universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Multi-Faceted Visual Data Analysis for Corpus Research“

verfasst von / submitted by

Asil Çetin

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

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 921
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Informatik
Betreut von / Supervisor:	Univ. Prof. Torsten Möller, PhD

Contents

Abstract

In this thesis we present a design study project for the development of an exploratory corpus analysis and visualization tool, which allows linguistics researchers to access large text corpora and gain insights to their research questions using an interactive multi-query dashboard. Our methodology follows an iterative process and a tight collaboration with domain experts to prioritise the most common tasks of the users and decrease the required technical skills for using the software. An inspection of existing tools in this field and the detailed analysis of tasks and users in this domain are documented as the basis of our work. In terms of evaluation, our tool receives very positive feedback from the users, and some case studies of example usages of our software are presented in this paper.

Abstract

In dieser Arbeit stellen wir ein Designstudienprojekt für die Entwicklung eines explorativen Korpusanalyse- und Visualisierungstools vor, das Sprachwissenschaftlern ermöglicht, auf große Textkorpora zuzugreifen und mithilfe eines interaktiven Multi-Abfrage-Dashboards Erkenntnisse zu ihren Forschungsfragen zu gewinnen. Unsere Methodik folgt einem iterativen Prozess in enger Zusammenarbeit mit Fachexperten, um die häufigsten Fragestellungen der Anwender zu priorisieren und die Benützung der Software auch für Anwender mit geringen technischen Kenntnissen zu ermöglichen. Eine Recherche zu den bestehenden Tools in diesem Feld und die detaillierte Analyse der Aufgaben und Eigenschaften der Anwender bildet die Grundlage unserer Arbeit. Bei Evaluierungen erhält unser Tool ein sehr positives Feedback von den Anwendern und einige Fallstudien zur Nutzung unserer Software werden in diesem Paper vorgestellt.

1 Motivation

The term "corpus" (plural: corpora) is used to describe any collection of written or spoken text. However, in reference to modern linguistics it implies a finite-size collection of texts in a machine-readable form, which has a certain representativeness for a certain language [29]. Today, corpus-based research is a key element of almost all language studies and the corpus is considered the default source for linguistics researchers [40].

The variety and the scale of text corpora can be diverse; from historical literature collections to parliament speech corpora, from a collection of tweets on a given hashtag to user reviews on an internet platform. Thus, corpus researchers may have significantly different expectations and varying research questions based on the corpora. This variety makes it an exciting challenge for visualization research projects to create suitable solutions for the given domain and the characteristics of the data source.

In this project the main domain of interest will be contemporary media corpora and their corpus-linguistic analysis. This type of language corpora generally consist of journalistic prose (newspaper and magazines articles, press releases, interviews and news stories transcribed from television etc.) and offer a unique contemporary language resource, which can be used in various lexicography and linguistics research projects [32].

As focus datasets we have two large contemporary media corpora: Austrian Media Corpus (one of the largest contemporary German language corpora) and CORPES (a reference corpus for 21st century Spanish language created by the Real Academia Española), and we collaborate with domain experts who work with these sources, which are described in detail in sections "Data" and "Users".

The researchers who work with such corpora need to provide empirical evidence in the form of data and offer replicability of their findings with making the corpora available for further exploration to other researchers [15]. However due to the technical limitations, it's usually not a straightforward task for linguistics researchers to gain access to corpora, install various software tools, process complex queries using the dataset, analyze the results to evaluate their hypothesis and visualize the findings in various dimensions for further exploration. During multiple sessions of user interviews we notice the common problem of high complexity and low usability of existing corpus analysis tools experienced by the users. Thus we focus on defining and solving the most important tasks of the users, which are presented in detail in section "Tasks".

Based on the guidelines and the methodology suggested Sedlmair et al. [37], this project is conducted as a design study with a heavy focus on the iterative collaboration with the users. This research approach, explained in detail in section "Methodology", allows the project to have a solid base of domain un-

derstanding, to abstract the most meaningful tasks and to apply effective visual encodings and algorithms.

The main contributions of this project are: 1) the detailed analysis of tasks and users in this domain, 2) the functional exploratory data visualization application, 3) evaluation of the iterative design and development process, and 3) the inspection of existing tools in this field. These contributions and the outcomes of this project are evaluated and documented in detail in sections "Evaluation and Case Studies" and "Lessons Learned".

2 Related Work

2.1 History of Corpus Analysis Tools

The roots of modern corpus linguistics can be tracked back to the long history of linguistics. However a major breakthrough happened in the 1970s and 1980s when it became possible to store texts in a machine readable form, and transport and analyze them electronically. Improvements in computer technology contributed to this development as more powerful and cheaper computational resources were available to linguistics researchers [22].

Figure 1: A historical example of pre-electronic hand-crafted corpora: Corpus of common Spanish words at Real Academia Española, Madrid.
CC-BY 4.0, Asil Çetin.

As the possibilities computers offered in this field have been enormous, corpora are nowadays divided into two types: pre-electronic and electronic corpora. Pre-electronic corpora (Figure 1) were created with intensive manual work, usually with pen and paper, and the analysis of these text collections was highly time consuming. An example for pre-electronic corpus analysis would be the concordance of the King James Version of the Bible, which was manually created by Alexander Cruden in the 18th century [23]. On the other hand, electronic corpora became the standard sources for corpus linguistic research in the current modern era, with "The Brown University Standard Corpus of Present-Day American English" (Brown Corpus) being one of the first significant examples of its kind [23].

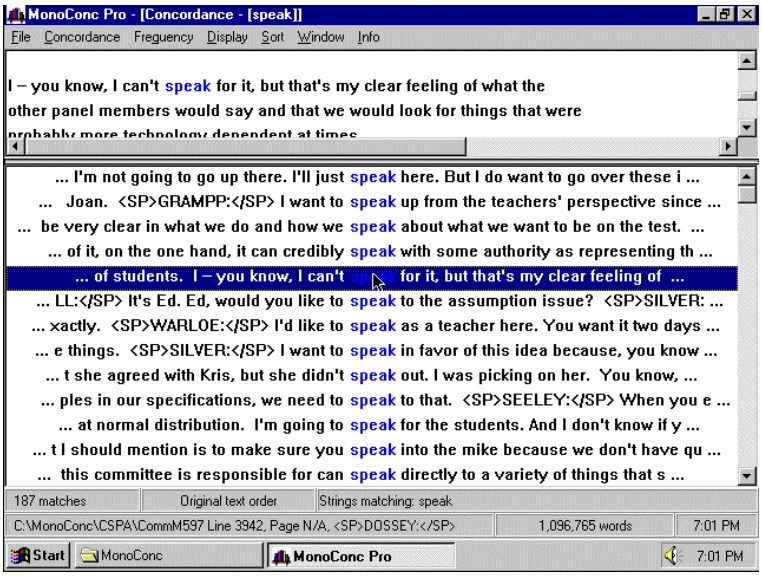

Figure 2: MonoConc [13][6], a desktop software developed for corpus analysis in early 2000s.

Increasing availability of electronic corpora opened the doors for a wide variety of corpus linguistic research questions to be answered by computational methods. Using electronic corpora as machine readable textual data sources, different corpus analysis tools have emerged. Looking back at the history of these tools, McEnery and Hardie divide these software into four generations [30]:

- 1st generation (from 1970s to 1980s): These tools were mostly running on a mainframe computer at a single site and output was typically a keyword-in-context (KWIC) concordance. Examples are CLOC (1978)[33], LOB (1978)[21],
- 2nd generation (from 1980s to 1990s): These tools took the advantage of the rise of the PCs and were available to be distributed to personal computers and different platforms. They offered still mostly KWIC concordances and some basic descriptive statistics. Examples are Kaye (1990)[24], Longman Mini-Concordancer (1989)[16] and Micro-OCP (1988)[19].
- 3rd generation (from 1990s to 2000s): These tools added many popular functionalities on top of concordances, which are still used today, such as frequency lists, keyword analysis and collocations. Support for different input sources, such as XML, and character support for different languages were introduced. Examples are WordSmith (1996)[36], MonoConc (2000)[13] (Figure 2) and AntConc(2005)[10].

- 4th generation (from 2000s until today): Most of these tools moved into the web browser to separate user interaction (client machine) and resource-intensive searching and processing (server). This allowed a more widespread usage of corpora and also a work-around for legal restrictions of distributing corpus data. Examples are SketchEngine (2004)[26], BNCweb (2008)[1], CQPweb (2008)[3].

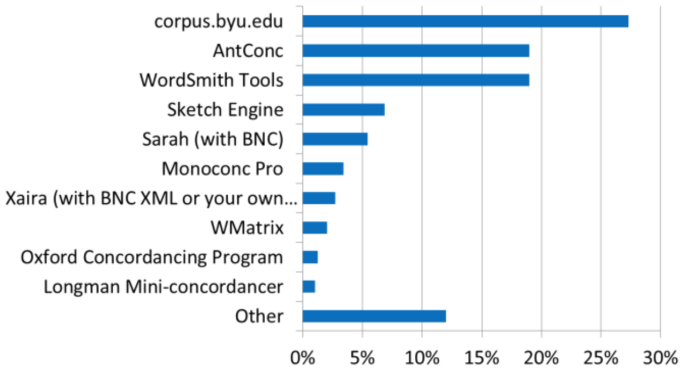

Figure 3: Answers to the question "Which computer programs do you use for analysing corpora?" surveyed by Tribble in 2012 [43][11]

Nowadays most of the popular corpus analysis tools can be considered belonging to the 4th generation, since the structure and functionalities of the most commonly used tools of today are consistent with the above given definition of this generation. However some tools are still developed and offered as desktop clients rather than web applications, as they build on their legacy as 3rd generation corpus analysis software. A user survey by Tribble [43] to the question of "Which computer programs do you use for analysing corpora?" from 891 responses reveals the tools or platforms, which are widely used by the corpus linguistics community (Figure 3).

In order to understand the advantages and drawbacks of these tools in terms of functionality, usability, availability and visualization capabilities, we will look deeper into some of the most relevant tools, which are actively used today in this domain in the following section.

2.2 Relevant Tools and Functionalities

In this section of related work we collected some of the most commonly used solutions for the analysis of text corpora. Our selection is based mainly on the survey by Tribble [43], however we replace some of the outdated or not available tools with newer ones, which were mentioned by our domain experts during the

interviews. The scope of these tools and their specific goals vary wildly and thus it's important to define the key take-away points from this survey. Therefore we inspect these related work in terms of the following criteria:

- Design Study: whether or not the software was created using an iterative design process with domain experts.
- User Evaluation: whether or not the developed software solution was evaluated by either expert reviews or lab studies.
- Multi-Faceted Dashboard: whether or not the tool offers a dashboard with different views to explore the corpus data in multiple dimensions (temporal, regional, media types and other relevant metadata). The importance of this criterion originates from the derived tasks listed in the "Tasks" section, which often require the analysis of multiple metadata dimensions simultaneously.
- Web-Based Application: whether or not the analysis system can be deployed as a web-based application, which could widen the cross-platform usability of the solution and possibly decouple server and client.
- Large Corpora: whether or not the solution is designed to handle and allow exploring large collections of annotated text documents for corpus linguistic research. As an arbitrary amount to describe "large corpora", especially for written media collections, we set the limit to a minimum of 1 million text documents, each with a news article length on average.
- Reusable: whether or not the source code of the final product is available for reuse or further development.

2.2.1 AntConc

AntConc is a freeware corpus analysis toolkit for concordancing and text analysis, and has been developed by Dr. Laurence Anthony, a Professor in the Faculty of Science and Engineering at Waseda University in Tokyo, Japan [12]. The program runs as a desktop software on operating systems such as Microsoft Windows, MacOS, and Linux, and was initially developed in Perl but afterwards refactored in Python and the Qt graphical user interface package [11].

