figure 31 illustrates the broad range of variation concerning the pronunciation of /q/, with circles in different colours representing the single values and different blue icons representing the various documented combinations of multiple values.

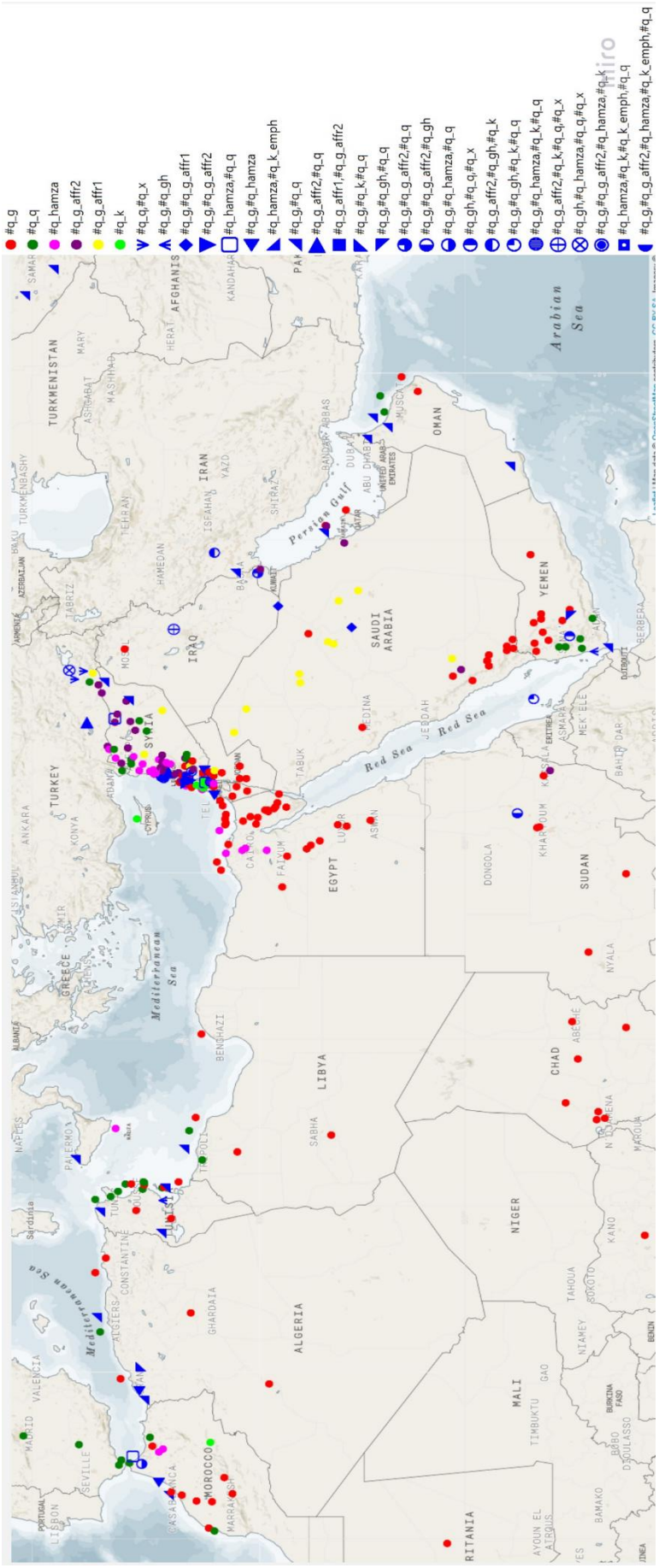

Figure 31 - Map of all database entries for /q/, the circles in different colours represent single values and the blue icons represent combined values icon.

4.1.2.1. q → q

The "classical" pronunciation of /q/ is that of a uvular stop. It is still found in many dialects, for instance, it is very common in the city and village dialects of the Maghreb. However, as we can see in Figure 32, q → q as a single value (green circle) is only represented 32 times all over the Arab World. However, it does not appear at all in the Gulf area, Saudi Arabia, Iraq, Jordan, Egypt, Eastern Libya, Sudan, and Sub-Saharan Africa. In addition to the occurrences as a single value, q → q in combination with another value was documented 27 times with an even wider geographical distribution. The only region where the realisation of qāf as q is completely missing is a large core area of the Arabic speaking world: Eastern Libya, Egypt, Sudan, Jordan, and Saudi Arabia (except for one occurrence in the Gulf).

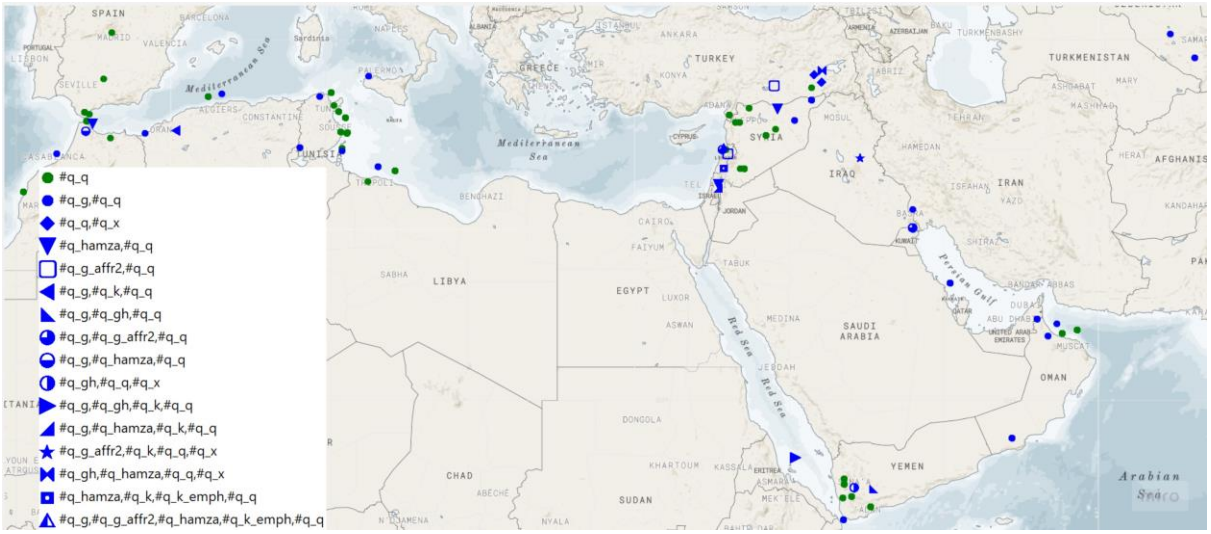

Figure 32 – The value q_q: combined values that contain q_q are shown in blue, the single value q_q is shown in green.

4.1.2.2. q → ʔ

A common reflex of /q/ is the voiceless glottal stop, which is typical of urban centres in the Levant and Egypt (e.g. Aleppo, Beirut, Damascus, Jerusalem, Cairo); however, it is also attested for other dialects. In our database, the value q → ʔ occurs both as a single value and as one of several variants. Figure 33 shows clearly that it is most prevalent in the Levant and lower Egypt where it was recorded in 33 locations, five of which are in Egypt and 28 in the Levant. Of the 28 occurrences in the Levant the majority (19) are as a single value, most of them in

Lebanon and Western Syria. In the other nine occurrences q → ʔ is one of several variants. The combined values occur mostly in Israel/Palestine, indicating a broad range of variation. The only other occurrence of q → ʔ in the Mashreq region beside the aforementioned locations, is one of four possible realisations of *qāf* in Siirt, Turkey.

With regard to the Maghreb (including Malta), the map shows a total of seven occurrences of q → ʔ, in three of which the glottal stop is documented as the only value, whereas the other four are cases of multiple values.

