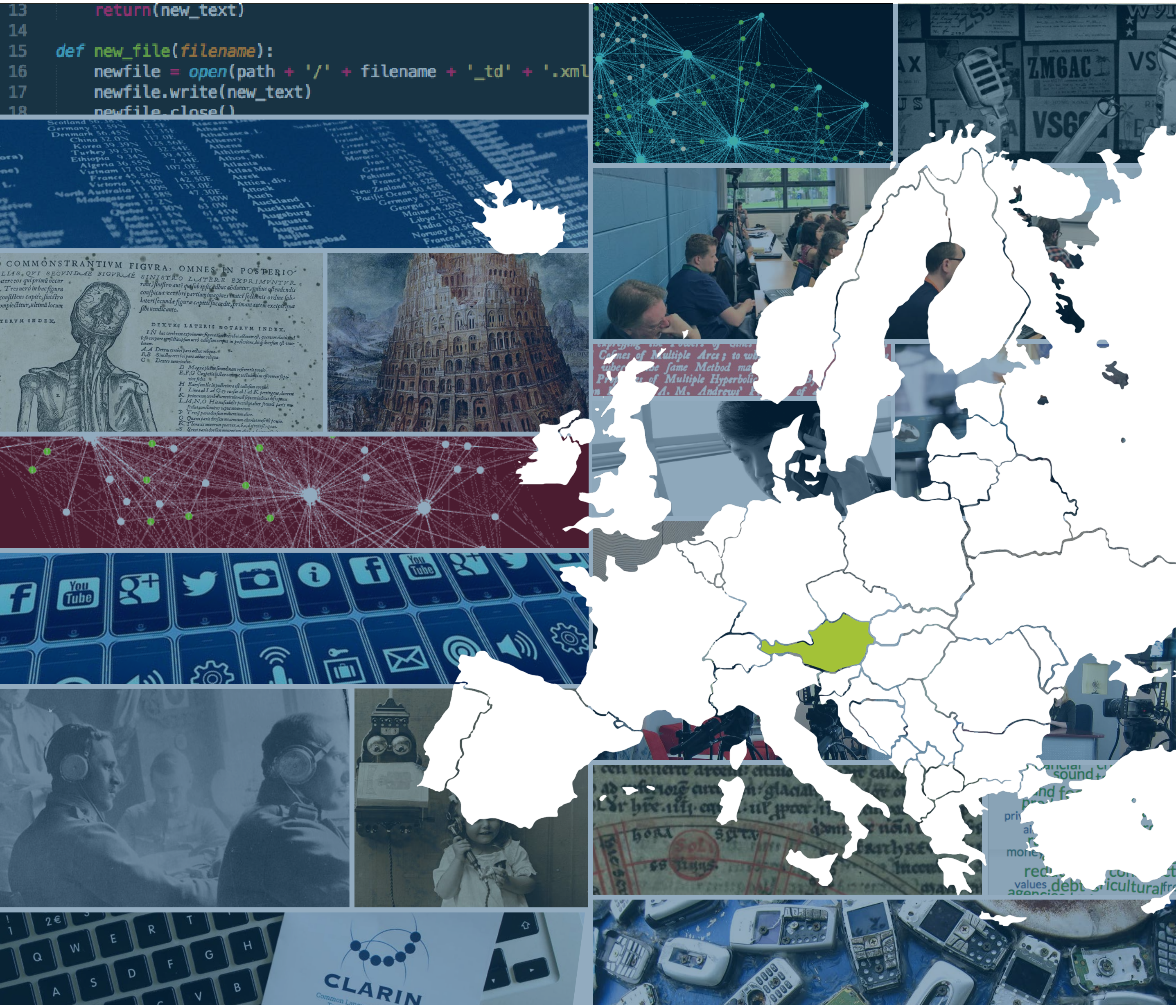
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Austria



Written and edited by Darja Fišer and Jakob Lenardič

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents Austria and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

Graz, Austria | photo by Thomas Quaritsch | Unsplash

Austria

Written by Darja Fišer and Jakob Lenardič

The Austrian CLARIN group is part of CLARIAH-AT,[1] a network of Austrian institutions participating in the two European research infrastructure consortia CLARIN and DARIAH. The network is comprised of eleven research departments at leading Austrian universities and heritage institutions, was a founding member of CLARIN ERIC, and is coordinated by Karlheinz Mörth, director of the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences.