- Design Study: Anthony describes AntConc's initial aim to be an easy-to-use tool for learners in a classroom context, who are not necessarily experienced researchers of corpus linguistics and often run personal computers with Windows/Mac/Linux in schools or colleges with a limited budget [10]. He mentions that the feedback received from the community after the success of the tool has influenced further development and future

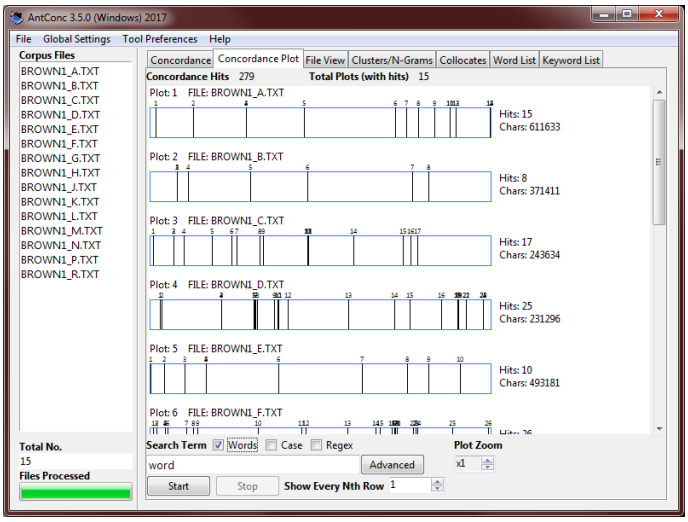

Figure 4: AntConc [12] displaying the concordance plot view on a separate tab on it's dashboard.

releases, however there is no methodological design study at the core of the design and development of the system, to the best of our knowledge.

- User Evaluation: Even though AntConc's developers mention that taking the needs of the potential users from corpus linguistics community very seriously [11], there is no official result of user evaluation provided, to the best of our knowledge.
- Multi-Faceted Dashboard: AntConc offers a rather traditional tabs-layout rather than a multi-faceted analysis dashboard for exploring and inspecting the results.
- Web-Based Application: AntConc is a desktop client, which runs on a personal computer and does not offer a decoupled web architecture.
- Large Corpora: Since AntConc is run on a personal computer and designed for individual learners, it has limitations when it comes to handling very large amount of text corpora and serving as a corpus engine.
- Reusable: AntConc is offered as a non-commercial (freeware) software product in binary form for personal use, however an open-source repository for the source code for reuse is not available.

2.2.2 corpus.byu.edu / english-corpora.org

Mark Davies, Professor of Linguistics at Brigham Young University in Provo, has developed a corpus architecture and web interface for the different text corpora, which he designed, collected, edited, and annotated mostly for the English language [17]. The web service was available initially under the domain of corpus.byu.edu - also colloquially referred as Mark Davies Corpora - and is currently redirected to a new domain of english-corpora.org. Tribble's survey (Figure 3) lists this web service as the most commonly used corpus analysis service by the researchers and in our user interviews we have also been told by some users that they use this service for their corpus linguistic research.

The corpus architecture of this tool uses Microsoft SQL Server and a relational database approach for storing the data related to corpora, and it's claimed to offer very good performance and scalability as it's used by more than 130,000 unique people each month [17]. Even though this web service offers a wide variety and large size of corpora, and allows users to make queries on these corpora, it's not possible to categorize this service as a generic tool for corpus linguistics, because it's not available to be downloaded and run as a tool by any means, and the users cannot upload and query their own corpora. Thus it would be more appropriate to categorize this platform as a web-based service, since it's not designed to be reused by researchers for their custom needs.

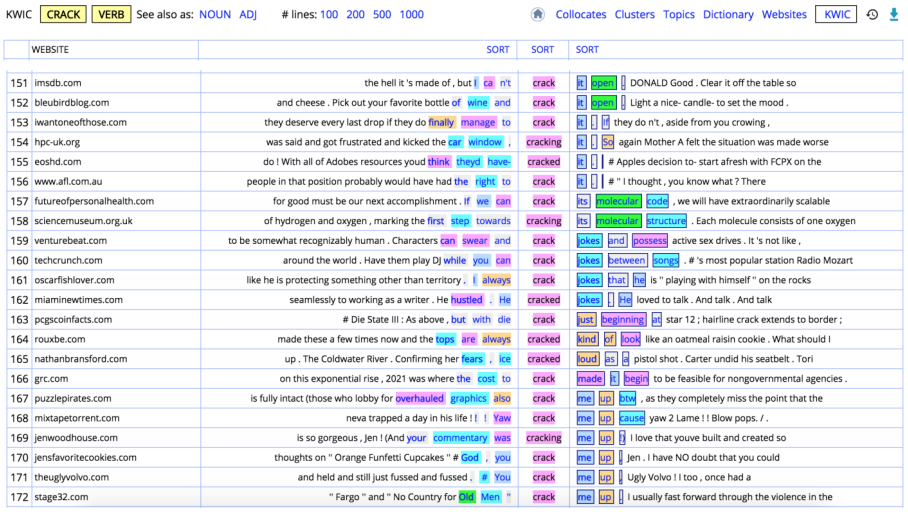

Figure 5: Keyword-in-context (KWIC) view on english-corpora.org [17]

- Design Study: Mark Davies, the creator of this platform, describes the development of this service as an iterative process, which has built over the previous improvements over time [18]. However there is no mention

of a design study, to the best of our knowledge.

- User Evaluation: There is no publicly available result of a user evaluation on this service, to the best of our knowledge.
- Multi-Faceted Dashboard: This service offers a rather traditional tabs-layout rather than a multi-faceted analysis dashboard for exploring and inspecting the results.
- Web-Based Application: This service is offered as a web platform.
- Large Corpora: This service has the capacity and scalability to handle large corpora.
- Reusable: This service is not downloadable or reusable.

2.2.3 CQPweb

No	Text			
1	AmE06_A01	years old with structures that could rise up to six stories and	house	offices , shops , and restaurants . When the council first voted
2	AmE06_A01	public unease with the Iraq war is jeopardizing Republican majorities in the	House	and possibly the Senate . Baker 's public appearances , whether by
3	AmE06_A01	after the elections , and one that did n't require the White	House	to give up itscampaign theme : while things might not be going
4	AmE06_A01	a campaign gimmick has turned into foreign-policy reality : the Bush White	House	has lost confidence in its ability to win the Iraq war .
5	AmE06_A01	the Iraq war . Acampaign to manage public perception of the White	House	's handling of the Iraq quagmire in advance of the November polls
6	AmE06_A03	Tuesday , Bush said earlier in the day during a joint White	House	press conference with Afghan President Hamid Karzai that he had ordered all
7	AmE06_A03	anticipated that not all of document would be declassified . On the	House	floor , Democratic Leader Nancy Pelosi of San Francisco called unsuccessfully for
8	AmE06_A03	Leader Nancy Pelosi of San Francisco called unsuccessfully for a rare secret	House	session to make the entire report available to all 435 members so
9	AmE06_A05	Senate will back the proposal , which is moving swiftly through the	House	. Sen. Alex Villalobos , the Republican majority leader from Miami ,
10	AmE06_A05	know I had . " Rep. Ralph Arza , chairman of the	House	PreK-12 Committee , said he thinks the class size changes working through
11	AmE06_A06	Most people think that we would have a hard time holding the	House	if we voted today , " says Charles Black , a senior
12	AmE06_A06	point in 1994 , weeks before they swept to control of the	House	and Senate during President Clinton 's first term , and higher than
13	AmE06_A06	has insisted that GOP candidates would survive a difficult political landscape because	House	elections traditionally turn on local issues , not national ones . However
14	AmE06_A06	" GOP advantages dwindle Democrats need 15 new seats in the	House	and six in the Senate to claim a majority . Non-partisan analysts
15	AmE06_A06	and Charles Cook estimate Democrats will gain at least 18 to 20	House	seats and possibly more . Taking over the Senate is a tougher
16	AmE06_A06	candidates are experiencing a late surge . The RNC and the GOP	House	and Senate campaign committees had a combined $77.6 million in cash on
17	AmE06_A06	buffeted by the scandal involving Foley 's lewd computer messages to former	House	pages . Democrats in several states -- including Ohio , Colorado ,
18	AmE06_A06	and vote , " she says . Foley was the third GOP	House	member to resign this year amid scandal . A fourth , Ohio
19	AmE06_A06	' Senate campaign chief , said in an interview . Former GOP	House	majority leader Dick Armey says small-government conservatives do n't believe the United
20	AmE06_A06	thousands of volunteers with Democratic-leaning voters in three Senate races and 50	House	contests . Eli Pariser , director of MoveOn 's political arm ,
21	AmE06_A08	is the exchange between Jim Webb and President Bush at a White	House	Christmas party . Mr. Webb did not want to pose with the
22	AmE06_A10	R-Ariz. , and a Vietnam-era torture survivor who had bucked the White	House	along with Senate Armed Services Committee Chairman John Warner , R-Va. ,
23	AmE06_A10	detainees . The agreement resolved the major sticking points between the White	House	and the senators . The senators persuaded the White House to leave
24	AmE06_A10	the White House and the senators . The senators persuaded the White	House	to leave intact U.S. obligations under the Geneva Conventions ' Common Article
25	AmE06_A10	humiliating and degrading treatment " of wartime detainees . But the White	House	and CIA gained the clarifications they were seeking in U.S. law ,

Figure 6: Keyword-in-context (KWIC) view on CQPweb [3]

CQPweb is developed as a web-based corpus analysis system at the Lancaster University and aims to offer a user-friendly interface to the corpus workbench (CWB) system, which was developed at the same university [2]. The system can be reached online at https://cqpweb.lancs.ac.uk and is available to be downloaded and installed as open-source software on any server or computer.

The popular web service of BNCweb and its user interface have been the influence for the development and release of CQPweb. Unlike BNCweb, which was a web platform only to be used by logged-in users with pre-defined corpora, CQPweb allows users to upload and query their own corpora, download the

software and use the system on their own server setup for their custom needs. It's a web-based application, which offers features such as concordances, query sorting, collocations, distributions, subcorpora creation and a simple query language. Based on these key facts, CQPweb can be categorized as a 4th generation corpus analysis system according to the criteria of McEnery and Hardie [30].

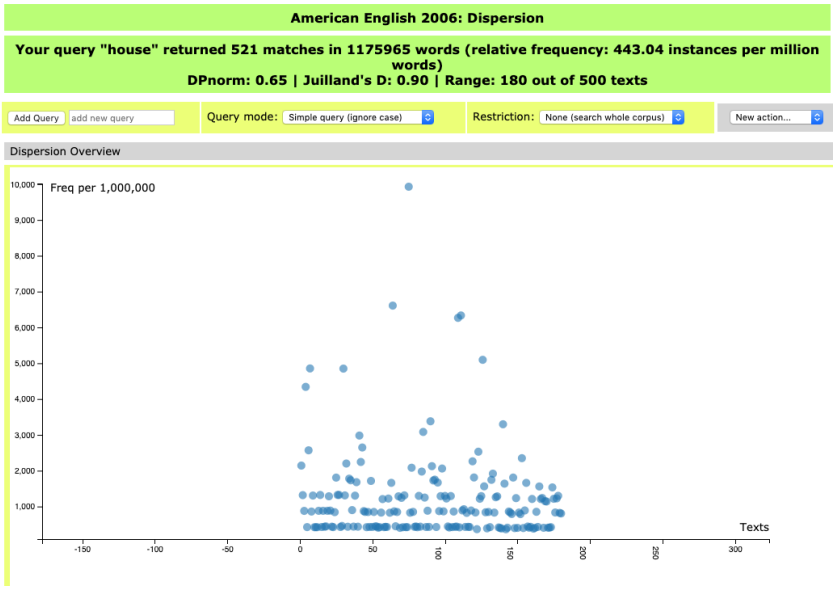

Figure 7: A basic dispersion scatterplot view on CQPweb displaying relative frequency per document on a given query [3]

- Design Study: CQPweb is built on BNCweb, which was created at the the English Department of the University of Zurich to serve the needs of some researchers [1]. The evolution of the tool led it to become a commonly used platform, however there is no design study at the hearth of the development of this tool, to the best of our knowledge.
- User Evaluation: An evaluation of the features offered by the earlier version of this tool, called BNCweb, is publicly available [1].
- Multi-Faceted Dashboard: This service offers a layout where every view is accessed on a separate page rather than a multi-faceted analysis dashboard for exploring and inspecting the results.
- Web-Based Application: This service is offered as a web application.
- Large Corpora: This service has the capacity and scalability to handle large corpora.
- Reusable: This service is downloadable and reusable.