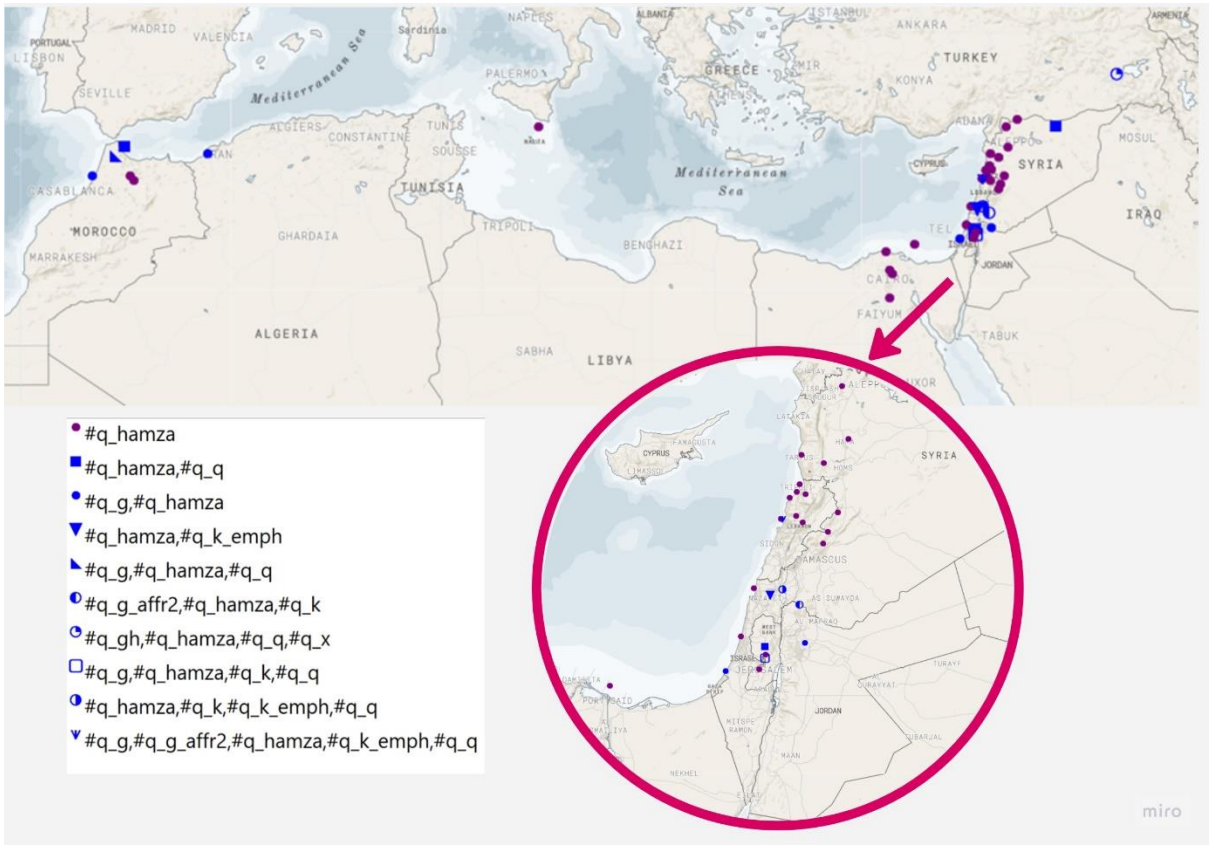

Figure 33 - The value q_hamza: the single value is shown in purple, the combined values containing q_hamza are shown in blue.

4.1.2.1. q → k, q → ḳ

The feature /q/ tends to surface as a voiceless velar stop in some dialects, mostly in Palestine; a rare feature is its velarized counterpart found in the Levant as illustrated in Figure 34. As we can see, q → k occurs both as a single value and in locations where it is one of several variants. As a single value, it occurs in nine different locations in Palestine/Israel (eight occurrences in the West Bank and one in Israel); and one occurrence each in Cyprus and in Tafilalet in Morrocco. As for combined values, it was documented in two locations in Israel/Palestine, in Irbid, Jordan as well as Baghdad, Iraq and Khuzestan, Iran and Saida, Algeria. This makes two occurrences in the Maghreb and 15 occurrences in the Mashreq, and one in East Africa, namely the Dahlak Archipelago in Eritrea.

Regarding the realisation of /q/ as ḳ, it is only documented in three places in the Levant: Karantina, Lebanon, and Nazareth and Tiberias, Israel. In all three locations it is only attested as one of several variants. The fact that it appears on the map only in those three locations and never as a single value indicates that it is a rare variant, and although it might appear in some specific lexemes in other dialects, these instances have been recorded as notes only, because there, it is used very sporadically. In general, a realisation is only displayed on the maps as a value when the source indicates it is used systematically.

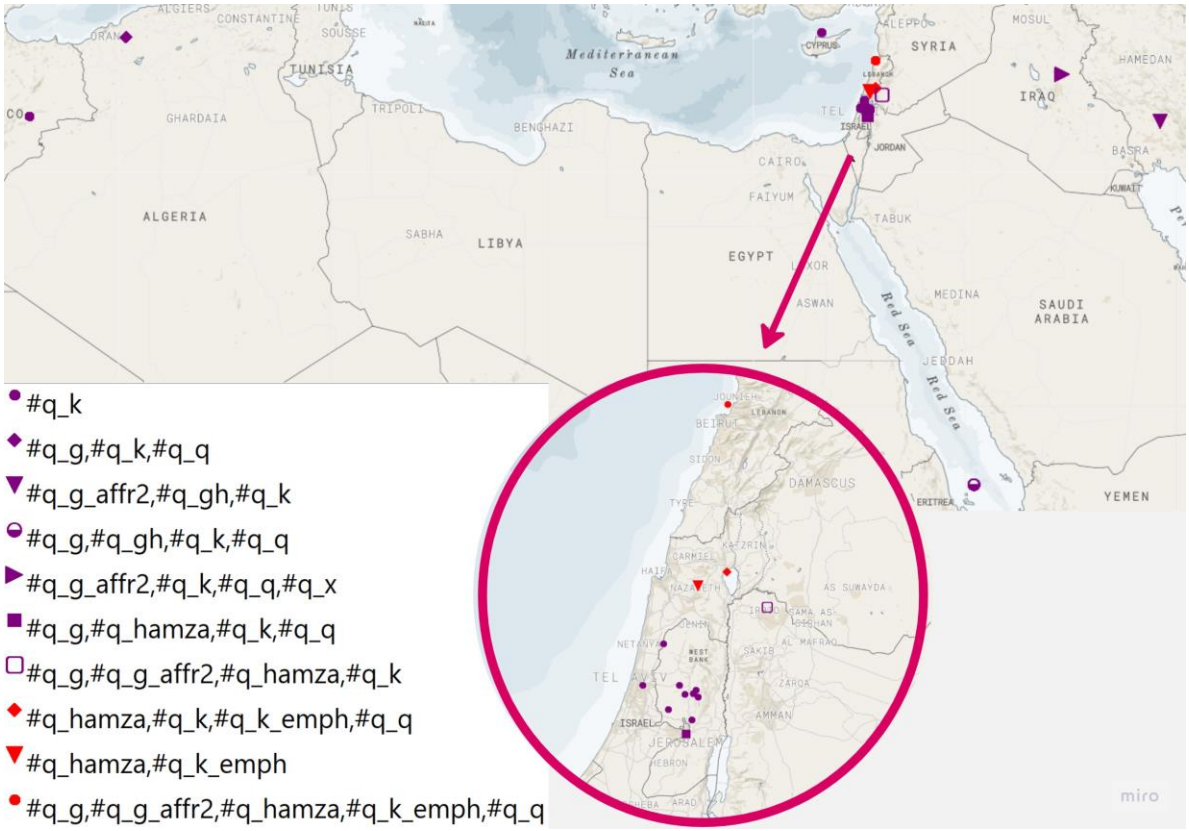

Figure 34 - Map of the values q_k and q_k_emph: the values containing q_k are represented by icons in purple, the values containing q_k_emph are represented by the red icons; the red rhombus represents a combined value that contains both q_k and q_k_emph.

4.1.2.2. q → x, q → ġ

The very rare realisation of /q/ as a voiceless velar fricative never occurs as a single value but is always combined with the value q → q (see Figure 35). This suggests that it is a fronted version of q. It is documented exclusively in the Mashreq, with a total of five occurrences, four of which are located in the Mesopotamian region (Southeast Turkey and Iraq) and one occurrence in Yemen.

The realisation of /q/ as a voiced velar fricative is slightly more used than its voiceless counterpart, totalling eight occurrences in the map, all of which are combined values. The occurrences are distributed over different regions but show a tendency to appear rather on the fringes of the Arabic speaking world: q → ġ appears three times in East Africa, once in Gabes, Tunisia, twice in Yemen and once each in Siirt, Turkey and Khuzestan, Iran.

