Asil Çetin (asil.cetin@oeaw.ac.at)
Austrian Centre for Digital Humanities at Austrian Academy of Sciences
Vienna, Austria

Report on Elexis Transnational Research Visit Grant at Real Academia Española

(Madrid, Spain, April 7 – April 27, 2019)

Project Title

Visual Data Analysis for Multi-Dimensional Corpus Exploration

About the Project

The aim of this project is to develop an explorative data visualization application to analyze and visualize regional language varieties and statistical differences in lexical uses of language. This project runs as a collaborative design study as part of my Master's thesis at the Computer Science Faculty of the University of Vienna and Austrian Centre for Digital Humanities. The software architecture of the application follows a decoupled service pattern separating data collection / curation, corpus access and frontend of the web application. Hence, a software architecture like that would offer the opportunity to reuse the application for different language sources and different corpus engines.

About the Data Sources

During the design, development and evaluation phases of this project the following two main data sources consisting of large corpora in two different languages are being used:

- **Austrian Media Corpus:** AMC was created as part of a cooperation between the Austrian Academy of Sciences and the Austrian Press Agency. It covers the entire Austrian media landscape of the past two decades, containing 40 million texts, constituting more than 10 billion tokens. AMC ranks among the largest collections of its kind as a contemporary German language corpora.

- **Real Academia Española:** The "Advanced Search Interface" of DLE 23, CORPES, CREA and CORDE are some of the query mechanisms of the Real Academia Española (RAE), which provide accurate linguistic data about varieties of Spanish language. RAE, with its affiliations in 22 hispanophone nations, offers the most extensive knowledge and data regarding the Spanish language.

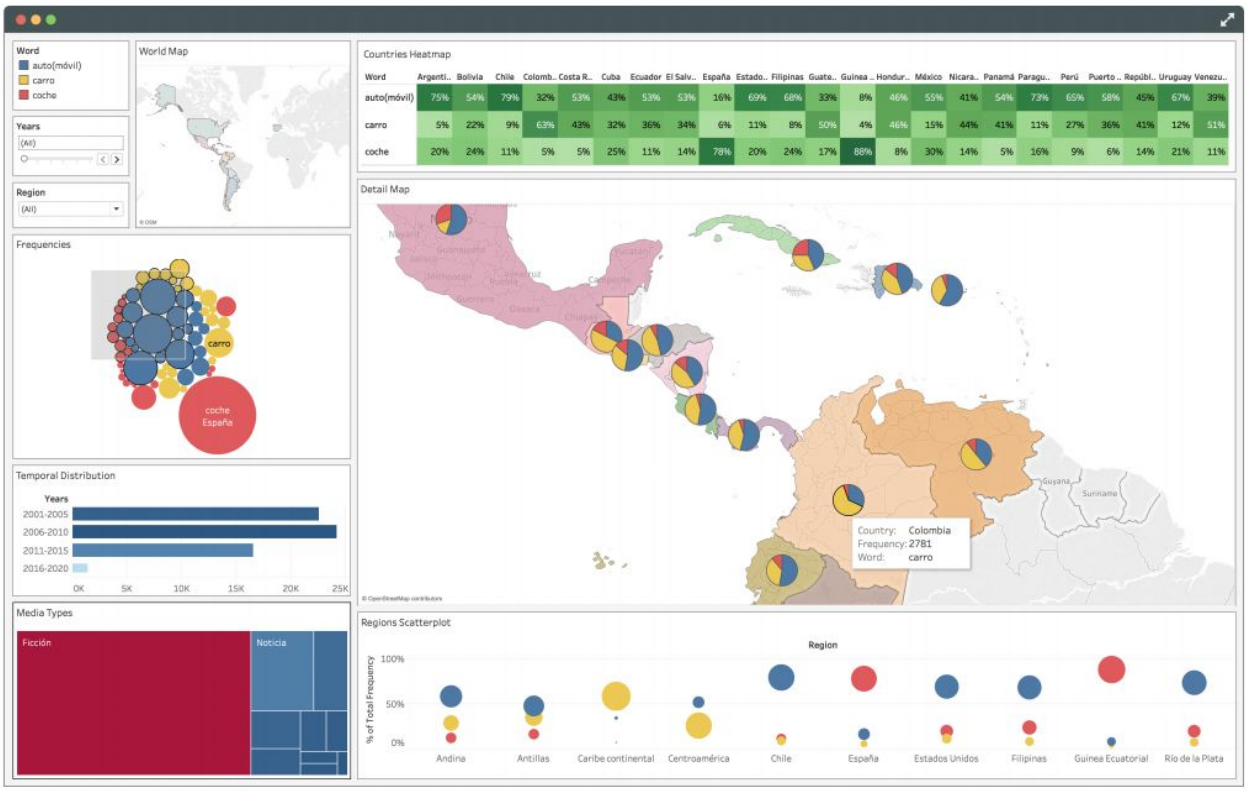

Figure 1: High-fidelity prototype using sample data from Corpus del Español del Siglo XXI of Real Academia Española. The application is intended be used with other languages / corpora as well.