2.2.4 LancsBox

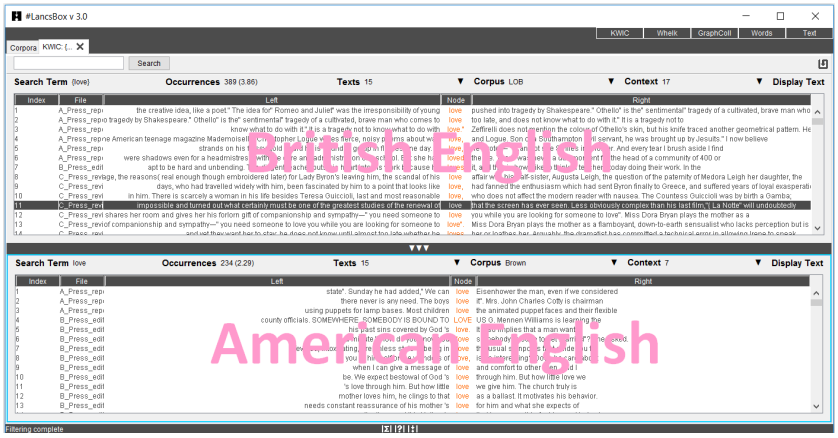

Figure 8: Multi-corpora KWIC comparison view on LancsBox [5]

LancsBox is developed at the Lancaster University by the project lead Vaclav Brezina and with the aim of a new-generation software package for the analysis of language data and corpora [5]. The system works as a desktop client on different operating systems such as Microsoft Windows, MacOS and Linux, and is available for free for non-commercial use. The system allows users to work with their own corpora and aimed to be useful for linguists, language educators, historians and people interested in language research as it can automatically annotate data for part-of-speech, offers multi-language support and some visualization functionalities [5].

- Design Study: There is no design study at the basis of the development of this tool, to the best of our knowledge.
- User Evaluation: A user evaluation of LancsBox is not publicly available, to the best of our knowledge.
- Multi-faceted Dashboard: This tool offers some layout elements to be shown together on some views however generally it applies a traditional tabs-layout rather than a multi-faceted analysis dashboard.
- Web-Based Application: This tool is offered as a desktop client and not available as a web-based application.
- Large Corpora: This tool has the capacity and scalability to handle large corpora.
- Reusable: This tool is downloadable for non-commercial use, however it's not reusable for further development as the source code is not available.

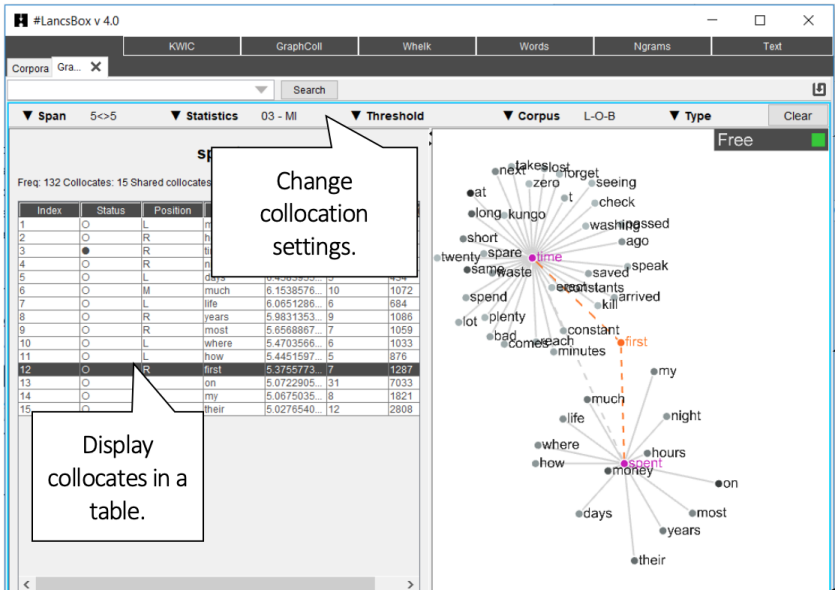

Figure 9: Collocations results as a table and a network graph on LancsBox [5]

2.2.5 MonoConc

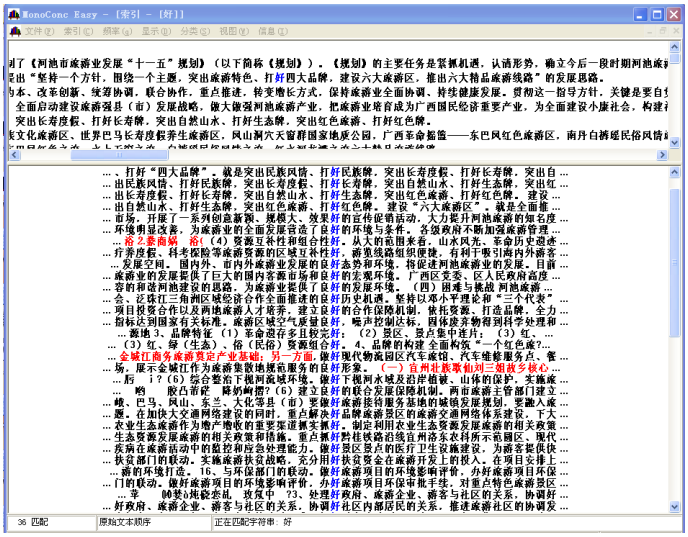

Figure 10: Multi-language support for corpus analysis displayed on a KWIC view on MonoConc [7]

MonoConc is developed by Michael Barlow and offered as a concordance /

text analysis desktop software for Microsoft Windows operating systems. The tool defines its target group as universities and schools for teaching and research, and claims to provide expert users with powerful and configurable functionalities as well [7].

In comparison to the other tools we have surveyed, MonoConc is a somewhat outdated tool, since it's latest stable release is more than a decade old, and based on it's architecture and features it can be categorized as a 3rd generation corpus analysis tool according to the criteria of McEnery and Hardie (2012) [30]. Even though the tool is relatively old, there are still some active users who are using it for their research, according the survey of Tribble [43].

- Design Study: For MonoConc, there is no design study at the basis of the development, to the best of our knowledge.
- User Evaluation: A user evaluation of MonoConc is not publicly available, to the best of our knowledge.
- Multi-Faceted Dashboard: MonoConc applies a layout based on separate pages rather than a multi-faceted analysis dashboard.
- Web-Based Application: This tool is offered as a desktop client and not available as a web-based application.
- Large Corpora: MonoConc is designed as a desktop client for personal users and thus lacks the capacity and scalability to handle large corpora and to be run on the server side.
- Reusable: The non-pro version of this tool is downloadable for non-commercial use, however it's not reusable for further development as the source code is not available.

2.2.6 SketchEngine

SketchEngine is currently one of the most commonly used corpus analysis tools and developed by the company Lexical Computing as it builds on the foundation of Adam Kilgarriff and Pavel Rychlý's early work. This tool takes its name from one of its core functionalities: the word sketches, a single page summary of a word's grammatical and collocational behavior [25]. SketchEngine is offered both as a web service and a web-based tool, and the currently developed versions are offered as commercial products. However an open-source version, called NoSketchEngine, combining the core components of SketchEngine, namely Manatee and Bonito, is available without some of the features of the commercial version such as word sketches, thesaurus and keyword computation [8].

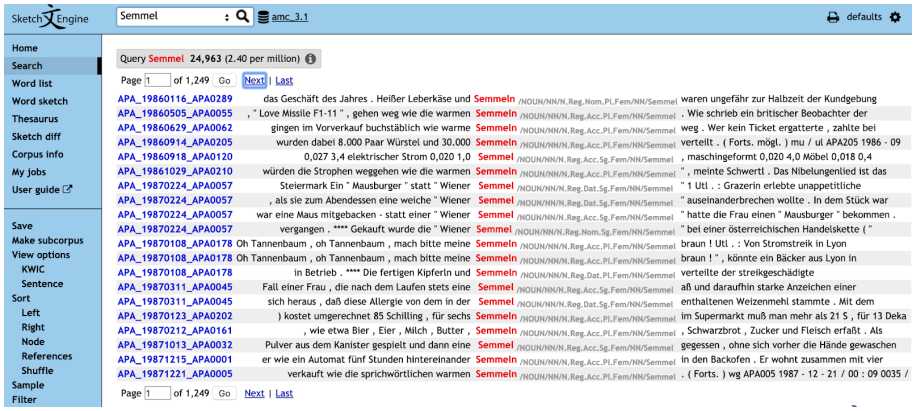

Figure 11: Keyword-in-context (KWIC) view displaying the results on a lemma search on SketchEngine.

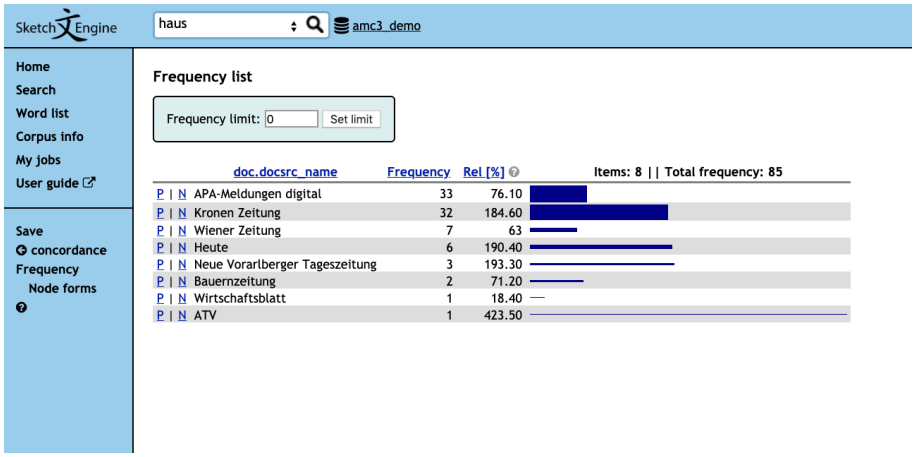

Figure 12: Frequency list view displaying the results as a table and a basic bar chart in SketchEngine.

- Design Study: There is no design study at the basis of the development and design of SketchEngine, to the best of our knowledge.
- User Evaluation: Developers of SketchEngine state that evaluation tests with users of Sketch Engine on different experience levels are currently in progress [27], however there are no results available to this date, to the best of our knowledge.
- Multi-Faceted Dashboard: SketchEngine applies a traditional pages-layout rather than a multi-faceted analysis dashboard.
- Web-Based Application: This tool is offered as a web-based application.
- Large Corpora: This tool has the capacity and scalability to handle large corpora.
- Reusable: Commercial version of this tool is not freely downloadable and reusable. However the open-source version NoSketchEngine, lacking the word sketch functionality, is available to download and reuse.