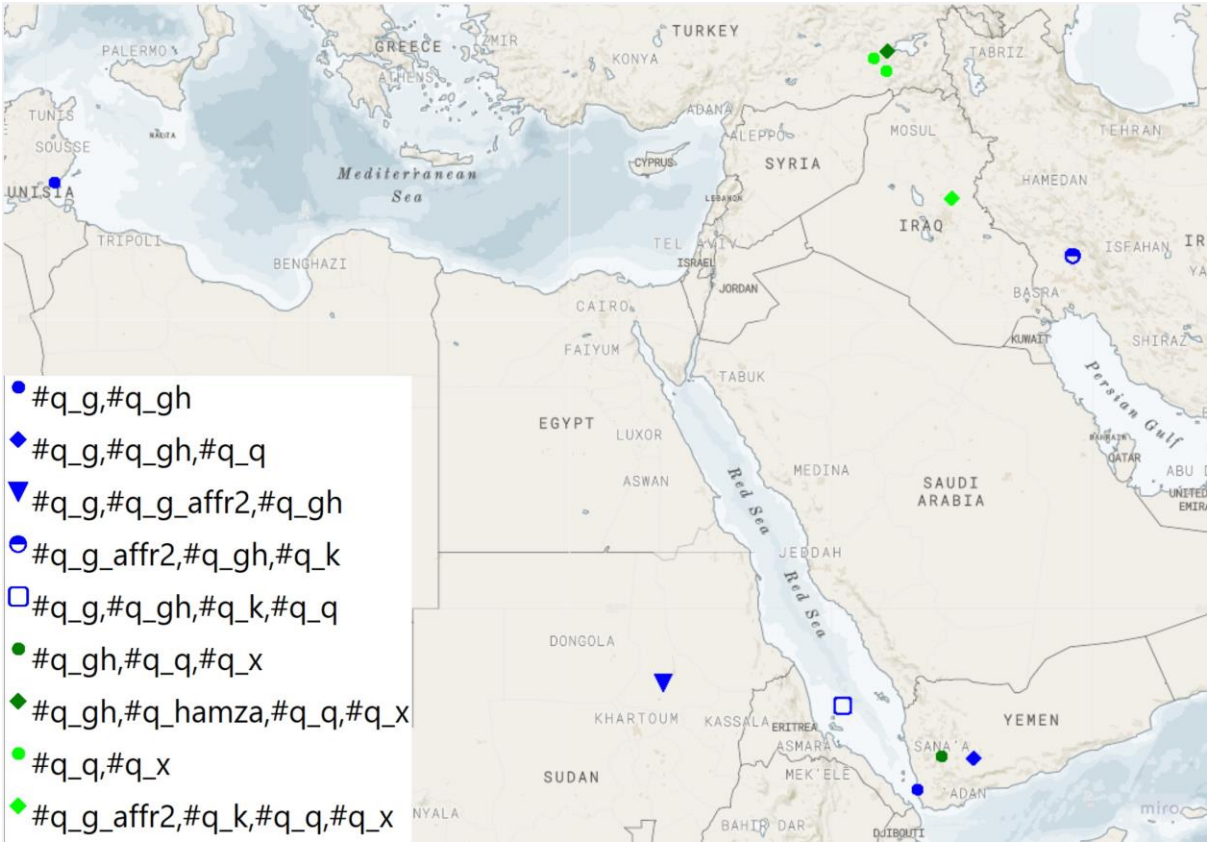

Figure 35 – Map for q_gh and q_x: values that contain q_gh are shown in blue, values that contain q_x are shown in lime, and values that contain both q_gh and q_x are shown in green.

4.1.2.3. q → g

The feature /q/ commonly surfaces as a voiced velar stop, which is regarded to be a hallmark of Bedouin-type dialects. The earliest explicit mention of the 'qāl-gāl' dialect split was made in the early 11th century by Ibn Sīna (d. 1037) (Blanc, 1966 as cited in Holes, 1991:667).

The map in Figure 36 illustrates that the realisation of /q/ as *g* is prevalent in the whole of the Arabic speaking world. Looking at the geographical distribution, it is clear that q → g as a single value is widely used, especially in sub-Saharan Africa, Egypt – particularly the Sinai -, Israel and West Jordan, and Northwest Yemen and Southwest Saudi Arabia. While it is also widely used along the North African coastline, there, we often find it combined with other realisations of /q/. The same can be said for the Levant and the Gulf area.

Overall, q → g was recorded 168 times in the database, making it by far the most widely used variant of /q/. Of these 168 occurrences, 59 are connected to a Bedouin tribe, thus q → g makes around 60% of all values connected to tribes, as we can see in Figure 39 in section 4.1.3. Additionally, of the 30 large cities investigated in this study (again, see section 4.1.3), in 14 of them q → g was recorded either as single value or in a combination of values, thus making up around 35% of all values recorded in these cities (see Figure 39). This means that while there is a tendency for /q/ to be pronounced g by members of originally nomadic tribes, this realisation is also frequently used in urban centres.

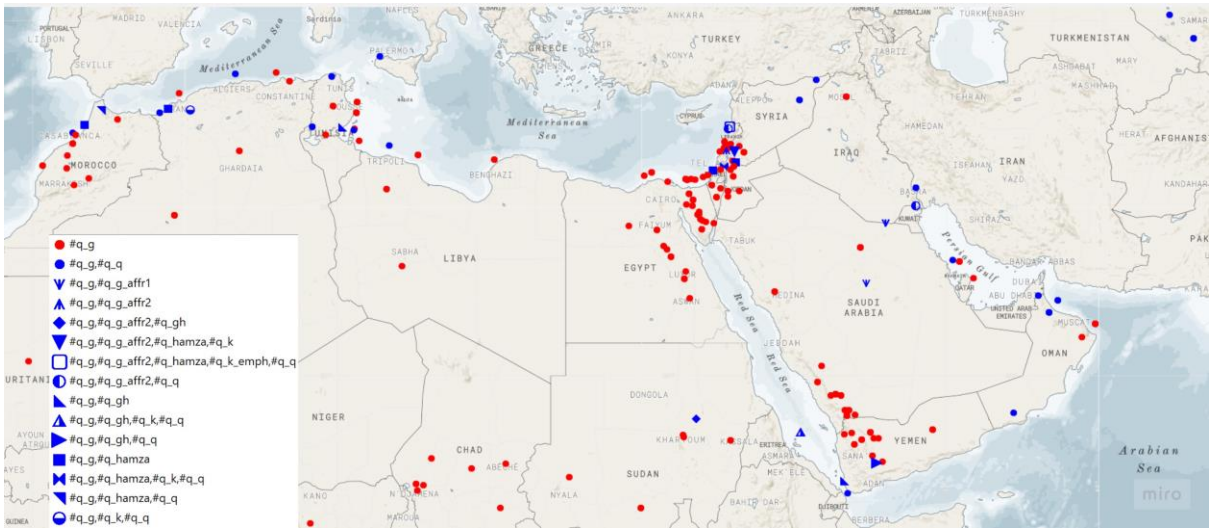

Figure 36 - Map that shows the value q_g: q_g as a single value is represented by the red circle, combined values that contain q_g are represented by blue icons.

4.1.2.4. Affrication

Many varieties of the Arabian Peninsula show affrication in specific phonetic environments additional to the realisation of /q/ as g: either *q > *g > ǵ [dz] (i.e. affrication 1) or *q > *g > ǧ [dʒ] (i.e. affrication 2). While the medieval Arab grammarians described a phenomenon they called *kaškaša*, Western scholars seem to be divided on whether to interpret these mentions as referring to both the feminine singular pronominal suffix –š(i) and the affrication of kāf and qāf to č/ć and ǧ/ǵ, or to the former only (Holes, 1991:652, , Johnstone, 1963:223). In modern Arabic dialectology, however, they are dealt with as two separate phenomena (Holes, 1991). The affrication of qāf and kāf is a widespread phenomenon and while Johnstone (1963)

attributes it mostly to Bedouin tribes, he also mentions several sedentary communities who use affricated variants. Holes explains the present geographical distribution of the variants with the two types of affrication or without affrication in the centre and eastern part of Arabian Peninsula and its verges "in terms of a chronological layering mirroring successive waves of migration from centre to the periphery of the peninsula" (Holes, 1991:665).

As Figure 37 shows, affricated variants occur only in the Mashreq region, the only exception being two occurrences in Sudan that are attributed to the tribe of the Rašayda, who according to Ingham have kept their Najdi dialect (Ingham, 1986:271) and thus seem to linguistically belong to the Mashreq rather than the Maghreb. This representation is also corroborated by WIBARAB team member Stanley Kochem, who conducted fieldwork with the Rašayda of Sudan earlier this year.

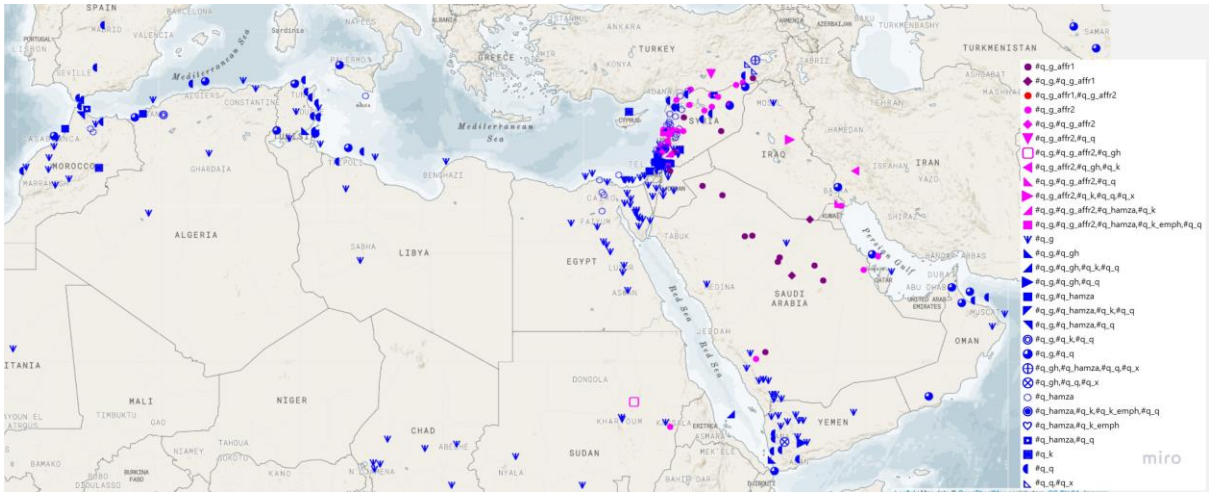

Figure 37 - Map of affricated variants of /q/: the blue icons represent non-affricated variants and purple, red and pink icons represent affricated variants.

