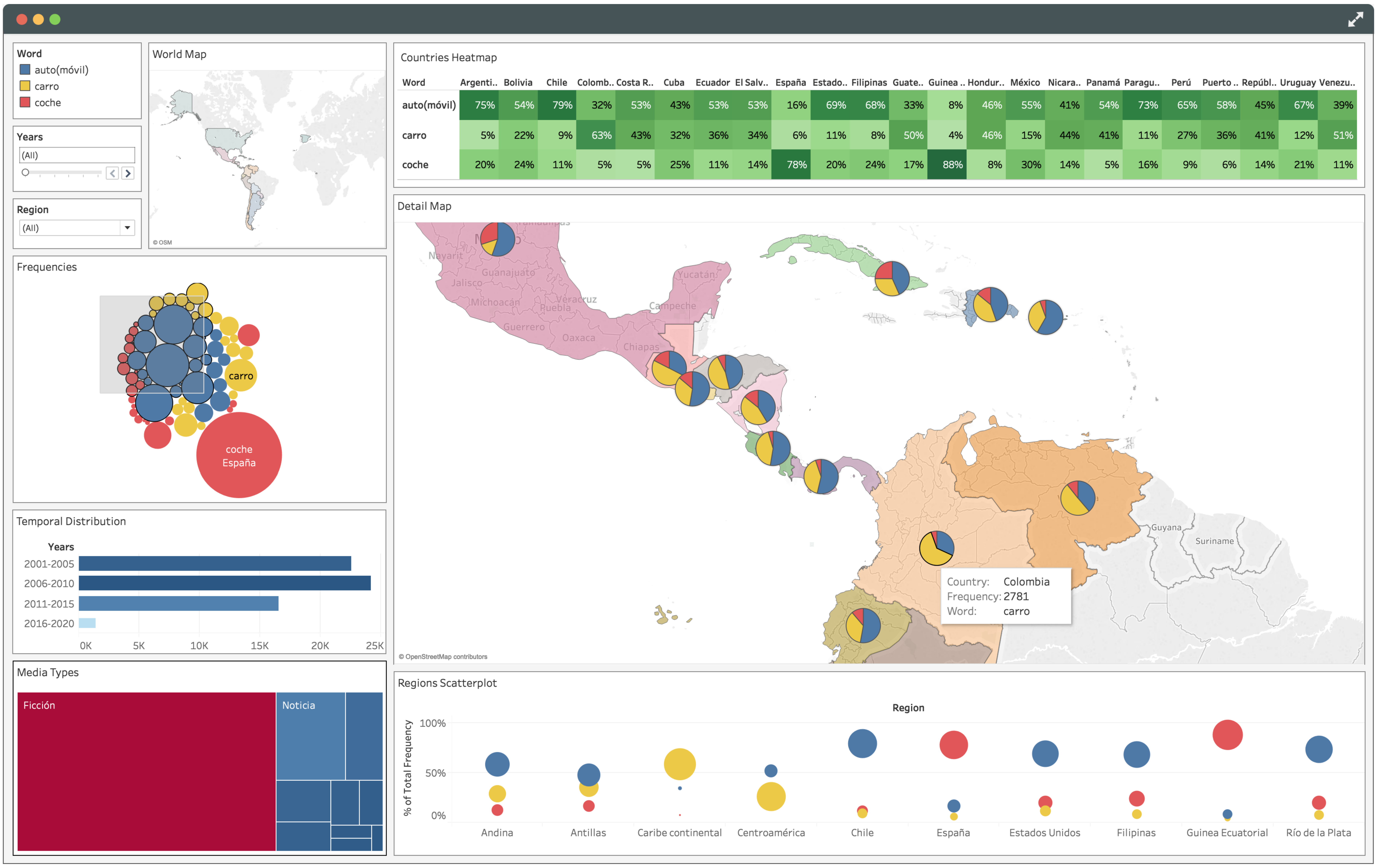
Visual Data Analysis for Multi-Dimensional Exploration of Language Varieties

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



Countries Heatmap

Word	Argenti..	Bolivia	Chile	Colomb..	Costa R..	Cuba	Ecuador	El Salv..	España	Estado..	Filipinas	Guate..	Guinea ..	Hondur..	México	Nicara..	Panamá	Paragu..	Perú	Puerto ..	Repúbl..	Uruguay	Venezu..
auto(móvil)	75%	54%	79%	32%	53%	43%	53%	53%	16%	69%	68%	33%	8%	46%	55%	41%	54%	73%	65%	58%	45%	67%	39%
carro	5%	22%	9%	63%	43%	32%	36%	34%	6%	11%	8%	50%	4%	46%	15%	44%	41%	11%	27%	36%	41%	12%	51%
coche	20%	24%	11%	5%	5%	25%	11%	14%	78%	20%	24%	17%	88%	8%	30%	14%	5%	16%	9%	6%	14%	21%	11%

Country: Colombia
Frequency: 2781
Word: carro

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

About the Project

The aim of this project is to develop an explorative data visualization application to analyze and visualize the regional language varieties and statistical differences in lexical uses of languages. The architecture of the application will follow a decoupled service pattern separating data collection / curation, corpus access and frontend of the web application. Such a software architecture would make it possible to reuse the application for different language sources and with different corpus engines.

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. Special thanks to the ELEXIS infrastructure and RAE for supporting this project in terms of expertise, resources and data sources.

universität wien

austrian centre for digital humanities

elexis european lexicographic infrastructure

About the Data

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

amc austrian media corpus

AMC, created as part of a cooperation between the Austrian Academy of Sciences and the Austrian Press Agency, 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.

User-Centric Research Approach

The focus group of this project consists of mostly young researchers and students of the fields of linguistics, philology and humanities. Following a user-centric research approach allows the project to have a solid base of domain understanding, abstraction of the most meaningful tasks and application of effective visual encodings and algorithms. In order to implement this nested model, many interviews with multiple people are conducted as an iterative practice of evaluation.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.