2.2.7 WordSmith Tools

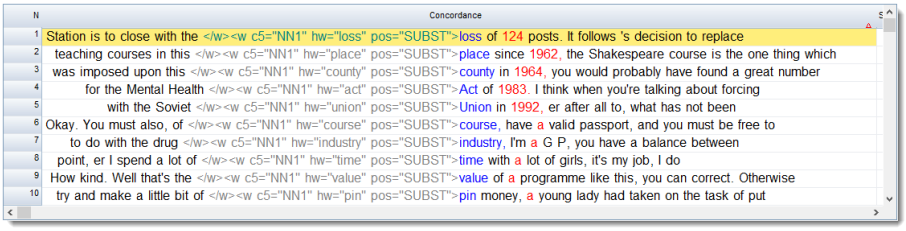

Figure 13: Concordancing capabilities of WordSmith Tools [34]

WordSmith Tools, continuously developed by Mike Scott since 1996, have been used by language teachers and students, who investigate language patterns in different languages, and by the Oxford University Press for their lexicographic work to prepare dictionaries [35].

This integrated suite of programs can be downloaded and run on a Microsoft Windows operating system. Other operating systems are not supported natively. The tool is intended to work with the own text collections of the users and provides functionalities such as concordancing, wordlisting, word-clustering and frequency analysis. The tool is offered as commercial software and the users can activate the downloaded copy with a license bought on the company's website.

- Design Study: For this tool, there is no design study at the basis of the development and design, to the best of our knowledge.

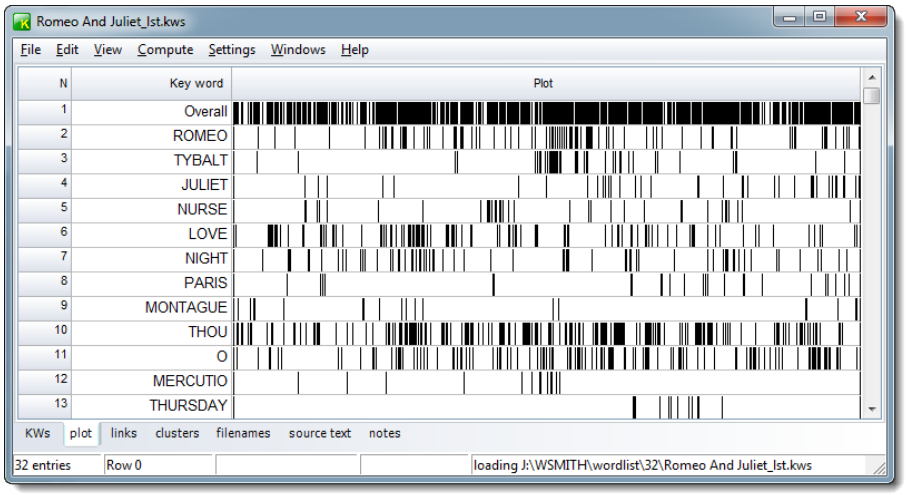

Figure 14: Keyword list plot view on WordSmith Tools [34]

- User Evaluation: There are no results available for user evaluation, to the best of our knowledge.
- Multi-Faceted Dashboard: This tool applies a layout based on separate pages and windows rather than a multi-faceted analysis dashboard.
- Web-Based Application: This tool is offered as a desktop client and not as a web-based application.
- Large Corpora: This tool has the capacity and scalability to handle large corpora.
- Reusable: This is a commercial tool and is not freely downloadable and reusable.

2.2.8 Wmatrix

Wmatrix was created by Paul Rayson in 2008 and has been continuously developed since. This web-based tool offers corpus analysis and comparison functionalities such as frequency lists, concordances and keyword lists [9]. The tool runs as a service on the servers of Lancaster University and can only be accessed with a valid user account, which either belong to the members of Lancaster University or the paying users with a valid yearly license.

The service allows users to upload their own corpora and conduct analysis such as frequency profiles, concordances, keywords, frequency lists, N-grams, c-grams and collocations [9].

Figure 15: Concordance view on the web-based service of Wmatrix2 [9]

- Design Study: There is no design study at the basis of the development and design of Wmatrix, to the best of our knowledge.
- User Evaluation: Even though official information is available about the evaluation of the statistical and algorithmic approaches behind Wmatrix [9], there are no results available for user evaluation, to the best of our knowledge.
- Multi-Faceted Dashboard: This tool applies a layout based on separate pages rather than a multi-faceted analysis dashboard.
- Web-Based Application: This tool is offered as a web-based service.
- Large Corpora: This tool has the capacity and scalability to handle large corpora.
- Reusable: This is a commercial tool and is not freely downloadable and reusable.

2.2.9 Voyant Tools

Voyant Tools, created by the lead of Stéfan Sinclair and Geoffrey Rockwell, is a web-based tool for analysis, reading and visualization of text collections. This

application aims to serve as a generic tool for text analysis and to help a wide range of users such as students, researchers, journalists and market analysts [39].

This web-based tool offers functionalities such as importing text documents in various formats, analysis of term frequencies and distributions, close reading and distant reading, as well as being open-source and freely available for download for personal usage [39].

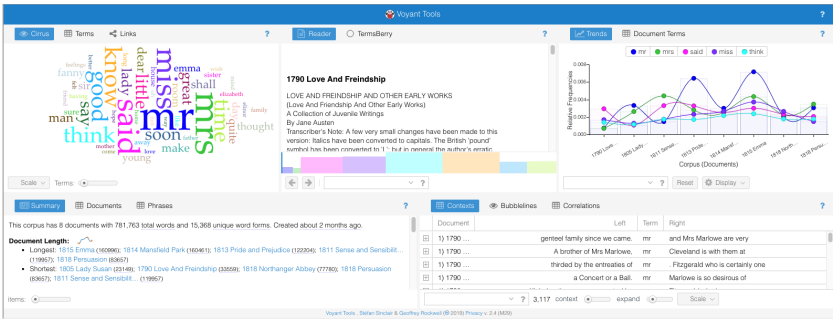

Figure 16: Dashboard layout for text analysis on Voyant Tools [4]

- Design Study: Even though Voyant Tools puts a substantial importance on the design principles, there is no design study at the basis of the development and design, to the best of our knowledge.
- User Evaluation: There are no results available for user evaluation, to the best of our knowledge.
- Multi-Faceted Dashboard: This tool applies a multi-faceted analysis dashboard with flexible components and various plugins.
- Web-Based Application: This tool is offered as a web-based application.
- Large Corpora: Based on our experiments this tool is not designed to provide the capacity and scalability to handle large corpora.
- Reusable: This tool is freely downloadable and reusable.

2.3 Related Work Conclusion

These surveyed publications (See: Table 1) yield interesting results in regard to the lack of usage of design study and evaluation methodologies in the design and development of corpus analysis tools. This observation is one of the key indicators for our visualization research project, which emphasise the necessity

Tool Name	D.Study	Evaluation	Dashboard	Web-based	L.Corpora	Reuse
AntConc	✗	✗	✗	✗	✗	✗
corpus.byu.edu	✗	✗	✗	✓	✓	✗
CQPweb	✗	✓	✗	✓	✓	✓
LancsBox	✗	✗	✗	✗	✓	✗
MonoConc	✗	✗	✗	✗	✗	✗
SketchEngine	✗	✗	✗	✓	✓	✗
WordSmith	✗	✗	✗	✗	✓	✗
Wmatrix	✗	✗	✗	✓	✓	✗
Voyant Tools	✗	✗	✓	✓	✗	✓

Table 1: Surveyed tools and services for corpus analysis solutions

of well-implemented design studies in regard to corpus and text visualization contributions.

CQPweb and Voyant Tools come out as better examples of methodological design and development from the surveyed tools as these mention somewhat close collaboration with users and give some explanations on decision making processes regarding the design and improvement of the tool. These surveyed examples may be helpful for our research in terms of showing methodological efforts in the domain of text and corpus visualization.

We observe that the lack of multi-faceted / multi-dimensional dashboards and the support for web applications are some of the main challenges and drawbacks of many solutions. This observation is consistent with the feedback we collected from multiple interviews with domain experts and linguistics researchers who try and use various tools during their projects. After a detailed analysis of the contributions with different layouts, we conclude that Voyant Tools offers some of the more complete combinations in terms of handling a wide variety of metadata. However, scalability and handling of large corpora remains a drawback for this tool.

Less than half of the surveyed contributions offer source code of their software with open-source licenses. Based on these observations and the tasks analysis that we present under the "Tasks" section, we conclude that none of the available solutions provide immediate answers to the problems our domains experts are facing and lack some features or capabilities for being a general purpose corpus visualization tool for large and annotated corpora, which can serve the needs of non-programmer users. This observation confirms the validity of our problem statement and the necessity of a visualization research in this area.

3 Methodology

In order to have a solid base for our methodology, we collected eight papers from the fields of visualization and human computer interaction (HCI). After careful reading analysis of many papers which propose various frameworks, guidelines and methodologies in these fields, we concluded that these selected papers (See: Table 2) offer the most comprehensive and consistent approaches, which should provide a basis for our work.

Publication	Year	Topic	Main contribution
Tory, Möller [41]	2004	Design	A research methodology for human factors in visualization design
Tory, Möller [42]	2005	Evaluation	Evaluation of visualizations with expert reviews
Shneiderman, Plaisant [38]	2006	Evaluation	Guidelines for conducting MILCs for information visualization
Valiati, Freitas, Pimenta [44]	2009	Evaluation	Review of MILCs-based usability evaluation methods for visualization tools
Munzner [31]	2009	Design	A four-layered nested model for visualization design and validation
Sedlmair, Meyer, Munzner [37]	2012	Design	Nine-stage design study methodology for visualization research
Isenberg, Isenberg, Chen, Sedlmair, Möller [20]	2013	Evaluation	A systematic review and categorization of evaluation practices
Lam, Tory, Munzner [28]	2018	Design	A framework for abstracting domain problems in visualization research

Table 2: Publications collected from design and evaluation methodologies.

Tory and Möller [41] present how visualizations offer humans cognitive support to accomplish complex data analysis tasks with visual information representations. Different methods for cognition support and their possible applications are showcased. It is interesting that this paper reports a limited amount of contributions (at that time 23 percent of TVCG papers) in visualization and computer graphics publications, which include a human factors component. Moreover only 2.8 percent of the abstracts were reported to mention a user study. It is wise to assume that these percentages have increased since the publication of this paper, however the argument for a human factors component in visualization research and the proposed methods are still highly valid.

The importance of validating the usefulness of visualization solutions for real people doing real tasks is emphasized by Tory and Möller [42] and if this is not the case even the most well intentioned and technically developed visual displays could be ineffective. Thus they argue that expert reviews are one of the more valuable techniques for evaluating visualization projects and offer insights on how and when to engage with expert reviews.

Shneiderman and Plaisant [38] present a method called Multi-dimensional In-depth Long-term Case studies (MILCs) which is well suited for evaluating visualization systems as well as human computer interaction solutions. This work also provides guidelines for conducting MILCs for visualization research and becomes relevant for understanding the importance of long-term evaluation in regards to our project.

Results of three case studies conducted as MILCs are reported and discussed by Valiati et al. [44] and they offer an insightful review on MILCs-based usability evaluation methods for visualization solutions. These reviews display the importance of conducting experiments with relevant users and focusing on tasks which cover the situations that a real user would face while using the visualization tool.