Figure 38 shows more closely how the two types of affrication are distributed geographically. Generally, we can see that q → g + ǧ (47 occurrences) is more common than q → g + ǵ (17 occurrences), although this could also be due to fact that the literature is not necessarily documenting the areas in question evenly. This in turn influences the number of varieties covered in the literature and therefore in the database. However, it is clear that there are strong regional differences in the distribution of the two types of affrication. While q → g + ǵ is mostly

found in central Arabia and sometimes in the Levant, q → g + ǧ is mostly found around the edges of the peninsula, mostly in the Gulf and the fertile crescent. This data corroborates Holes's claim about layers of migration from the centre to the periphery each spreading the variant of /q/ that was in use in central Arabia at the time in question. According to him, q → g + ǧ must have been spread at the latest in the 18th century by migrations from Najd and the southwestern highlands of Saudi Arabia to the Gulf littoral, the Syrian desert and Mesopotamia, and q → g + ǵ developed in the region after these migrations (Holes, 1991:666f).

What is notable about the different types of affrication is that they usually don't occur together, i.e. a community of speakers uses only one variant of affrication (if any). However, an exception to this rule seems to be the Bani Ṣaxar tribe in Jordan, where a member of the WIBARAB team, Antonlla Torzullo, was able to record both variants among members of the tribe, who, however, belong to two different branches of the tribe (Torzullo, in publication). Unfortunately, due to the unfinished phase of data input, especially for varieties on which WIBARAB is conducting fieldwork, this is not visible in the map at this stage.

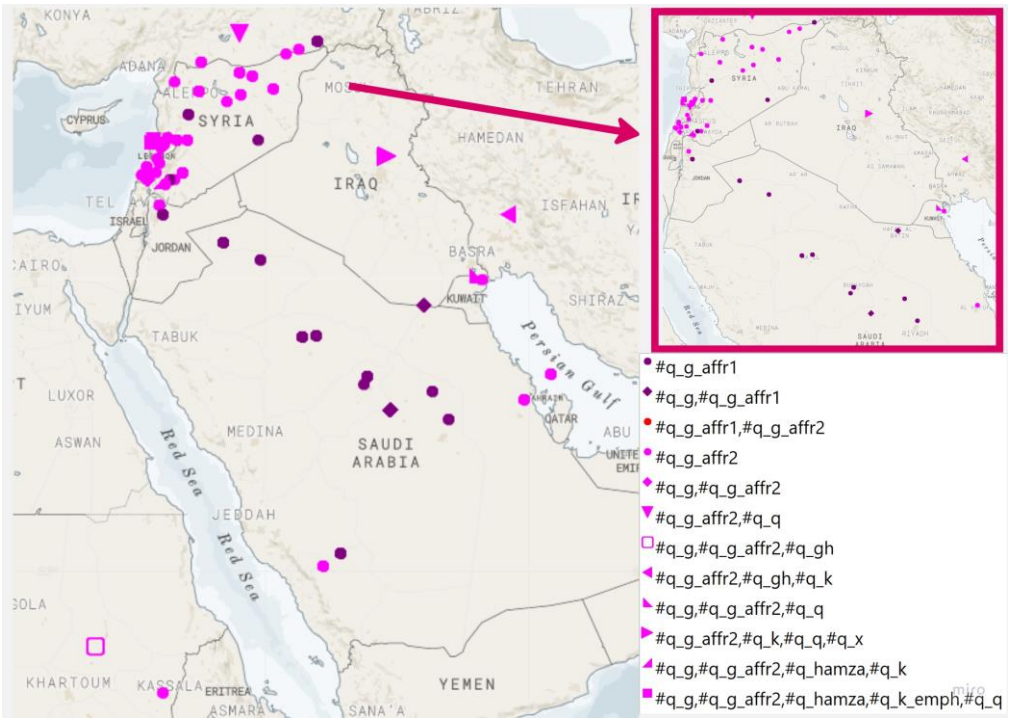

Figure 38 - This map shows the different types of affrication: affrication 1 is purple, a combined value of affrication 1 and two is red and different values containing affrication 2 are pink.

4.1.3. City (Sedentary) versus Tribe (Bedouin)

To shed some light on the question whether lifestyle is mirrored in linguistic realities, I chose 30 cities from around the Arab World to compare them to the varieties that are attributed to Bedouin tribes. I selected one or two large cities from each Arabic speaking country considering their size and the availability of data, i.e. for each country where data was available, I chose the capital and, in some cases, a second large city. Although these cities do not form a comprehensive list of Arabic speaking cities which would be necessary for an analysis on a larger scale, they are a sample that is representative of the diversity of the realisations of *qāf* and therefore suffice for the purposes of this study.

Table 3 - 30 cities in the Arab World: how their dialects are classified according to VICAV and the different realisations of /q/ that were recorded in them.

Casablanca (no entry)	q_q, q_g	**Cairo** (not specified)	q_hamza	**Baghdad** (not specified)	q_q, affr2, q_k, q_x, q_g
Rabat (not specified)	q_hamza	**Alexandria** (sedentary)	q_g	**Basra** (Bedouin-type)	q_q, q_g
Algiers (not specified)	q_q, q_g	**Jerusalem** (sedentary)	q_hamza	**Kuwait City** (Bedouin-type)	affr2
Jijel (sedentary)	q_g	**Ramallah** (not specified)	q_q, q_hamza	**Manama** (no entry)	affr2
Tunis (not specified)	q_q	**Amman** (not specified)	q_g, q_hamza	**Doha** (no entry)	q_g
Sousse (not specified)	q_q	**Irbid** (not specified)	q_g, q_hamza, affr2, q_k,	**Riyadh** (no entry)	affr1
Tripoli LBY (no entry)	q_q (Jews), q_g	**Beirut** (no entry)	q_hamza	**Abha** (not specified)	q_g
Benghazi (not specified)	q_g	**Tripoli LBN** (sedentary)	q_hamza	**Sana'a** (not specified)	q_g
Khartoum (sedentary)	q_g	**Damascus** (no entry)	q_hamza	**Aden** (sedentary)	q_q
Kadugli (no entry)	q_g	**Aleppo** (sedentary)	q_hamza	**Muscat** (no entry)	q_q

Looking at Table 3, we can see the general degree of variation of the realisation of /q/ in the Arabic speaking world mirrored also in its larger cities. Despite the diversity it is, however, possible to discern some regional trends. For instance, in the coastal cities of the Maghreb both q → q and/or q → g are the predominant variants recorded in our database. In the Sudanese cities Khartoum and Kadugli (which was included in this list due to the availability of data more than because of its size or regional importance) only q → g was recorded, which corresponds to the almost exclusive use of q → g in Eastern and Central Africa as illustrated by Figure 36 in section 4.1.2.3. Similarly but to a lesser degree, we see this kind of correspondence between the area in general and its larger cities in the Levant, where the cities, although showing a fair amount of variation, predominantly use q → ʔ, while the hinterland exhibits a widespread use of q → ʔ, but also the most variation on the entire map – a fact that might also be connected to the ample amount of research and studies done on Arabic varieties of this region that cannot be matched by any other region in its quantity.

With regard to the cities included in this analysis, Baghdad shows, on the first glance, the greatest diversity with five different variants of /q/, an impression that is put into perspective when looking into the data in more detail: while the Muslim majority of Baghdad's inhabitants use almost exclusively q → g, the small Christian minority uses four different values. With regard to the cities of the Gulf Area and the Peninsula, they do show some variation between themselves, yet it is notable that, with the exception of Muscat and Aden, all of them exhibit a voiced realisation of /q/.

In addition to geographical factors, it would also be worth looking at how the dialects of the selected cities are traditionally classified to determine whether there is a correlation between the realisation of /q/ and the classification of a dialect. At a later point in the development of the WIBARAB database, this will be possible as there will be a feature dedicated to the traditional classification of the different varieties included in the database. Unfortunately, however, the data collection and input phase has not been concluded for this feature in time to include it in this thesis. Still, in order to get a rough idea, I examined the VICAV language profiles for all 30 cities to see how their dialects are classified there. Unfortunately, only nine of the profiles contained a formal classification; of these, seven were classified as 'sedentary'

or ‘urban’ (Aleppo, Aden, Khartoum, Tripoli (LBN), Jijel, Jerusalem, and Alexandria), and two as ‘Bedouin-type’ (Basra and Kuwait). It is interesting to note that of the seven cities whose dialect is classified as ‘sedentary’, three (Alexandria, Jijel, Khartoum) have q → g as their sole realisation of /q/. In the case of Alexandria, this has to be put into context by adding, on the one hand, that the *feature value observation* containing this value also includes the comment “g is receding”, and on the other hand, that the source which reports this value (Behnstedt & Woidich, 1985) was published almost four decades ago. This makes it possible if not likely, that the then receding *g* has since receded and been replaced by the prestigious Cairene variant *ʔ*.

With regard to the data that is attributed to Bedouin tribes, it shows that there is a clear preference of voiced realisations of /q/. As we can see in Table 4, around 35% of the occurrences of q → g recorded in the database, around 50% of all recorded occurrences of q → g + affrication 1, and around 65% of all recorded occurrences of q → g + affrication 2 are connected to Bedouin tribes.

Table 4 - This table shows how often the different realisation of /q/ were recorded for Bedouin tribes and in the database as a whole, and how many percent the of all the values recorded in the database are attributed to tribes.