The main Austrian infrastructure for language resources is the Resource Centre for Humanities Related Research in Austria (ARCHE) CLARIN Centre Vienna, which is hosted by the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences. Operating as a B-certified repository that received the Data Seal of Approval in 2014, ARCHE is a language-resource portal that offers researchers the opportunity to deposit and host language resources, language data and tools.

The focus of many activities of the group has been non-standard and historical language varieties. Among the corpora that can be accessed through the centre are the historical Mecmua corpus, which consists of language data from the early modern Ottoman period, the VICAV corpus, which pools language data for the research of Modern Spoken Arabic, and the Austrian Baroque Corpus (described on page 7), which is a specialised historical corpus of German texts from the memento mori genre written in the Baroque period.

[1] http://digital-humanities.at/en/dha/

Part of CLARIAH-AT | Back row: Gerlinde Schneider, Walter Scholger, Claudia Resch, Johannes Spitzbart, Friedrich Neubarth, Martin Hagmüller, Christiane Fritze, Gernot Kubin. Front row: Vesna Lušicky, Tanja Wissik, Helmut Kowar, Ursula Brustmann, Karlheinz Mörth. (Photo by Mehmet Emir, CC-BY 4.0).

Additionally, ARCHE also hosts DictGate, a platform for exchanging lexicographic tools and freely accessible lexical data such as small dictionaries of vernacular Arabic and a bilingual Persian-English dictionary that is still in development.

State-of-the-art tools that have been developed at the Austrian consortium can be freely obtained from the Centre as well. One such tool is the Viennese Lexicographic Editor (described on page 6), an XML-based dictionary writing system which serves as a user-friendly and dynamic tool for compiling and editing digital dictionaries. Another one is the SMC browser, which is a web application that offers users the ability to efficiently explore the Component Metadata Infrastructure by visualising its data.

As a very successful user involvement activity, the Austrian consortium hosts the ACDH Tool Galleries (described on page 9) three times a year. The Tool Galleries are workshops in which established scholars give lectures on the usage of digital tools relevant in digitally grounded humanities research. The consortium thereby promotes knowledge-sharing among its participants and propagates a multifaceted approach to research that is crucial when working in the digital humanities.

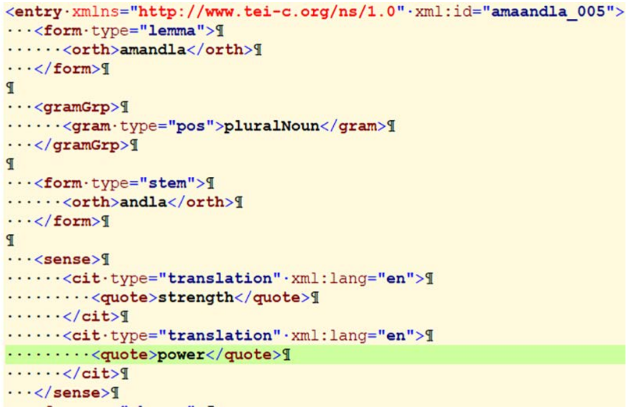
Viennese Lexicographic Editor

Written by Darja Fišer and Jakob Lenardič

The Viennese Lexicographic Editor[2] has been developed by the Austrian Centre for Digital Humanities and is a standalone XML editing system that is designed for collaborative work on lexicographic data. The tool can be freely downloaded and updated versions are uploaded regularly.