Munzner [31] present a four-layers nested model for visualization design and validation. This model is highly relevant for our work since it offers prescriptive guidance for choosing appropriate methods for both design and validation for different phases of a visualization research project.

A systematic overview of the evaluation practices from visualization publications is provided by Isenberg et al. [20], which categorizes the most common evaluation goals and methods in the visualization community. Eight presented evaluation scenarios are extracted from an extensive literature review of visualization publications, and different scenarios bring up different study goals and research questions. This guideline is valuable for our work since it may provide support for reaching decisions about the most effective evaluation of a certain visualization research scenario that we are facing.

Sedlmair et al. [37] define the concept of design studies, give successful and unsuccessful examples from their past work, and lay the guidelines for designing a visualization system with this approach as they propose a nine-stage framework as the methodology for design studies. In our work we use this framework and the proposed design study methodology as the basis of our research concept.

Lam et al. [28] present a framework to bridge the gap between low-level actions and higher-level context of analysis goals based on analysis reports derived from 20 design study papers published at IEEE InfoVis 2009-2015. The proposed framework has two axes illustrated by nine analysis goals and each goal is placed under the axes of specificity (Explore, Describe, Explain, Con-

firm) and number of data populations (Single, Multiple). This framework serves a valuable role in our design study as the abstraction and analysis of domain problems is a crucial starting point in any visualization research project.

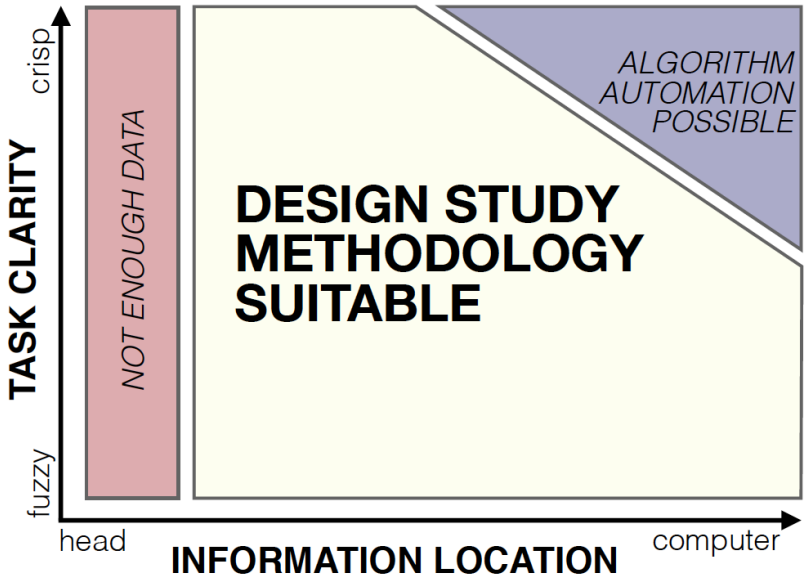

Figure 17: Suitability for the application of the design study methodology based on task clarity and information location [37]

A design study is defined as a project in which a specific real-world problem faced by domain experts is analyzed by visualization researchers and a visualization system is created to support solving this problem [37]. In the next stage, the proposed design is validated, and the visualization researchers reflect about lessons learned in this process.

After analyzing the suitability for design study methodology based on task clarity and information location (Figure 17) in our project, we conclude that even though the information location is mostly in computers i.e electronic corpora, the clarity of tasks in this specific domain is rather fuzzy and user interviews to analyze and clarify the most importing tasks would be needed. These specific steps and the lessons learned from our selected methodology publications will lead the way for the next phases in our research project.

4 Domain Characterization

4.1 Data

In this design study project we collaborate with domain experts who work with two large contemporary media corpora:

- Austrian Media Corpus (AMC), created as part of a cooperation between the Austrian Academy of Sciences and the Austrian Press Agency, covers almost the entire Austrian print media landscape of the past two decades, containing 40 million documents, constituting more than 10 billion tokens. AMC ranks among the largest collections of its kind as a contemporary German language corpus.
- CORPES is a reference corpus for 21st century Spanish language, which is created and maintained by the Real Academia Española and its affiliations in 22 hispanophone nations. CORPES offers one of the most extensive data regarding the Spanish language with a total of 285.000 documents and 286 million tokens.

In order to conduct this collaborative design study it's crucial to be in contact with domain experts and researchers of the fields of linguistics and humanities, who work with the contemporary media corpora listed above.

In this paper the data and query resources from Real Academia Española were made available thanks to a travel grant by ELEXIS, the European Lexicographic Infrastructure, which allowed an onsite visit to RAE's historical and current infrastructure and a great deal of knowledge exchange in terms of corpus analysis tools and technologies.

4.2 Users

In this section we describe the characteristics of our users, their current research topics and the workflows of conducting their research.

4.2.1 User Profile

We select and contact various linguistics researchers and students, who use and/or need to use corpus analysis and visualization tools for their projects. The general profile of our users can be described as follows:

- Age: 20+
- Gender: Not a factor
- Education: University education in linguistics / humanities
- Research Topic: Projects in research institutes in relation to studies and investigations of linguistics and change in language
- Technical Skills: Average computer user. No / low programming skills.

With 10 selected interviewees we conducted multiple rounds of interviews paying close attention to their specific background and research interests. As shown in detail on the table of interviewed users (See: Table 3), all our interviewees fit into the user profile defined above and even though they might have different data to analyze and different research questions, they have a set of common tasks in their research projects, which are detailed in the upcoming sections.

4.2.2 Research Topics and Workflows of the Users

Analysis of Regional Linguistic Varieties: (User1, User4, User7)

- Starting with a defined set of words and some defined regions.
- Querying the corpus with these variables and evaluating hypotheses.
- Extracting data, documenting results and/or proposing new hypotheses.

Diachronic Lexicographic Analysis: (User2, User5)

- Focusing on a time period and some defined topics / set of words.

User	Education	Corpus	Current Topic
User1	MA Translation-Studies	AMC	Comparing 30 pairs of sentences/words with German and Austrian varieties in terms of regions and time.
User2	MA German Philology	AMC	Searching for articles on the topic of "gender-equal language", and looking at how the topic has been covered.
User3	PhD Linguistics	Custom	Comparison of two corpora in Spanish language from Mexico to investigate political influences in language, mostly focusing on collocations and high-freq. words.
User4	MA Linguistics	Custom	Working with an ancient language in Mexico and mapping its etymology and regional changes.
User5	BA English Philology	AMC	Focus on the word "Gutmensch", understanding the political context, looking at what kind of articles this word appears in, context research.
User6	PhD Linguistics	CORPES	Investigating the contemporary Spanish corpora to find out the relationship of social factors in language and spelling changes over time and regions.
User7	PhD Linguistics	AMC	Multiple projects using AMC for comparing regional varieties and/or spelling mistakes.
User8	MA Theater, Film and Media Studies	AMC	"Gedenkjahr" Project. Analyse how politicians are talking about this special topic in media.
User9	PhD Translation-Studies	AMC / ParlAT	Investigating language change using AMC and ParlAT (A collection speeches in the Austrian Parliament) corpora in Austria in relation to politics, media and social factors.
User10	MA German Philology	AMC	Compound words in standard German language. E.g.: first word + s + second word. Regional differences are important.

Table 3: Table of users interviewed in this design study (names of the interviewees are anonymized for the protection of personal information).

- Investigating the change in frequencies and reviewing the context in concordances.
- Extracting data, documenting results and/or proposing new hypotheses.

Political and Social Discourse Analysis: (User3, User8, User9)

- Focusing on sources (authors, politicians, media etc.) and some defined topics.
- Evaluating the relationship between frequencies and social events / phenomena.
- Extracting data, documenting results and/or proposing new hypotheses.

Analysis of Orthographic Change: (User6, User10)

- Starting with a defined set of words and some focus time span and/or regions.
- Querying the corpus with these variables and evaluating hypotheses.
- Extracting data, documenting results and/or proposing new hypotheses.

4.3 Tasks

As suggested in the multi-level typology of abstract visualization tasks by Brehmer and Munzner [14] the main pillars of the abstract tasks typology are "Why?", "How?" and "What?", defined from the user's perspective. As we derive and list the tasks for our tool in this section, we base our task definitions to answer these three questions for each research interest category (Figure 18).

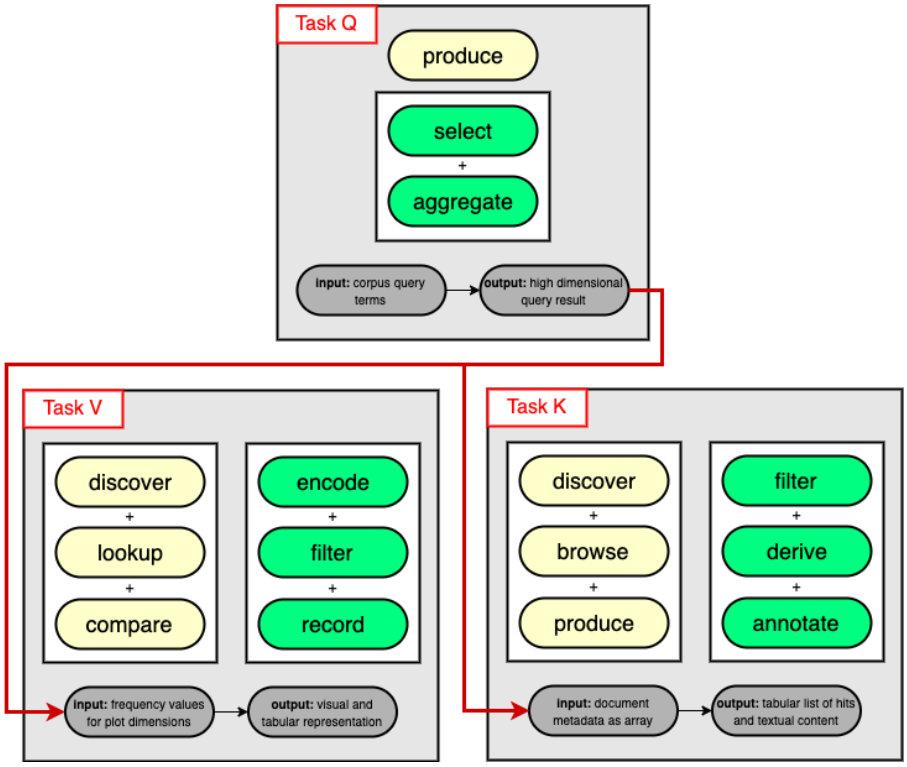

Figure 18: Task typology of our visual corpus analysis tool based on Brehmer and Munzner [14]. In this diagram we use the original color scheme of this methodology, where yellow colored nodes represent "Why?", green "How?" and gray "What?" (input and output). Task Q, V and K are main task blocks which are referenced by varying but related derived tasks.

We group the tasks, which need to be fulfilled by our tool, in these three main task blocks with specific sub-nodes:

- Task Q: Query building tasks, which take corpus query terms defined by the user as input and allow the user to produce results with the help of multi-query input area with selection options.

- Task V: Visualization tasks for the discovery, lookup and comparison of frequency values in different dimensions such as time, region, newspaper sources and sections. These values will be encoded, further filtered and exported by the user. Output of this group of tasks can be visual (image as SVG) or tabular (data as CSV).
- Task K: Keyword-in-context based tasks, which serve the purpose of discovery in documents, browsing large collections and producing new sets of articles (subcorpora). The means of achieving this type of tasks can be via filtering, deriving and annotating. The input for this block is the document metadata as array and the output is a tabular list of hits with actual textual content.