Value	Occurences	Tribes	Total
affr1	occurences	8	17
	percent	47,06%	
affr2		31	47
		65,96%	
q_g		59	168
		35,12%	
q_gh		1	8
		12,50%	
q_q		3	75
		4,00%	
q_hamza		1	43
		2,33%	
q_k_empl		1	3
		33,33%	
q_k		0	18
		0,00%	
q_x		0	5
		0,00%	

The distinct inclination of varieties linked to Bedouin tribes to use voiced realisations of /q/ that can be seen in Table 4, is reinforced when checking what percentage of the 104 Bedouin entries is made up by voiced variants of /q/. As illustrated in Figure 39, q → g, q → g + ǧ, and q → g + ġ make up almost 95% of all values connected to tribes. However, when looking at the cities, the picture is more varied. While the voiceless variants constitute a slight majority with around 54% of all values, the value that was recorded most often is still q → g.

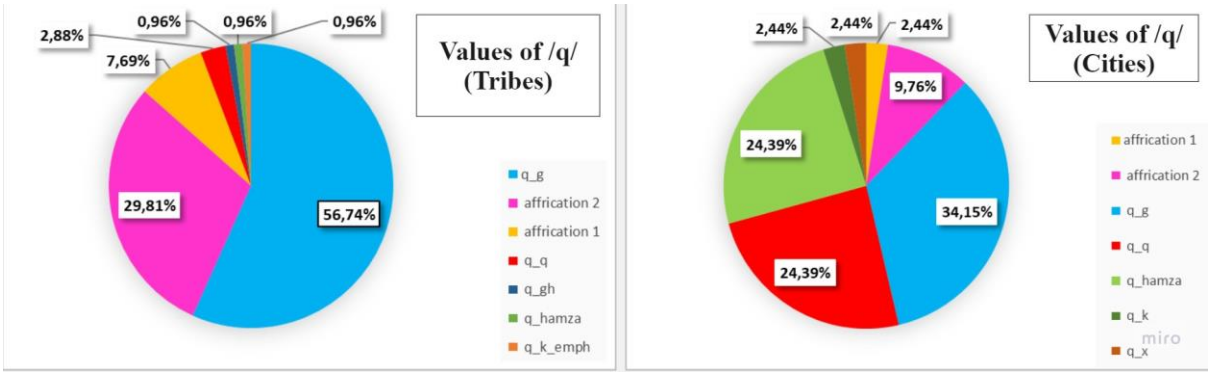

Figure 39 - This Figure shows the relative occurrences of the different realisations of /q/ between Bedouin tribes and large cities.

4.1.4. Country Comparison

4.1.4.1. Comparing Yemen, Morocco and Lebanon

I chose these three countries for a comparison mostly for their geographical location: Lebanon in the centre of the Arabic speaking world, and Morocco and Yemen on almost opposite ends of it.

While Yemeni Arabic is known for its vast diversity and unique characteristics (Vanhove, 2009) in general, this is not really confirmed when looking at the realisations of /q/, especially in comparison with Morocco and Lebanon. According to Figure 40 at the end of this section, there are four different realisations found in Yemen (two single values, two in combinations): q → g, q → q, q → ġ, and q → x. The first two are by far the most frequent and occur not only in Yemen but also in both Morocco and Lebanon. However, q → ġ and q → x are in this comparison unique to Yemen, where they only occur in combination with other values. It is important to note that the frequency in which the values occur varies significantly. The

realisation q → g occurs 15 times (14 times as a single value, once combined with other values), q → q occurs seven times (five times as a single value and twice in combinations), q → ġ occurs twice and only as a combined value, and finally, q → x occurs once as a combined value. According to the dialect areas described in Behnstedt's dialect atlas (2016:478), q → q is found in the southern part of the Tihāma region, as well as in the Ḥugariyyah, the Southern part of the k-dialect area, and Aden. The combined values featuring q → ġ and q → x are also found in the Ḥugariyyah, and the southern part of the k-dialect area. The rest of the recorded locations in Yemen, i.e central and northern Yemen and Wadi Hadramawt, all exhibit the value q → g.

For Morocco, the database recorded the values q → g, q → q, q → ʔ, and q → k. Although Morocco exhibits the same number of realisations as Yemen, they come in more different combinations: in addition to all four values occurring on their own, there were four different combinations of q → g, q → q, and q → ʔ; only q → k occurs just once and as a single value. However, the fact that q → g and q → q is shared with both Yemen and Lebanon, and q → ʔ is a shared value between Morocco and Lebanon, makes Morocco the country with the least unique traits in this comparison. Moreover, it is worth looking deeper into the data of Morocco, since by doing so it becomes apparent that some of the values are connected to specific groups of speakers. For instance, when looking at q → ʔ we find that out of the five occurrences, the majority are connected to specific person groups, namely Jewish communities and women. The realisation q → ʔ is recorded on the one hand for women in both Chefchaouen (males have q → q) and Masmouda (males have both q → g and q → q), and on the other hand for the Jewish communities of Sefrou (no other value recorded) and Rabat (for Rabat q → ʔ was also recorded for the Muslim community except the ZʕIr who have q → g). In addition to that, the only occurrence of q → k is linked to the Jewish community of the Tafilalet. Hence, we can say that q → ʔ and q → k are realisations of /q/ which are used predominantly by (functional) minorities, i.e. some women and Jews, while q → q and/or q → g are used predominantly by the (functional) majority population, i.e. Muslim men and some women. This shows the importance of an elaborate data model that allows us to store and process metadata about the recorded varieties, their speakers and locations.

In Lebanon, the database recorded five different values: q → ʔ, q → q, q → g, q → ḳ, and q → g + ğ. All of these values occur both as single and as combined values, except q → ḳ which is recorded only once in a combination. The most widely used variants in the varieties included in the WIBARAB database are q → ʔ and q → g + ğ, both occurring eight times. With regard to the geographical distribution of the values, we can see a tendency of q → g + ğ occurring in eastern Lebanon, close to the Syrian border, where the presence of originally nomadic and semi-nomadic tribes has been documented, whereas q → ʔ is recorded in northern Lebanon and rather closer to the coast. Indeed, when giving the data a closer look, it becomes evident that the actual distinction is not geographical, but pertaining to social group: while all eight occurrences of q → g + ğ are recorded for Bedouin tribes, all but one non-tribal communities use the reflex q → ʔ. This provides the clearest and most credible indication for a linguistic Bedouin-sedentary split of all the data examined in this analysis, even if only on a regional level.

The remaining three values are more rarely used. The value q → q was recorded four times, the value q → g three times, and the value q → ḳ only once. What is especially interesting here is the location of Karantina, a neighbourhood of Beirut, where WIBARAB team member Ana Iriarte Díez conducted fieldwork. In this location, all five variants occur. In fact, the realisation q → ḳ only occurs in Karantina and nowhere else in Lebanon. A possible explanation for this could be the fact that the area received a great influx of Palestinian migrants in the past 80 years and the reflex q → k frequently occurs in Palestinian rural varieties, which, however, does not account for the velarisation. Generally, when we have a closer look at the data, we see that the different reflexes are used by different age groups: q → g and q → g + ğ are used only by old and middle-aged speakers, q → q is used only in specific lexical items and not by young speakers, q → ʔ and q → ḳ is used by young and middle-aged speakers, but also occasionally by old speakers. This different usage across different age groups indicates that the language is changing. By the example of Karantina and the ample amount of sociolinguistic data that is available for this location, we can see how important the granularity of the WIBARAB fieldwork data is for understanding the underlying dynamics of complex communities.