A powerful and adaptable tool, the Viennese Lexicographic Editor provides a flexible environment for navigating through and working with complexly-annotated dictionary entries. Researchers can use the tool either to directly access the data in XML (Figure 1) or to edit them by means of an easy-to-use graphical database interface (Figure 2). Furthermore, the tool offers different ways to visualise the data and checks for well-formedness whenever researchers save their entries. Through a special module, the Viennese Lexicographic Editor allows its users to access and integrate external language resources, such as corpora and other dictionaries.

The Viennese Lexicographic Editor has established quite a tradition since it was first used within a glossary-building student project. It has been used as the key piece of software for the compilation of electronic dictionaries, such as the TUNICO Dictionary, within the international Viennese Corpus of Arabic Varieties and the Linguistic dynamics in the Greater Tunis Area projects. In turn, these dictionaries, which are freely available through the DictGate platform, have served as an important resource for comparative dialectological research of language varieties and have been used in language teaching courses, thereby facilitating a cooperative approach to lexicography and lexicology within the digital humanities and social sciences.



```
<entry·xmlns="http://www.tei-c.org/ns/1.0"·xml:id="amaandla_005">
···<form·type="lemma">¶
······<orth>amandla</orth>¶
···</form>¶
¶
···<gramGrp>¶
······<gram·type="pos">pluralNoun</gram>¶
···</gramGrp>¶
¶
···<form·type="stem">¶
······<orth>andla</orth>¶
···</form>¶
¶
···<sense>¶
······<cit·type="translation"·xml:lang="en">¶
·········<quote>strength</quote>¶
······</cit>¶
······<cit·type="translation"·xml:lang="en">¶
·········<quote>power</quote>¶
······</cit>¶
···</sense>¶
```

Figure 1: Editor for TEI dictionaries (screenshot by Karlheinz Mörth).

[2] https://www.oeaw.ac.at/acdh/tools/vle/

The ABaC:us Corpus

Written by Darja Fišer and Jakob Lenardič

The Austrian Baroque Corpus (ABaC:us)[3] is a digital collection of printed texts from the Baroque era, with the bulk of the data from the period between 1650 and 1750. Since 2015, the core corpus has been freely available through the ABaC:us web application, which is provided by the Austrian Centre for Digital Humanities and serves as the first corpus-based application for viewing well-documented language data from the Baroque period. The texts within the collection are predominantly characterised by religious topics, and include morality lectures by Abraham a Sancta Clara, who was one of the most successful preachers in the German-speaking area in the 17th century.