Based on the previously listed research topics and workflows of the users, we are able to derive the following generalized tasks, which are the main requirements of conducting a particular research question:

4.3.1 Analysis of Regional Linguistic Varieties

Derived tasks:

- Display frequency results divided by regions on geographic map (Task V)
- Compare queries of different words / phrases with each other (Task V)

4.3.2 Diachronic Lexicographic Analysis

Derived tasks:

- Display temporal trends of frequency results (Task V)
- Display the keyword-in-context in different time periods with sorting and filtering (Task K)

4.3.3 Political and Social Discourse Analysis

Derived tasks:

- Display sources / topics of frequency results (Task V)
- Search and filter by other words using the keyword-in-context results (Task K)

4.3.4 Analysis of Orthographic Change

Derived tasks:

- Query words / phrases with regular expressions, case sensitivity etc. (Task Q)
- Display temporal, regional and source distributions (Task V)

4.3.5 Common Tasks

Derived tasks:

- Allow querying multiple words / phrases at the same time (Task Q)
- Annotation of findings / results directly on the KWIC view with input-, single- or multi-selection (Task K)
- Create sub-corpus based on a selection of documents / queries from the dashboard (Task K)
- Compare results with the total partitions / distributions of the corpus (Task V)
- View and export general descriptive statistics about the corpus (size, frequencies, partitions, textual description etc.) (Task V)
- Download the results from single views as CSV and/or XLS (Task V)
- Export the visualizations as images (Task V)

4.3.6 Data Curation Tasks

These are some additional tasks from the perspective of the data provider and curator (in our case the maintainers of AMC and CORPES corpora), which are important to note for the decisions related to the "Implementation" section, but not directly connected to the above described user task typology.

- Single web-application to disseminate different corpora to the users
- Ability to adapt to different APIs of various corpus engines easily

5 Design Decisions

As stated in the Methodology section, our design process involves many iterations and feedback rounds with various domain experts to make the necessary decisions regarding our tool. In the following sections these iterative steps will be described with some examples.

5.1 Low-Fidelity Prototypes

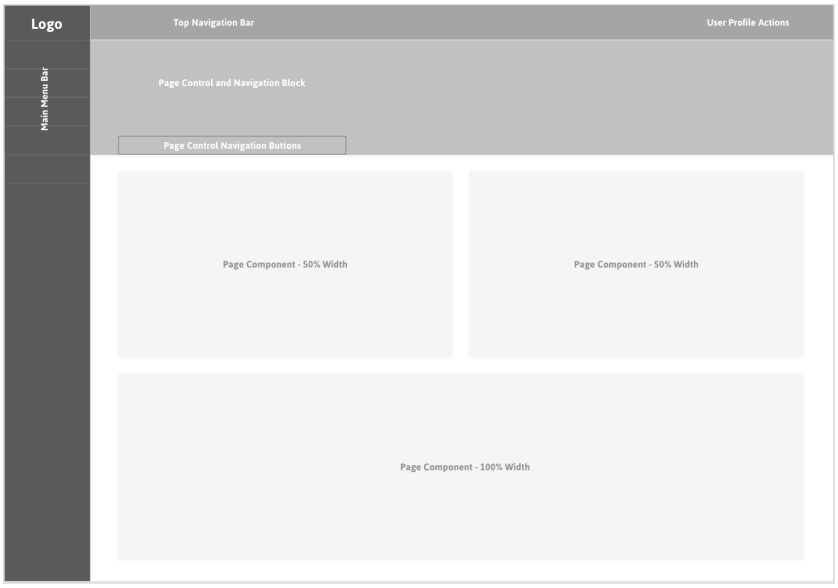

Figure 19: Initial wireframes for our web-based application.

In our design process the first step was to come up with a wireframe, which would be fitting to the needs of our application based on the previously derived tasks. On top of this wireframe we add some fundamental elements, which our users would need to use the most. At this stage we focus on some main components such as on corpus search and comparison views.

As we plan to increase the amount of detail and add interactions on data visualizations in the next steps, it is our aim to get feedback from our interviewees and from some HCI / VIS experts before we bring any design into a more "high-fidelity" level with UI features such as colors, typography and interactions.

The initial wireframes for our web-based application (Figure 19) consist of some fundamental layout elements. The main menu bar for navigating between

different pages is placed on the left side as a top level hierarchical element. This vertical bar has a logo on top, which is linked to the main page of the application, and the top level menu items underneath.

Our second level hierarchical element is the top navigation bar, which has functional links and buttons depending on the page, and user profile actions on the right side. Placement of these navigation bars are based on the current convention of web applications, which are used heavily by all levels of users. These decisions are successfully evaluated in the next iteration of the user interviews.

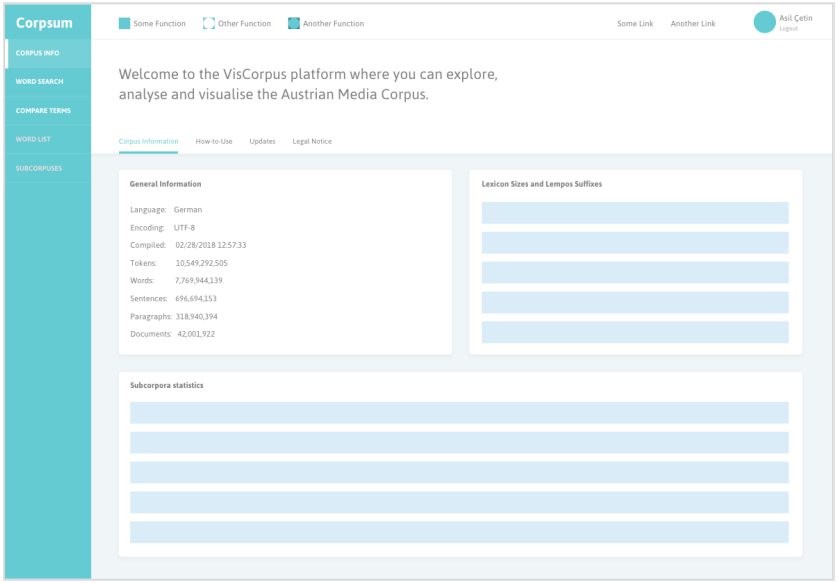

Figure 20: Start page with corpus information on the low-fi prototypes.

Main areas of the layout, which cover most of the screen space, are reserved for the contextual components and page controls (Figure 20). These elements are placed in varying sizes and order depending on the page. Their wrapper containers are designed with the principles of responsive design, so that the application can be viewed and used in a wide variety of screen sizes and devices.

As covered in the Related Work section, concordances and KWIC components are the main views of corpus analysis tools. In the long tradition of corpus linguistics research and its community, these views are mostly the starting points of any corpus linguistic research and thus we include these components in the first view after a query is made. The low-fidelity prototypes include buttons and navigation elements, which serve to fulfill some of the derived tasks such as viewing the results of multiple queries at the same time, sorting and filtering results on KWIC view and creating subcorpora based on a selection made on the KWIC view (Figure 21).

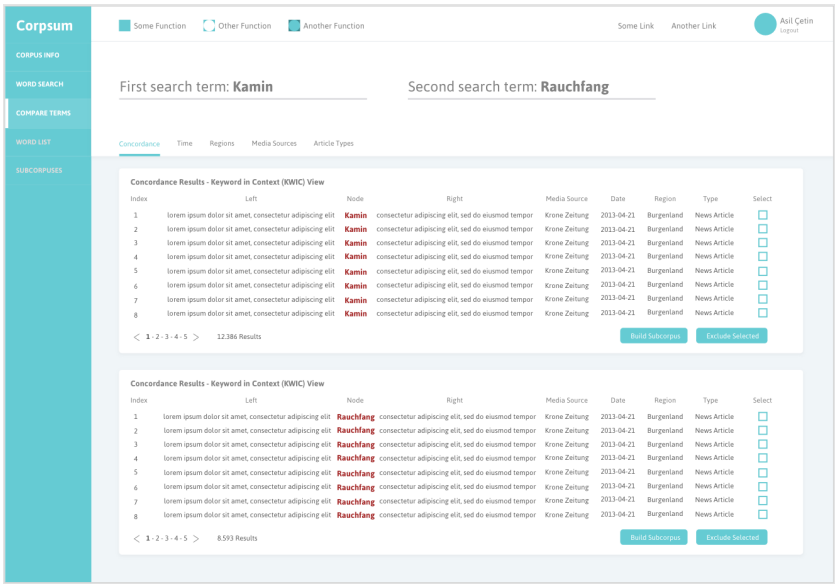

Figure 21: Keyword-in-context (KWIC) view with multi query option displayed on the low-fi prototypes.

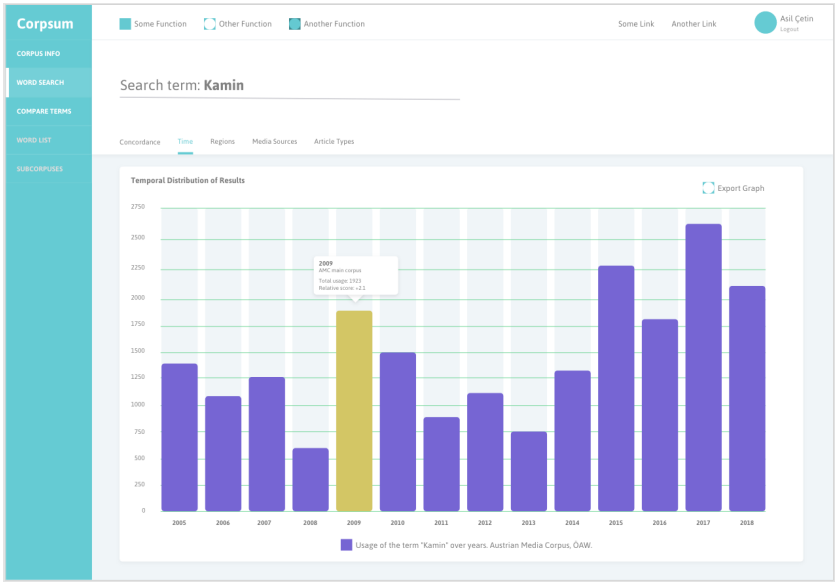

Figure 22: Temporal distribution of hits based on a given query is displayed on the low-fi prototypes.

Based on these low-fidelity prototypes we received feedback from our domain experts as well as some VIS and HCI experts. Even though the initial layout proposal was perceived as modern and clearly structured, one of the main criticisms was the tabs-layout for switching between and comparing different dimensions and distributions of frequencies of metadata on a given query (Figure 22). This feedback was one of the most productive inputs we received at this stage, because moving away from the tabs-layout allowed us to achieve one of the main requirements for building a multi-faceted interactive dashboard.

A quite positive feedback at this stage was the ability to make multiple queries and perform comparisons in various metadata dimensions, which most of the participants mentioned as a lacking feature in their currently used software tools.

Based on the analysis of the feedback received at this stage, we continue with high-fidelity prototypes, where we implement some of the most important components into a clickable prototype, which used a sample dataset to be processed.