Summary of the Research Visit

My research visit at the Real Academia Española through the ELEXIS travel grant started by the second week of April 2019 and lasted three weeks. In this timespan the main aim of my project was to develop and design an explorative data visualization application to analyze and visualize corpus linguistic data.

Since my project runs as a collaborative design study and focuses on the target group of researchers of the fields of linguistics and humanities, it's crucial to be in contact with domain experts of these fields.

During my research visit at the Real Academia Española it was possible to fulfill both of these expectations for my project: accessing some of the largest and well annotated corpora available today and exchanging knowledge and feedback with domain experts during various stages of my research.

I've worked during the weekdays at the offices of the Real Academia Española in Madrid's El Viso quarter, which are allocated for departments of computational and corpus linguistics, lexicography and software development. My main supervisor was Mr. Jordi Porta-Zamorano, PhD and it was great opportunity to work with him on a daily basis and profit from his decades long experience in this field and at the Real Academia Española.

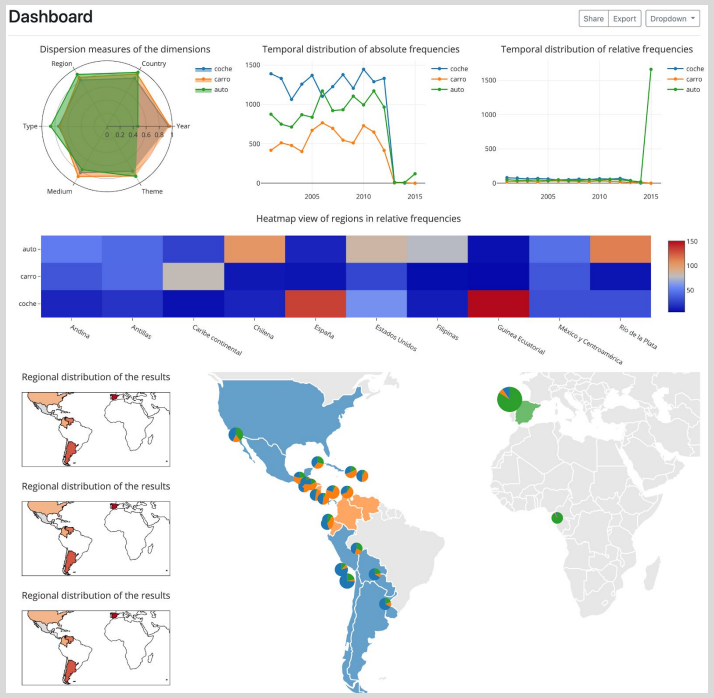

Figure 2: The tool being implemented as an interactive web-app with real-time connection to RAE's CORPES API.

Some of the other colleagues who I've worked with and exchanged knowledge on daily basis were José Luis Sancho, MSc, Rafael Ureña, MSc and Juan Romeu, PhD. During and after each iteration of design and implementation it was possible to consult with these colleagues to review the strengths and weaknesses of various choices.

At the end of my research visit I had the opportunity to make a final presentation about my project with the attendance of ~20 colleagues, which made it possible to gather feedback from various domain experts.

Results and Future Work

The results of this highly productive research visit in terms of software development, design study and knowledge exchange can be listed as follows:

- Accessing large and well annotated linguistic data
- Developing the software tool using the real-time connection the available resources
- Getting valuable input from domain experts on a daily basis during design / development

- Presentation and live-demonstration session with feedback from colleagues at RAE
- Setting up a basis for future collaboration regarding this project with RAE

Currently the results in terms of software tool are still in progress and the public git-repository of the web-application can be found at: https://github.com/asilcetin/corpsum. In the upcoming months the application will be under further development and the results of both development and the design study will be documented in a paper. This paper and detailed documentation about the tool will be made available / linked at the main readme page of the git repository.