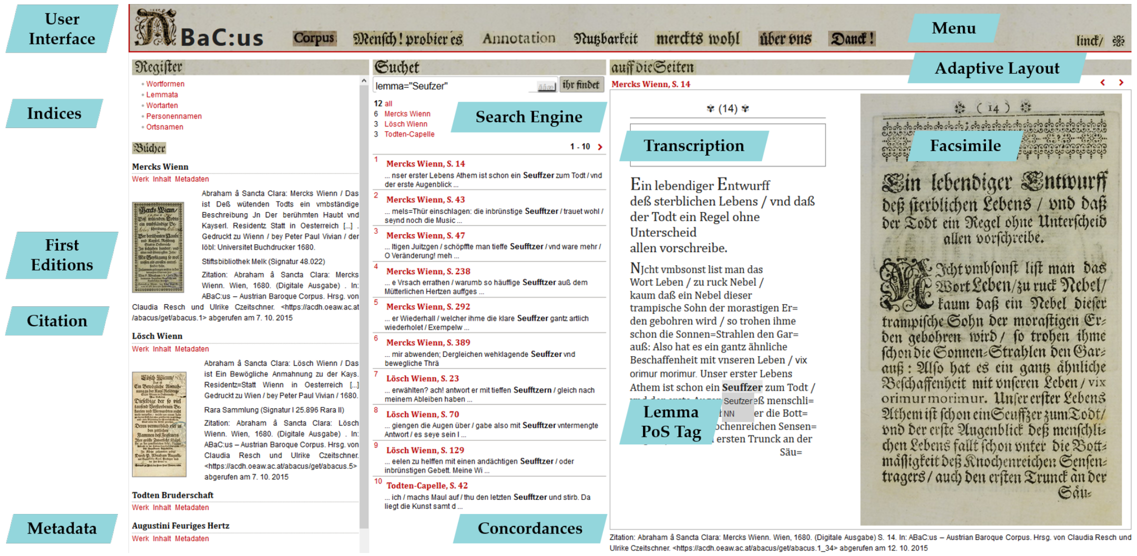


Figure 2: ABaC:us web application (screenshot by Claudia Resch).

The collection consists of 200,000 tokens. Its rather smaller size is due to the fact that computer-generated annotations of Baroque texts, whose stylised orthography and variations in spelling cause a large amount of mismatches by extant taggers, requires copious amounts of additional manual editing. However, the data that are part of the corpus are very richly annotated with mark-up applied to chapters, headings, paragraphs, and named-entities. Apart from PoS-tagging, the corpus is annotated with lemma information, which means that each word form is linked to its base form. Lemma information is a crucial part of the corpus, as it enables researchers to easily identify all occurrences of a word despite the existence of many competing spelling variants and inflected forms.

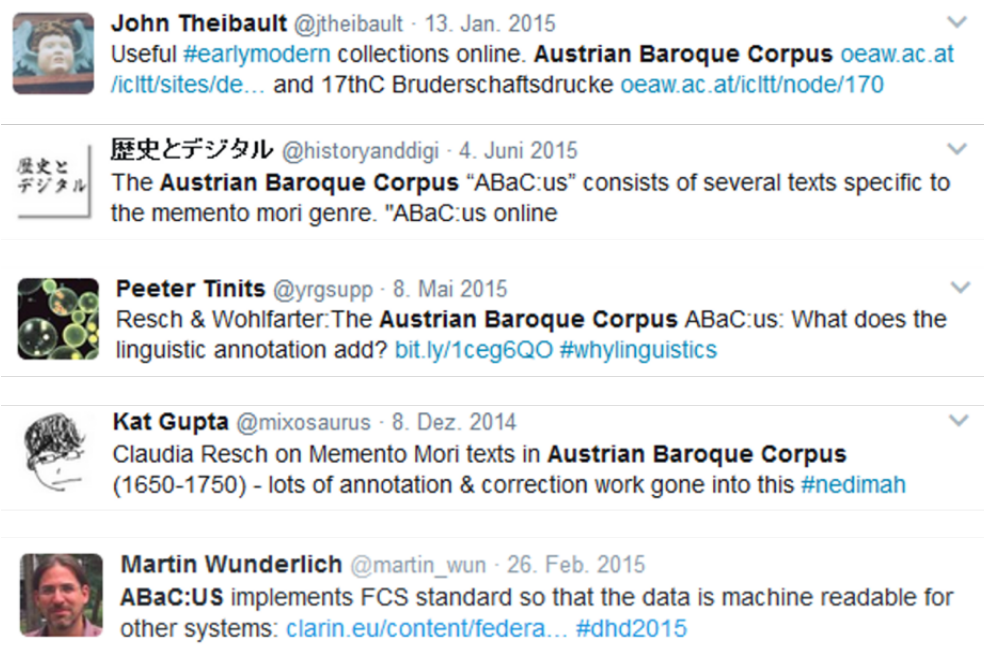
Because of ABaC:us, scholars are for the first time able to explore the vocabulary as well as linguistic structures of the works attributed to Abraham a Sancta Clara in a corpus-based approach. Moreover, the detailed linguistic annotation allows for unbiased research of Sancta Clara, who is portrayed as a particularly linguistically-talented writer in literary history.