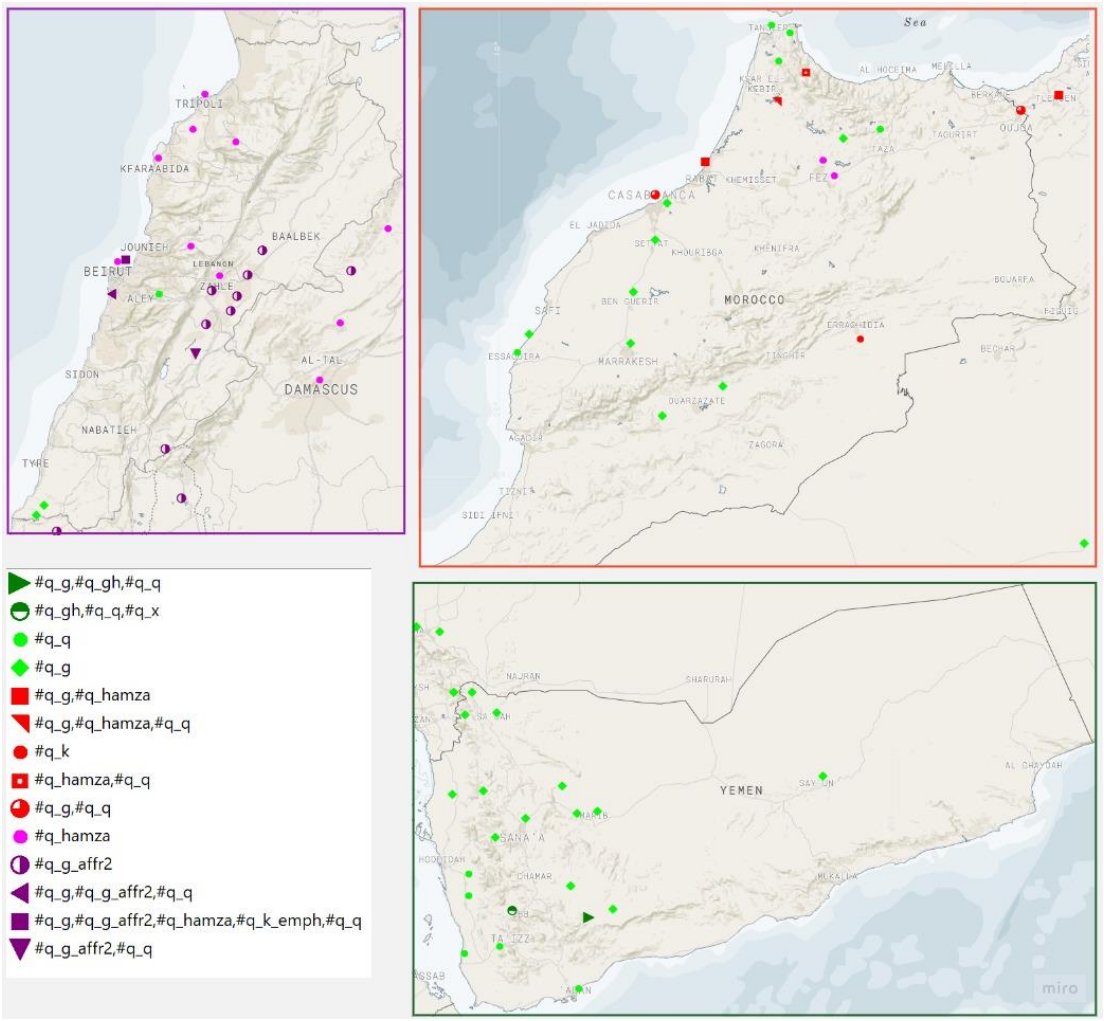

Figure 40 - Close-up of three countries: Lebanon, Morocco, and Yemen. The lime-coloured icons are shared between all three countries, the pink circle is shared between Lebanon and Morocco, the green icons represent the values found in Yemen, the red icons represent the values found in Morocco, and the purple icons represent the values found in Lebanon.

4.1.5. Conclusion of the Analysis

In this section I want to briefly summarise the conclusions of the analysis and answer the research questions that I formulated in the beginning of this chapter.

1) How are the different realisations of qāf distributed over the Arabic speaking world? Are there areas that are more or less diverse than others? Generally, we see a wide range of variation in most parts of the Arabic speaking world. However, the data in the maps show that there are some areas that are more diverse than others. Especially the Levant shows a large

amount of variation which indicates that in this region, diverse Bedouin, rural and urban communities have lived together in great proximity for a long time, but which might also be connected to the fact that the quantity of research done on the varieties in the Levant cannot be matched by any other region. A contrasting tendency can be seen both in East and Central Africa, and to a lesser degree on the North African coastline. Both regions exhibit only a small number of different realisations of /q/, except for a few locations each, which show more variation. Generally, we can see that the Mashreq is more diverse than the Maghreb: there are several values (q_x, affrication1, affrication2, q_k_emph) that are recorded only in the Mashreq, while there are no values that are unique to the Maghreb.

2) How are the realisations of qāf q→g and q→g + affrication, traditionally labelled as Bedouin, geographically distributed? Does the distribution of this values allow us to establish correlations between different lifestyles and/or social groups? While q → g is distributed all over the Arabic speaking world, reflexes with affrication are only found in the Mashreq, not in the Maghreb. Between the different types of affrication it is clear that q → g + ǵ is found in central Arabia and q → q + ǧ mostly in the north and east of the Peninsula.

With regard to the correlation between reflexes of /q/ and lifestyle/social identity, the data presented here indicates on the one hand, that lifestyle and social identity of speakers do not necessarily correlate strictly with linguistic classifications. This is shown for example by the varieties of urban centres whose populations live a sedentary lifestyle, but which still can be classified as Bedouin-type dialects and exhibit a "typically" Bedouin feature like a voiced realisation of /q/. On the other hand, we see that while usage of a voiced realisation of /q/ is significantly more prevalent in Bedouin speakers, it is by no means exclusive to them.

3) Comparing different countries, what tendencies can we see? Do these local tendencies align with the tendencies of the whole Arabic speaking world? In terms of geography, we can see that although Yemen, Morocco and Lebanon are located at great distance from each other, they share several values, especially Yemen and Morocco, both of which exhibit q →g and q → q as their main values. In contrast, each of the countries has one or two values that are not shared with the others. Generally speaking, when zooming in on different countries it becomes evident

how important it is to have metadata on the speakers in order to understand linguistic dynamics in detail. In terms of local tendencies compared to tendencies of the whole Arabic speaking world, we see for instance that while in Lebanon, there are strong indications that the Bedouin-sedentary split is mirrored in linguistic facts (at least concerning the realisation of /q/), this cannot be said about the Arabic speaking world in its entirety, based on this analysis.

These conclusions can be made by looking almost solely at the data and the map function of the WIBARAB database. Of course, in order to achieve a deeper understanding of the reasons why and how the different realisations of *qāf* and their geographic distribution developed in this way, further research would need to be conducted with additional tools and methods. This, however, is beyond the objective of this thesis whose main aim is merely to demonstrate how the database can assist a solid, data-driven linguistic analysis.

5. Chapter 5: Conclusions

In parallel to the two different dimensions of methodology of this thesis explained in chapter 2 (i.e. the methodological considerations underlying the database and the methods applied in this thesis), this chapter offers two different dimensions of conclusions. On the one hand the conclusions pertaining to the database itself, and on the other hand the conclusions pertaining to the thesis and its research objectives. Although the former were partly mentioned in the chapters and sections containing the documentations of the database and the analysis, I will briefly summarise them in this section, before moving on to the conclusions of the thesis.

5.1. Conclusions on the WIBARAB Database

The general aim of our database is to facilitate and enhance linguistic analysis and to make a wide range of data easily accessible and quickly comparable to scholars and anyone interested in variation within spoken Arabic. The main objectives of the WIBARAB database are therefore twofold: On the one hand, this database was primarily developed to assist in solving WIBARAB's research questions and backing up findings with data. Thus, it will help the WIBARAB team to answer the project's research questions and confirm or refute the existence of common linguistic trends among Bedouin communities across the Arab World. On the other hand, the database will undoubtedly facilitate comparative analyses on various other topics in Arabic dialectology and linguistics in the future. Thus, it will enable other researchers to answer other sets of relevant questions for Arabic dialectology and Arabic linguistics generally, e.g. on the Maghreb-Mashreq divide, on religion as a sociolinguistic factor, etc.

In order to achieve these objectives, the WIBARAB team developed a complex data model which was moulded by the methodological considerations that arose during the process. Thus, the development of the data model was shaped by three essential factors: (1) the overall wish to remain objective and avoid replicating authors' and/or researchers' interpretations, (2) the complex relation between the spoken varieties, locations and the speech communities, and (3) the diversity of the sources and the resulting diversity of the data. Considerations of the first and second factors provided the initial idea of the data model with the notion of *feature value*

observation at its heart. However, it was mainly the third factor which shaped the details of the data model to a greater extent. The essential purpose of the data model is to provide a 'template' or a 'skeleton' that is flexible and dynamic enough to be customised in order to faithfully reflect the complexity of the methodological and (socio)linguistic variation we found in the sources.

Although the database is not yet equipped with a user interface and therefore cannot yet unfold the full potential of its planned analytic functions, I included an analysis of the feature 'Reflexes of *Qāf*' into the thesis for two reasons. First, I illustrated that even with the basic analytic functions the database possesses in its current state, it is already a useful tool for linguistic analysis as it allows the user to access a relatively large set of data, to search this data for specific information, and to visualise the results on a map. Second, the analysis presented in chapter 4 of this thesis was the first analysis of this dimension carried out with the database and thus served as a trial run for the WIBARAB database. The analysis clearly shows the importance of the flexibility of our data model, which allows us to store and process complex data and metadata about the recorded varieties, their speakers, and their locations. The analysis also shows that the granularity of information provided by the complex data model, particularly with regard to the WIBARAB fieldwork data, is of great importance for understanding underlying dynamics governing linguistic variation.

5.2. Conclusions of the Thesis

Now, I will move on to the conclusions pertaining to this thesis and particularly its three research objectives: (1) documentation of the database and its construction, (2) illustration of the specificities of a TEI-based database, and (3) demonstration of the potential of the database's analytic function.

1) **Documentation of the Database and its Construction:** In this thesis, I provided a detailed description of the database and its construction. For this, I documented the stages in the construction of the databases. These can be divided into processes which have already been completed and their results; processes which are currently ongoing; and elements which are planned to be included into the database. The documentation

of the first stage of the database's construction includes a thorough description of the sources of the database, the challenges which arose in the process of developing the data model, the data model itself, and how it addresses these challenges. The documentation of the currently ongoing processes includes mainly the data encoding workflow with careful descriptions of the data input, and the revision and validation processes. Finally, the documentation of the elements that are planned to be included focuses to a large degree on the database's analytic functions, while the next steps for the construction of the database are only briefly outlined. This detailed description of the WIBARAB database and of the processes behind its creation does not only serve as a part of the general documentation of the whole WIBARAB project, but also aims in particular at contributing to the field by providing guidance for potential future databases dealing with Arabic dialectology. In fact, the aim to serve as a model for other databases not only includes databases of Arabic, but databases of any other language with similar characteristics. For this reason, in parallel to the technical details, I tried to include as many practical and visual examples as possible.