5.2 High-Fidelity Prototypes

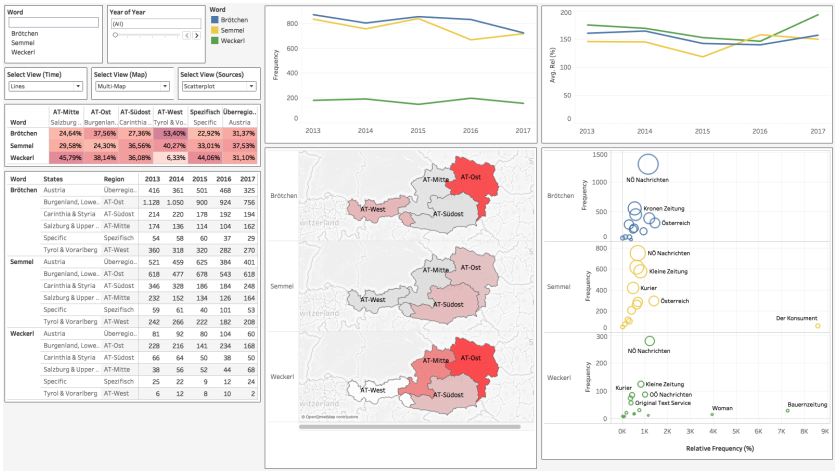

Figure 23: High-fi prototype using a sample data extracted from AMC.

At this stage of the project we focus on the visual analysis dashboard concept for our tool, which is aligned with the feedback received from the previous stage. We determine that the KWIC view will be on the main screen and offer some navigation, subcorpus creation and annotation features. Thus we save this component for the final design implementation and only focus on the interactive connected visualization components for the metadata query results (Figure 23).

In these prototypes we defined alternatives for some visualization components for different metadata dimensions such as temporal, regional and discourse distributions. For the temporal dimensions a simple line chart with multiple traces is compared against an area chart with the same data. A multi-map choropleth is compared against a single-map choropleth with pie-charts as state data points for the regional dimension. For the discourse analysis we have a scatter plot, which displays relative and absolute frequencies per media source, compared against a bar graph using the same data (Figure 24).

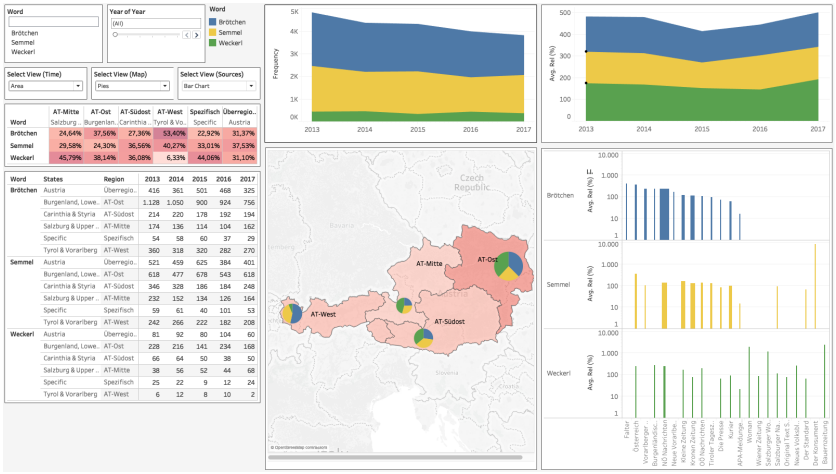

Figure 24: Alternative visualization components in the high-fi prototype using sample data extracted from AMC.

The way we conduct evaluation for these prototypes are based on one-to-one interviews with each participant following a set of predefined tasks and questions. In order to implement a clickable and interactive high-fidelity prototype, we use sample data extracted from AMC and CORPES, and we bind the data to a multi-view dashboard in the visualization software Tableau. Each interviewee receives the control of the dashboard on a desktop PC with a monitor, keyboard and mouse, and is asked to complete some tasks using this dashboard. Between each task we ask the interviewees questions to get their input regarding whether a certain visualization component is useful for solving this task. In order to avoid bias regarding the comparison of the alternative visualization components, we shuffle the order in which different layouts are presented to the interviewees.

The feedback we gathered from this round of interviews and the lessons learned regarding our design can be summarized as the following:

- When multiple terms are compared with each other, a line chart for the temporal distribution is preferred.

- When a topic consisting of multiple terms is observed over time, an area chart for the temporal distribution is preferred.
- A multi-map choropleth is perceived to be easier to read and understand compared to the single-map choropleth with pie-charts.
- A scatterplot for the distribution of media sources is generally preferred over the bar graph layout since it makes the axes for the relative and absolute frequencies more explicit.
- However, some users mention an improved ability to compare a specific media source in multiple queries in the bar graph since the data points for the same media source is vertically aligned.

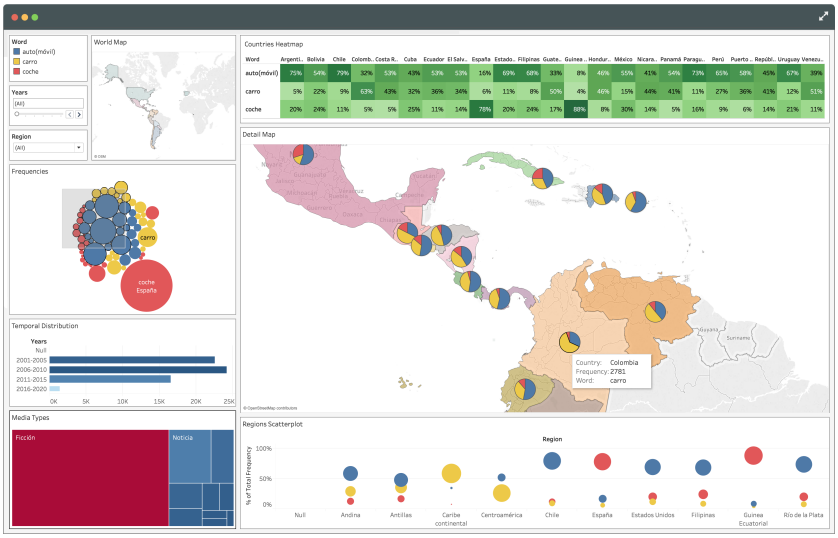

Figure 25: Dashboard prototype with altered visualization components as a high-fi prototype using sample data extracted from CORPES Spanish corpus.

At the end of this round of interviews we ask the interviewees some questions to evaluate the direction our tool is heading. One of the questions is "How would you rate the usefulness of this tool for your research?", which is to determine the direct impact our solution can bring to our users' active research (Figure 26). Another question is to determine the novelty of the design and functionalities we propose and thus we ask "How many of these functionalities are existent in the tools you know or use?" (Figure 27). Answers to these questions are in five point Likert scale, where 1 is the lowest and 5 is the highest rating. 8 users took part in this specific evaluation round.

Generally we consider the feedback to be very positive at this stage. We observe that even some interviewees who seemed uninterested during the first

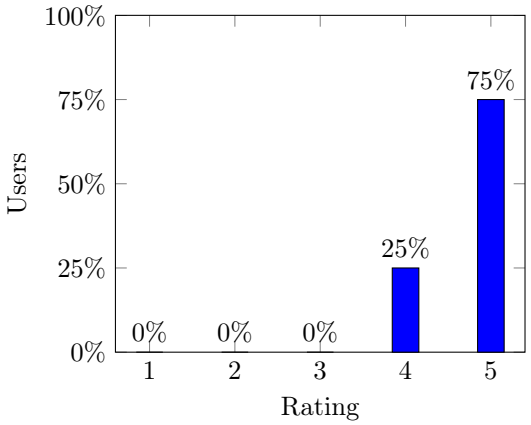

Figure 26: Answers to the question "How would you rate the usefulness of this tool for your research?" on a five point Likert scale (1 = very low, 5 = very high).

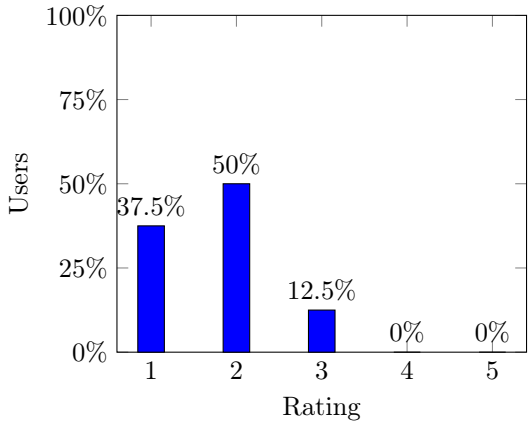

Figure 27: Answers to the question "How many of these functionalities are existent in the tools you know or use?" on a five point Likert scale (1 = very few, 5 = all of them).

round of interviews were very excited after using the prototype themselves and asked about when it will be available for them to use.

5.3 Final Design

Based on the previous iterations of our design and multiple rounds of interviews, we implement our tool as a web application. Detailed information about the technical decisions and structure of the application is given in the section "Implementation". When moving from the prototypes to the actual production application on a web environment, we use this opportunity to tweak and improve many existing elements in our design based on the latest feedback we gathered.

The web application has the "Corpus Info" screen as starting page (Figure 28). As discussed in the "Low-Fidelity Prototypes" section, the left side menu bar provides navigation and corpus selection functions. The top search bar allows users to enter multiple queries easily using a corpus query guide.

Our main query interface is the "Corpus Analysis" screen, which displays the results of multiple queries at the same time on a single dashboard (Figure 30). On this screen we provide visual analysis components for total relative and absolute frequencies, word forms, the most frequent multilevel combinations, KWIC view with annotation functions, temporal and regional frequencies, and distribution of media sources and newspaper sections.

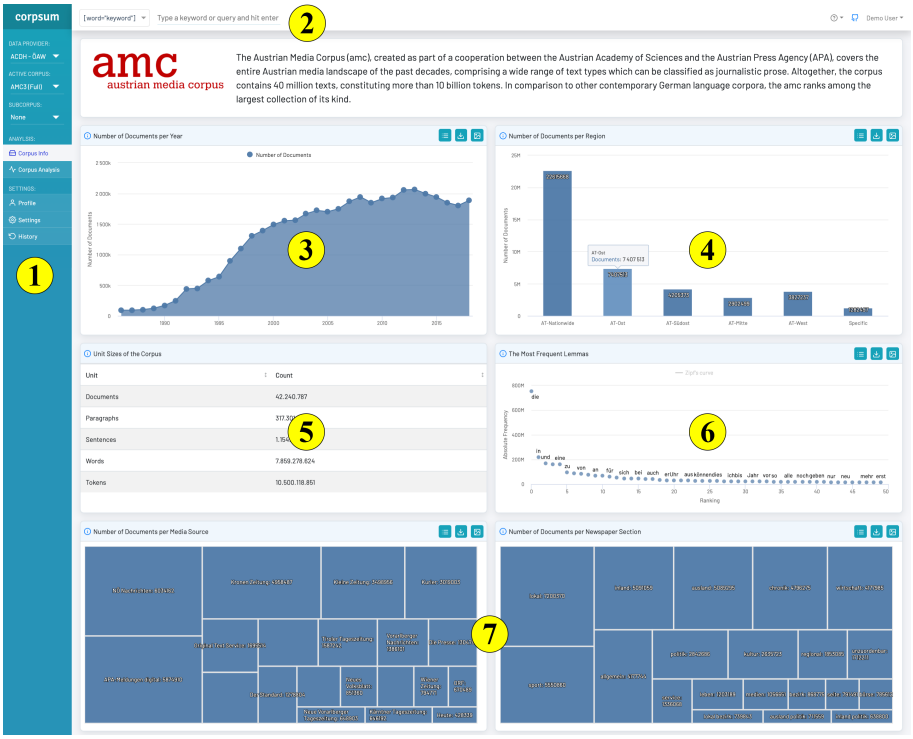

Figure 28: "Corpus Info" screen as the starting page displaying relevant information about the selected corpus: (1) sidebar navigation with corpus and subcorpus selectors, (2) top search bar with query guide, (3) an area chart displaying number of documents per year in the selected corpus, (4) a bar chart displaying number of documents per region, (5) a table listing the unit sizes of the corpus, (6) a scatter plot displaying the most frequent lemmas in the corpus, (7) tree-maps displaying the distribution of documents per media source and newspaper section.