[3] https://www.oeaw.ac.at/acdh/tools/abacus/

The ABaC:us project has broken new ground since it allows scholars to combine methods from historical studies—whether they be literary, theological or purely historical—with a corpus-based approach that enables access to richly-annotated data. Indeed, reactions from literary scholars, (computer) linguists, religious scholars and historians have shown how interdisciplinary the interest in ABaC:us is and how many different fields of research across the digital humanities hope to benefit from the free availability of this enriched resource.

The following is a selection of published papers on the ABaC:us corpus:

Resch, C. (2017). "Etwas für alle" – Ausgewählte Texte von und mit Abraham a Sancta Clara digital. In Zeitschrift für digitale Geisteswissenschaften 2017. http://www.zfdg.de/2016_005.
Resch, C. and Czeitschner, U. (2017). Morphosyntaktische Annotation historischer deutscher Texte: Das Austrian Baroque Corpus. In Digitale Methoden der Korpusforschung in Österreich (= Veröffentlichungen zur Linguistik und Kommunikationsforschung, 30, 39–62.
Resch, C., Czeitschner, U., Wohlfarter, E., and Krautgartner, B. (2016). Introducing the Austrian Baroque Corpus: Annotation and Application of a Thematic Research Collection. In Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age.
Resch, C. and Wolfgang, U. Dressler. (2016). Zur Pragmatik der Diminutive in frühen Erbauungstexten Abraham a Sancta Claras. Eine korpusbasierte Studie. In Linguistische Pragmatik in historischen Bezügen, 235–249.



John Theibault @jtheibault · 13. Jan. 2015
Useful #earlymodern collections online. **Austrian Baroque Corpus** oeaw.ac.at /icltt/sites/de… and 17thC Bruderschaftsdrucke oeaw.ac.at/icltt/node/170

歴史とデジタル @historyanddigi · 4. Juni 2015
The **Austrian Baroque Corpus** “ABaC:us” consists of several texts specific to the memento mori genre. "ABaC:us online

Peeter Tinits @yrgsupp · 8. Mai 2015
Resch & Wohlfarter:The **Austrian Baroque Corpus** ABaC:us: What does the linguistic annotation add? bit.ly/1ceg6QO #whylinguistics

Kat Gupta @mixosaurus · 8. Dez. 2014
Claudia Resch on Memento Mori texts in **Austrian Baroque Corpus** (1650-1750) - lots of annotation & correction work gone into this #nedimah

Martin Wunderlich @martin_wun · 26. Feb. 2015
ABaC:US implements FCS standard so that the data is machine readable for other systems: clarin.eu/content/federa… #dhd2015

Figure 3: Tweets on ABaC:us by humanities scholars.