2) **Illustration of the Specificities of a TEI-Based Database**: Since (linguistic) databases can be built using different technologies, this thesis also aims to illustrate the specificities and advantages of TEI-based databases. The main characteristic of TEI is its flexible and extensible structure which allows highly customised data models. In the case of the WIBARAB database, this permitted us to shape the data model according to our needs especially in terms of granularity and the challenges that arose from the data and the complexity of the project's research questions. Specifically, this is demonstrated by our data model, which is based on the *feature value observation*, an element we customised particularly to fit both the type of information upon which the database is built, and the database's objectives. Another example of how TEI allowed us to tailor the data model to our need is the source representation element which provides information on the original spelling of a feature value and thus is used to reflect the different transcription systems found in the wide range of the database's sources. Similarly, the person group element makes it possible to preserve sociolinguistic information on specific groups of speakers, which is especially

important for in-depth analyses. In addition to TEI being flexible, it is also a community standard and was used in previous projects on Arabic dialectology such as TUNICO and VICAV. Therefore, it provides us with the opportunity to integrate our database into – and benefit from – existing research infrastructures. To sum up, the two main advantages of using TEI for the construction of the WIBARAB database are its flexibility and the fact that TEI is a widely used standard. Both of these advantages can be equally beneficial for future databases in the field of Arabic linguistics and linguistics in general.

3) **Demonstration of the Potential of the Database's Analytic Functions:** The potential of the database's analytic functions, even though they are not yet fully developed, can be seen partly when looking at the analysis presented in the thesis and mainly when looking at the planned functions. With regard to the question of how these functions will help answering WIBARAB's research questions, the thesis outlines first concrete ideas and approaches such as the possibility of examining feature bundles, comparing linguistic data with traditional classifications, comparing different areas or person groups, as well as the fact that the database provides access to granular data on a number of hitherto under- or unstudied varieties. While the database cannot be used to its full potential yet, the analysis presented in the thesis still demonstrates some of its analytic capacities: In the analysis of /q/, the database allowed me to compare data, both on a large scale and in detail, to isolate specific information and to find correlations within the data. Generally, with its description of the database's scope and our flexible data model, as well as the example of an analysis, this thesis demonstrates how the WIBARAB database has the potential to contribute not only to Arabic dialectology, but also Arabic linguistics, Arabic sociolinguistics and even general linguistics.

With these three objectives – the documentation of the database and its construction, the illustration of the specificities of a TEI-based database, and the demonstrations of the database's analytic functions – the thesis integrates both technical and linguistic aspects which clearly places it on the interface of Arabic linguistics and digital humanities and connects

established research fields with modern technologies. Therefore, this thesis operates as a bridge between the more traditional approaches of Arabic dialectology and newer ideas and methods of the digital humanities. To summarise, the thesis serves both as documentation and as a trial run for the database side of the WIBARAB project and as a contribution to the growing body of literature on digital humanities tools in (Arabic) linguistics.

6. Bibliography

6.1. Publications

al-Sharkawi, M. (2010). *The ecology of Arabic: A study of Arabicization*. Brill.

Al-Farabi, A.-N. M. I.-M. I.-T. (1969). *Kitāb al-ḥurūf* (M. Mahdi, Ed.). Dar El-Mashreq. https://usearch.uaccess.univie.ac.at/primo-explore/fulldisplay/UWI_alma21559823470003332/UWI

Asiri, Y. M. (2008). Relative clauses in Rijal Alma' dialect (South-west Saudi Arabia). *Proceedings of the Seminar for Arabian Studies, 38*, 71–74.

Bagatin, M. (2019). La variation linguistique selon Ibn Ḫaldūn. *Romano-Arabica*, *19*, 211–226.

Behnstedt, P. (1997). *Sprachatlas von Syrien. I: Kartenband*. Harrassowitz.

Behnstedt, P. (2016). *Dialect atlas of North Yemen and adjacent areas*. Brill.

Behnstedt, P., & Geva-Kleinberger, A. (2019). *Atlas of the Arabic dialects of Galilee (Israel): With some data for adjacent areas*. Brill.

Behnstedt, P., & Woidich, M. (1985). *Die ägyptisch-arabischen Dialekte. Band 2: Dialektatlas von Ägypten*. Reichert.

Behnstedt, P., & Woidich, M. (2010). *Wortatlas der arabischen Dialekte* (Vol. 100). Brill.

Benkato, A. (2019). From Medieval Tribes to Modern Dialects: On the Afterlives of Colonial Knowledge in Arabic Dialectology. *Philological Encounters*, *4*, 2–25.

Blanc, H. (1966). Les Deux Prononciations Du Qāf D'Après Avicenne. *Arabica*, *13*(2), 129–136.

Cadora, F. J. (1992). *Bedouin, village and urban Arabic: An ecolinguistic study*. Brill.

Cantineau, J. (1936). Etudes sur quelques parles de nomades arabes d'Orient 1. *Annales de l'Institut d'Études Orientales, 2*, 1–118.

Cotter, W. (2022). The Arabic dialect of Gaza City. *Journal of the International Phonetic Association, 52*(1), 122–134. https://doi.org/10.1017/S0025100320000134

De Jong, R. (2000). *A grammar of the Bedouin dialects of the northern Sinai littoral: Bridging the linguistic gap between the eastern and western Arab world*. Brill.

Ech-Charfi, A. (2018). The politics of language standardization and the nature of Classical Arabic. *Folia Orientalia*, *55*, 61–81. https://doi.org/10.24425/for.2018.124679

Edzard, L. (2009). Qāf. In K. Versteegh (Ed.), *Encyclopaedia of Arabic Language and Linguistics* (Vol. 4, pp. 1–3). Brill.

Fischer, W., & Jastrow, O. (1980). *Handbuch der arabischen Dialekte*. Harrassowitz.

Heath, J. (2002). *Jewish and Muslim Dialects of Moroccan Arabic*. RoutledgeCurzon.

Heine, B., & Kuteva, T. (2005). *Language contact and grammatical change*. Cambridge University Press.

Herin, B. (2019). Traditional dialects. In E. Al-Wer & U. Horesh (Eds.), *The Routledge Handbook of Arabic Sociolinguistics* (pp. 93–105). Routledge.

Holes, C. (1991). Kashkasha and the fronting and affrication of the velar stops revisited: A contribution to the historical phonology of the peninsular Arabic dialects. In A. S. Kaye (Ed.), *Semitic studies in honor of Wolf Leslau on the occasion of his eighty-fifth birthday, November 14th, 1991* (pp. 652–678). Harrassowitz.

Holes, C. (2013). An Arabic Text from Sūr, Oman. In R. de Jong & C. Holes (Eds.), *Ingham of Arabia: A Collection of Articles Presented as a Tribute to the Career of Bruce Ingham* (Vol. 69, pp. 87–107). Brill.

Ibn Taġrībirdī, Ǧamāl ad-Dīn. (1963). *An-Nuǧūm az-zāhira fī mulūk miṣr al-Qāhira.* Vol. 8, Cairo : Wizārat aṯ-Ṯaqāfa.

Ingham, B. (1986). Notes on the dialect of the Āl Murra of eastern and southern Arabia. *Bulletin of the School of Oriental and African Studies*, *49*, 271–291.

Jannidis. (2017). *Digital Humanities*. J.B. Metzler.

Johnstone, T. M. (1963). The affrication of "kaf" and "gaf", in the Arabic dialects of the Arabian Peninsula. *Journal of Semitic Studies*, *8*, 210–241.

Kaye, A. S., & Rosenhouse, J. (1997). Arabic dialects and Maltese. *The Semitic Languages*, 263–311.

Lombezzi, L. (2019). The Spoken Omani Arabic of 'Ibrī: A "Crossing Point" in Gulf Dialects. In C. Miller, A. Barontini, M.-A. Germanos, J. Guerrero, & C. Pereira (Eds.), *Studies on Arabic Dialectology and Sociolinguistics: Proceedings of the 12th International Conference of AIDA held in Marseille from May 30th to June 2nd 2017*. IREMAM. http://books.openedition.org/ iremam/3878

McNeil, K. (2019). Tunisian Arabic Corpus: Creating a written corpus of an "unwritten" language. In T. McEnery, A. Hardie, & N. Younis (Eds.), *Arabic Corpus Linguistics*. Edinburgh University Press.

Parkinson, D. (2019). Under the Hood of arabiCorpus. In *Arabic Corpus Linguistics* (pp. 17–29).

Pierazzo, E. (2015). Textual Scholarship and Text Encoding. In *A New Companion to Digital Humanities* (pp. 307–321). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118680605.ch21

Procházka, S. (2020). *WIBARAB - ERC Advanced Grant 2020: Research Proposal [Part B1]*.

Procházka, S. (2020). *WIBARAB - ERC Advanced Grant 2020: Research Proposal [Part B2]*.

Procházka, S. (2023, June 7-9). *Linguistic conservatism and Bedouin-type Arabic: Between myth and reality* [Paper presentation]. 1st WIBARAB Symposium, Vienna, Austria.