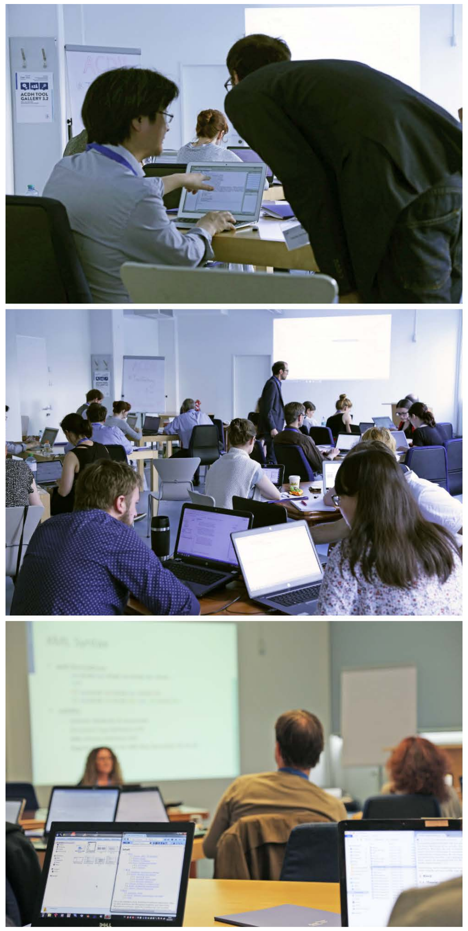
The ACDH Tool Gallery

Written by Darja Fišer and Jakob Lenardič

The training of new and experienced researchers plays a very important role in the digital humanities and social sciences, guaranteeing the far-reaching utilisation of computation tools and digital methods. In Austria, the ACDH Tool Galleries[4] are exemplary cases of such training endeavours. The Tool Galleries are user involvement events organised three times a year by the Austrian Centre for Digital Humanities (ACDH). At these, the developers of tools and experienced professionals share their theoretical and practical knowledge on tools that are designed for digital humanities users. The events take the form of morning lectures and hands-on sessions in the afternoons; the practical work done at these is particularly valuable, especially since it offers the attendees the chance to immediately consult with tool experts if they encounter a problem.

Eleven Tool Galleries have been organised so far, and each has been dedicated to presenting a different subfield in computational linguistics and related tools. For instance, the second Tool Gallery, which took place on 13 October 2015, focused on the importance and use of basic linguistic annotation and was intended for linguists and professionals from all disciplines. Annotation work on the Austrian Baroque Corpus (ABaC:us), which is presented on page 7, and the Austrian Media Corpus was presented, while Marie Hinrichs and Claus Zinn from the University of Tübingen gave a talk on Weblicht, which is a fully functional processing chain that brings together linguistic tools such as tokenisers, part-of-speech taggers, and parsers.

Although the Tool Galleries were originally intended as a service for employees of the Austrian Academy of Sciences, the format was soon extended to a much larger audience that now includes students and academics at all career stages. By organising Tool Galleries three times a year, the Austrian Academy hopes to achieve a regularity and continuity that will serve as a model of researcher training.



ACDH Tool Gallery (photo by Sandra Lehecka, CC-BY 4.0).

[4] https://www.oeaw.ac.at/en/acdh/events/event-series/

Stephan Procházka

Stephan Procházka is a linguist working at the Department of Oriental Studies at the University of Vienna who has collaborated with the Austrian CLARIN consortium in the interdisciplinary TuniCo project, which focused both on researching the linguistic dynamics of the greater Tunis area as well as producing a dictionary of Tunis Arabic and a corpus of transcribed texts. The interview was conducted by e-mail correspondence by Jakob Lenardič and edited by Darja Fišer.

1. Your main research interests lie in Arabic studies. What initially attracted you to the field and what excites you most today?

Initially I was mainly attracted by the fascinating Arab history and the rich material culture such as arts and architecture. For many years now, spoken Arabic varieties have become my main field of interest and research. The so-called dialects are not only interesting for linguists, but also vehicles of a multifarious oral culture ranging from traditional Bedouin poetry to hip-hop songs in the suburbs of Arab megacities.

2. How did your collaboration with the Austrian CLARIN begin and how has it influenced your own work and the way you perceive contemporary Arabic studies?

My collaboration with CLARIN began in 2011 when I was looking for a competent partner to build a kind of platform for Arabic dialectology. I found that in the then ICLTT, which was the fore-runner of the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH-OeAW). From this cooperation many projects such as VICAV emerged.

3. Your most recent project was the interdisciplinary TuniCo[5] project in which you and your team investigated the linguistic dynamics in the greater Tunis area. Could you briefly describe the methodological framework of the project, and highlight its impact for digital humanities and social sciences?

The project was based on the analysis of data gathered during two longer fieldwork campaigns in Tunis. Texts that had been transcribed from the recordings of conversations among young people were the core of our analysis. These texts formed the basis of both lexical and grammatical research, the latter mainly in the field of syntax.

4. Has your analysis of contemporary Arabic spoken by young speakers from different backgrounds revealed any interesting societal trends or culturally specific characteristics?

Yes, we found out that remarkable changes have happened during the last few decades. Young men in particular increasingly show features in their speech which are stigmatised and mostly connected with low-class people from the countryside. They deliberately choose these features to set themselves apart from the mainstream culture. Young educated women, on the other hand, have a preference for using many French words and phrases to show that they are modern and open-minded.

5. Can you describe the two main resources that were developed in this project? What kind of advantages do they bring to your fellow researchers in the field?

We produced a dictionary of Tunis Arabic that comes in a digitally reusable form and lives up to modern IT standards. It contains a very wide range of lexical data, from “historical” vocabulary taken from previous studies to up-to-date youth language taken from our interviews and rap songs, ca. 8,500 entries. It is currently the largest and technically most advanced online dictionary of a spoken Arabic variety worldwide. Together with the other VICAV dictionaries, it is the only such product that is freely available for future research and at the disposition of all researchers. The second resource is a corpus that consists of 24 transcribed texts with ca. 100,000 words. This corpus is linked to the dictionary and thus gives users direct access to the relevant dictionary articles and allows them to understand the Tunisian original. The inclusion of a large number of conversations is one of the innovative traits of our corpus approach, as there are extremely few corpora of spoken varieties of Arabic which include dialogues.

6. How does Arabic, or rather its varieties, fare in the digital context? Are language resources and tools for Arabic readily and widely available? Are there any difficulties specific to automatic processing of Arabic and its varieties? Is there any essential tool or resource that is still missing for Arabic?

Digital language resources for Arabic in general and its spoken varieties in particular, both data as well as tools, are, for several reasons, still under-represented in comparison to many other languages. A major problem is that automatic processing of Arabic, for instance part-of-speech tagging, is more complex because of the characteristic Arabic script that does not indicate short vowels. Arabic varieties are only written in informal settings and lack any standard orthography, which further complicates automatic processing.

[5] https://tunico.acdh.oeaw.ac.at/

7. Have your fellow researchers in the field embraced language technologies in their research frameworks? What is the potential of using language technologies for Arabic studies?

Many scholars in the field are still sceptical about language technologies. However, I see very high potential for my field of research, particularly in the fields of lexicography and syntax. While several treebanks have become available for Modern Standard Arabic, there remains much to be done for the spoken varieties of Arabic.

8. How is the available infrastructure provided by the Austrian consortium or CLARIN ERIC beneficial for your research? Could you highlight a CLARIN tool or resource that has been especially helpful for your work? Would you like to point out anything that could be improved in the future?

The cooperation has been excellent and the available infrastructure very satisfying. The main CLARIN tool for me is the Viennese Lexicographic Editor which from the very beginning facilitated the work in the project. The Vienna CLARIN Centre takes care of the entire resource publication side of our projects, provides both for hosting and preservation of research data, and has always been very helpful in setting up web-interfaces. Especially in our work on the corpus-dictionary interface, the infrastructure of CCV and ACDH-OeAW proved to be very useful.

9. What do you see as the biggest strength of Austrian CLARIN?

They are really interested in cooperation with the humanities and very user orientated. Their interest in further development of their infrastructures in concrete research projects opens up unprecedented synergies, and allows us to move our research in entirely new directions.

10. Where would you like to see CLARIN ERIC 10 years from now?

I think we all would like to have more freely available resources, data and tools that can be used by all researchers, can easily be adapted to the needs of a wide range of fields and projects. While many tools have become available, we still have a long way to go in terms of usability. Finally, I would like to say that I regard CLARIN’s user involvement activities as a very important part of our activities. While much has already been achieved, there are still many in various fields of the humanities who are not aware of recent developments. My vision of CLARIN in 10 years from now is that all young researchers are sufficiently aware of the possibilities the pan-European infrastructure consortia provide, and that the new digital methods are taught in introductory seminar courses on a regular basis, which will eventually lead to wholly new research questions and results.

Vienna, Austria | photo by Andreas N. | Pixabay

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number:
CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Written by
Darja Fišer and Jakob Lenardič

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure

CLARIN 2018