Prochazka, T. (1988). *Saudi Arabian dialects*. Kegan Paul International.

Qafisheh, H. A. (1977). *A short reference grammar of Gulf Arabic*. Univ. of Arizona Press. http://ubdata.univie.ac.at/AC07732304

Reinhardt, C. (1894). *Ein arabischer Dialekt, gesprochen in ʻOmān und Zanzibar*.

Shaaban, K. A. (1977). *The phonology of Omani Arabic* [PhD Thesis]. The University of Texas at Austin.

Torzullo, A. (2022). The Emergence of a Mixed Type Dialect: The Example of the Dialect of the Bani ʿAbbād Tribe (Jordan). *Languages*, *7*(1), 9. https://doi.org/10.3390/languages7010009

Torzullo, A. (in publication). New insights on a camel-breeder Bedouin dialect of Jordan. *AIDA (Association Internationale de Dialectologie Arabe) 14 Proceedings*.

Trudgill, P. (2009). *Contact, isolation, and complexity in Arabic*. In E. Al-Wer & R. de Jong (Eds.), *Arabic Dialectology—In honour of Clive Holes on the Occasion of his Sixtieth Birthday* (pp. 173–185). Brill.

Vanhove, M. (2009). Yemen. In Versteegh, Kees (Ed.), *Encyclopaedia of Arabic Language and Linguistics* (Vol. 4, pp. 750–758). Brill.

Versteegh, K. (Ed.). (2006). *Encyclopedia of Arabic Language and Linguistics. 4 vols.* Brill.

Watson, J. C. E. (1993). *A Syntax of San'ānī Arabic*. Harrassowitz.

Watson, J. C. E. (2011). Arabic dialects (general article). In S. Weninger (Ed.), *The Semitic languages: An international handbook* (pp. 851–895). De Gruyter Mouton.

Webster, R. (1991). Notes on the dialect and way of life of the Āl Wahība Bedouin of Oman. *Bulletin of the School of Oriental and African Studies, 54,* 473–485.

Wehr, H. (1979). *A dictionary of modern written Arabic (Arabic—English)* (J. M. Cowan, Ed.; 4th edition, considerably enlarged and amended by the author). Harrassowitz.

Woidich, M. (2006). *Das Kairenisch-Arabische: Eine Grammatik*. Harrassowitz.

Younes, I. (2018). Linguistic retentions and innovations amongst a camel-breeder tribe of northern Jordan. *Wiener Zeitschrift Für Die Kunde Des Morgenlandes, 108,* 265–274.

Younes, I., & Herin, B. (2013). Un parler bédouin du Liban: Note sur le dialecte des ʿAtīğ (Wādī Xālid). *Zeitschrift für Arabische Linguistik, 58,* 32–65.

6.2. Digital Resources

Abouzahr, H. (n.d.). *Lughatuna: The living Arabic Project*. Available online at https://www.livingarabic.com/. Accessed on 10.10.2023.

ArabiCorpus. (n.d.). Available online at https://arabicorpus.byu.edu/. Accessed online on 23.10.2023.

DARIAH Working Group. (2020). *TEI Lex-0.* Available online at https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html. Accessed on 30.10.2023.

Dryer, M. & Haspelmath, M. (eds.). (2013). *WALS Online (v2020.3)*. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at https://wals.info). Accessed on 20.10.2023.

GeoNames. (n.d.). Available online at https://www.geonames.org/. Accessed on 05.11.2023.

GitHub. (n.d.). Available online at https://github.com/. Accessed on 20.10.12023.

Magidow, A. (2015). *Database of Arabic Dialects.* Available online at https://www.database-of-arabic-dialects.org/. Accessed on 14.10.2023.

McNeil, K. & Faiza, M. (2010-). *Tunisian Arabic Corpus (TAC): 895,000 words*. Available online at http://www.tunisiya.org. Accessed on 23.10.2023.

Michaelis, S. & Maurer, P. & Haspelmath, M. & Huber, M. (eds.). (2013). *The Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. Available online at https://apics-online.info/. Accessed on 05.11.2023.

TEI Consortium, eds. (n.d.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.6.0. last updated 4th April 2023. TEI Consortium. http://www.tei-c.org/Guidelines/P5/. Accessed on 17.04.2023.

TUNICO. (n.d.) Available online at https://tunico.acdh.oeaw.ac.at/. Accessed on 15.10.2023.

Vienna Corpus of Arabic Varieties. (n.d.). Available online at https://vicav.acdh.oeaw.ac.at/ Accessed on 01.05.2023.

Visual Studio Code. (n.d.). Available online at https://code.visualstudio.com/. Accessed on 15.10.2023.

WIBARAB. (n.d.). Available online at https://wibarab.acdh.oeaw.ac.at/ Accessed on 15.09.2023.

6.3. WIBARAB Database

Procházka, S., Iriarte Díez, A. (2023). *Chapter for the Reflexes of Qāf.* Available online at https://github.com/wibarab/featuredb/blob/main/010_manannot/features/features_q.xml. Accessed on 19.10.2023.

Procházka, S., Mörth, C., Schopper, D., Iriarte Díez, A., Engler, V. & Doppelbauer, J. (2023). *WIBARAB FeatureDB ODD*. Available online at https://github.com/wibarab/featuredb/blob/main/802_tei_odd/featuredb.odd. Accessed on 25.10.2023.

WIBARAB Feature Database. (n.d.). https://github.com/wibarab/featuredb. Accessed on 26.10.2023.

7. Appendix

7.1. Abstract (English)

A vital part of the ERC funded WIBARAB (What Is Bedouin-type Arabic?) project is its TEI-based database which is being constructed to facilitate linguistic analysis of a great number of Arabic varieties. The WIBARAB database will include approximately 125 features, mainly from phonology, morphology, and syntax, but also a selected set of lexical and phraseological items. The finished database will contain linguistic data from approximately 300 publications, fieldwork campaigns and personal communications with experts on around 320 Arabic varieties. Many of these varieties are spoken by (formerly) nomadic communities due to the project's research focus on the Bedouin-sedentary split of Arabic. The WIBARAB database will not only be the first all-Arabic database of this size, but it will also be the first one to focus on the varieties spoken by Bedouins.

This thesis serves both as a documentation of the database and the first two years of its development, and as a trial run of the database's preliminary analytic functions. It contains a thorough documentation of the database's different types of sources and data, the methodological considerations and challenges that arose during the development of the data model, and a detailed description of the data model itself. It also describes the data encoding process and the analytic functions the database will ultimately have. Some of these functions are also demonstrated in an analysis of the linguistic feature 'reflexes of *qāf*' which focuses on the geographical distribution of the different reflexes of *qāf*, a comparison of urban and Bedouin varieties, and a comparison of three countries. The main objectives of the thesis are the documentation of the database and its construction, the illustration of the specificities of a TEI-based database, and the demonstrations of the database's analytic functions.

7.2. Abstract (Deutsch)

Ein wichtiger Teil des ERC-finanzierten Forschungsprojekts WIBARAB (What Is Bedouin-type Arabic?) ist die TEI-basierte Datenbank, die zur linguistischen Analyse einer großen Anzahl an arabischen Varietäten entwickelt wird. Die WIBARAB-Datenbank wird ungefähr 125 Features, hauptsächlich aus Phonologie, Morphologie und Syntax; aber auch eine Auswahl von lexikalischen und phraseologischen Phänomenen enthalten. Die Datenbank inkludiert im Endausbau linguistische Daten aus ungefähr 300 Publikationen, Feldforschung und persönlichen Kommunikationen mit Expert*innen zu rund 320 arabischen Varietäten. Aufgrund des Forschungsschwerpunkts des Projektes, der auf der Dichotomie zwischen Beduinen- und Sesshaftendialekten liegt, werden viele der inkludierten Varietäten in (ehemals) nomadischen Gemeinschaften gesprochen. Die WIBARAB-Datenbank wird nicht nur die erste Datenbank ihrer Größe, die ausschließlich arabische Varietäten beinhaltet, sondern auch die erste, die sich auf beduinische Varietäten spezialisiert.

Die vorliegende Masterarbeit dient sowohl der Dokumentation der Datenbank und der ersten zwei Jahre ihrer Entwicklung als auch als Probelauf ihrer vorläufigen Analysefunktionen. Die Arbeit beinhaltet eingehende Beschreibungen der verschiedenen Quellen- und Datentypen, der methodologischen Überlegungen und Herausforderungen, die sich während der Entwicklung des Datenmodells ergaben, und des Datenmodells selbst. Sie beschreibt auch den Prozess der Datenkodierung und die analytischen Funktionen, die die Datenbank schlussendlich besitzen wird. Einige dieser Funktionen werden in einer Analyse des linguistischen Features „Realisierung von *qāf*“ aufgezeigt. Die Analyse ist auf die geographische Verteilung der Realisierungen von *qāf,* den Vergleich urbaner und beduinischer Varietäten und einen Vergleich zwischen drei Ländern fokussiert. Die wichtigsten Ziele dieser Masterarbeit sind die Dokumentation der Datenbank und ihrer Entwicklung, die Illustration der Eigenschaften einer TEI-basierten Datenbank und die Demonstration der analytischen Funktionen der Datenbank.