All of these components come with interaction buttons on top, which allow to get more information about a specific visualization, download the data, export the visualization as SVG and view the data as table. Most of these components offer brushing, filtering and various interaction possibilities.

Based on the needs of our users in regard to annotation capabilities, we offer an enhanced annotation screen, which offers diverse and customizable annotation features integrated directly to the KWIC component (Figure 29). This enhanced annotation screen is currently used by the researchers of a collaboration project between University of Vienna and Austrian Academy of Sciences, where a media analysis to understand the characteristics of news reporting on algorithms, robotics and artificial intelligence in Austria is conducted.

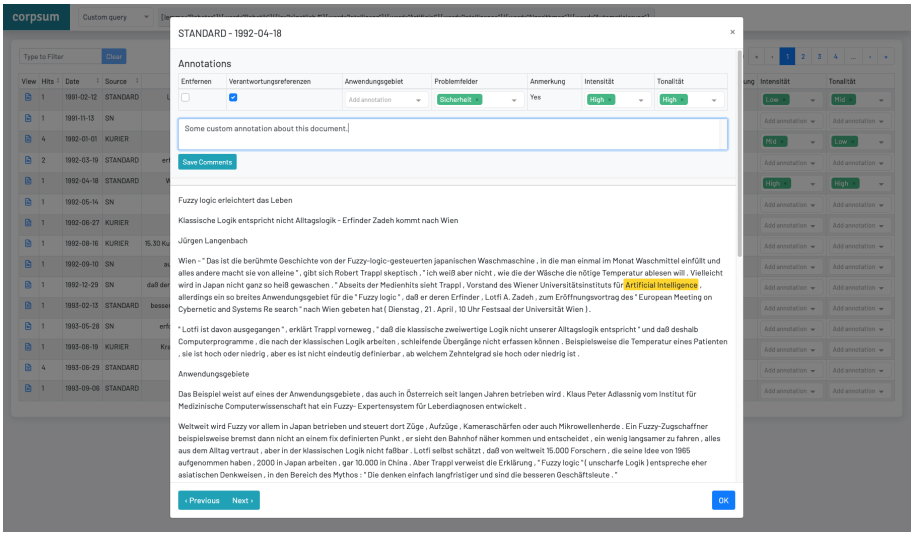

Figure 29: Enhanced annotation capabilities integrated on the KWIC component: the document viewer modal is opened after the selection of an article, and the annotations can be created or edited with various input areas and selectors above the article text.

The visual analysis functionalities of our tool are inspected in detail and evaluated further in the section "Evaluation and Case Studies", where we display some of the representative case study examples using our tool.

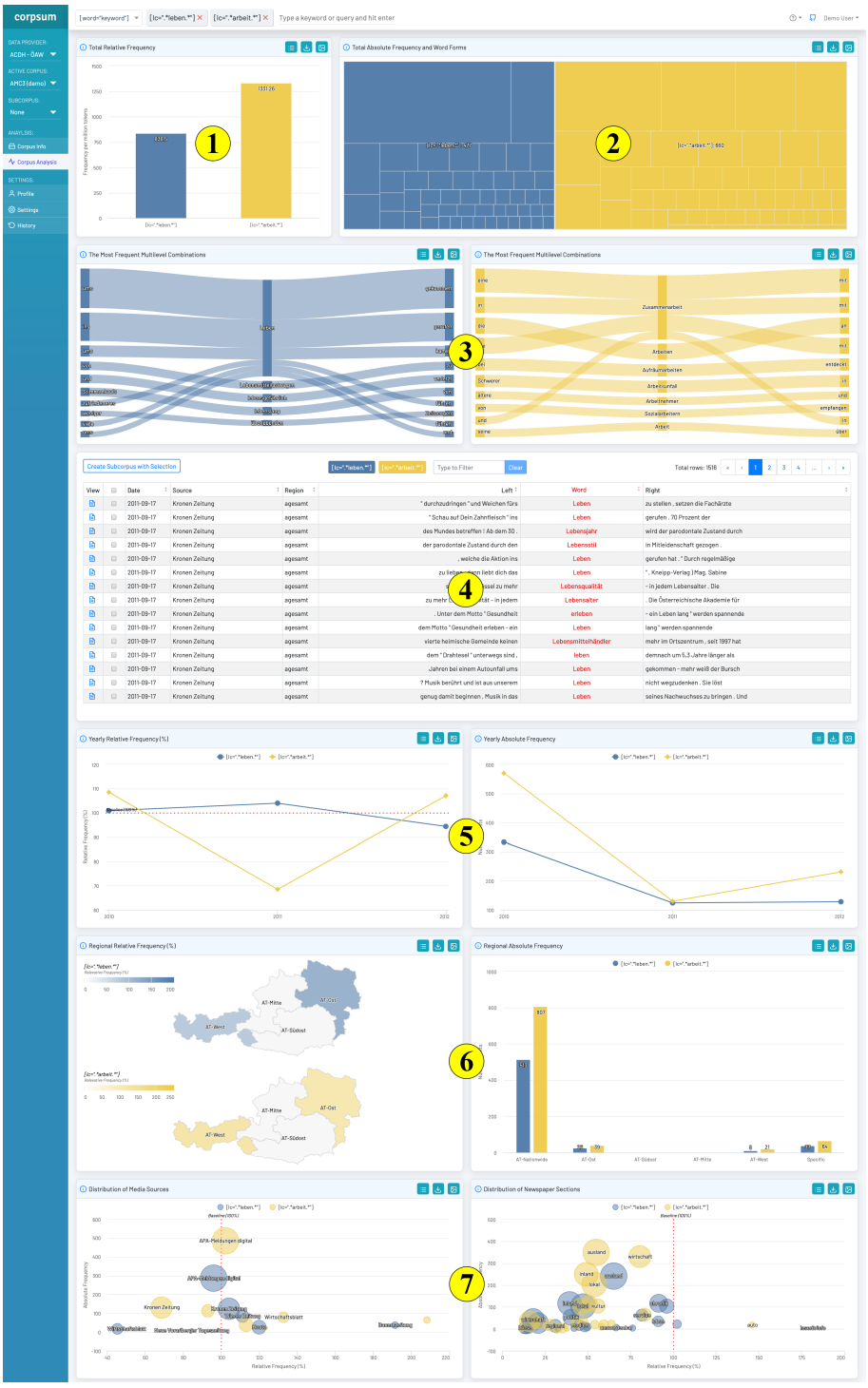

Figure 30: "Corpus Analysis" screen displaying the results of multiple queries on a single dashboard: (1) a bar chart displaying the relative frequencies for two separate queries, (2) a tree-map displaying total absolute frequencies and word forms inside sub elements, (3) sankey diagrams showing the most frequent combinations (one token left and right), (4) the KWIC view displaying documents as results with filtering and subcorpus creation features, (5) line charts displaying temporal relative and absolute distributions, (6) a choropleth map and a bar chart displaying regional relative and absolute distributions, (7) scatter plots displaying the distribution of media sources and newspaper sections with dimensions of relative frequency, absolute frequency and total original document size.

6 Implementation

The implementation of our tool follows a decoupled web application pattern, meaning that the corpus processing engine, annotation database, HTTP cache and the web-based visual analysis tool are separate from each other and communicate over the network connection using various predefined functions in representational state transfer (REST) protocols. (Figure 31).

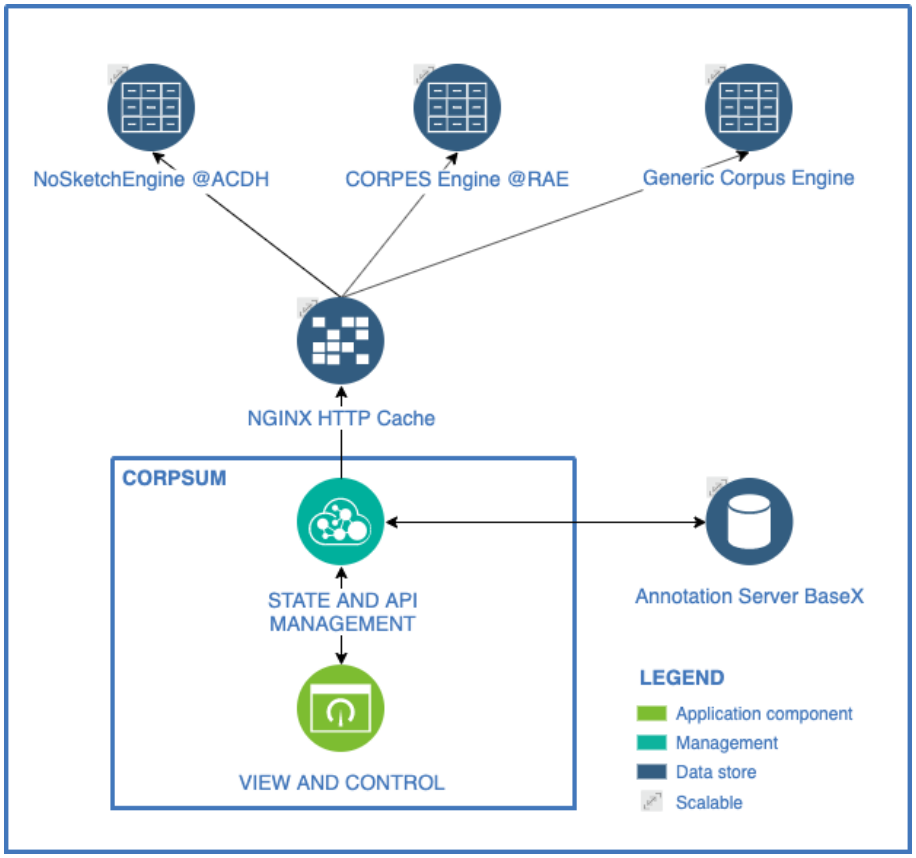

Figure 31: Decoupled web application architecture of our corpus analysis tool.

This type of system architecture allows our tool to function independently from a corpus processing engines and databases, as a state and API management interface takes care of the communication between various services with our application. This interface is designed to be as flexible as possible so that different corpus engines from different service providers can be bound with our tool easily.

As a JavaScript framework VueJS is used based on a NodeJS runtime en-

vironment. Various libraries and dependencies for state management, routing, frontend components and more are used as dependencies from the NodeJS package manager, npm. The up-to-date list of all dependencies and their versions can be found in the package.json file in our tool's code repository. The dependencies defined in this file at the time of writing this document can be seen below:

```
"dependencies": {
  "@johmun/vue-tags-input": "^2.0.1",
  "axios": ">=0.18.1",
  "bootstrap": "^4.3.1",
  "bootstrap-vue": "^2.0.0-rc.20",
  "highcharts": "^7.1.1",
  "highcharts-vue": "^1.3.1",
  "vue": "^2.6.10",
  "vue-feather-icons": "^4.22.0",
  "vue-multiselect": "^2.1.6",
  "vue-router": "^3.0.1",
  "vue-text-highlight": "^2.0.6",
  "vue-tour": "^1.1.0",
  "vuex": "^3.1.0"
},
"devDependencies": {
  "@vue/cli-plugin-eslint": "^3.5.0",
  "@vue/cli-service": "^3.6.0",
  "@vue/eslint-config-airbnb": "^4.0.0",
  "babel-eslint": "^10.0.1",
  "compression-webpack-plugin": "^3.0.0",
  "eslint": "^5.8.0",
  "eslint-plugin-vue": "^5.0.0",
  "eslint-config-airbnb": "^17.1.0",
  "node-sass": "^4.11.0",
  "sass-loader": "^7.1.0",
  "vue-template-compiler": "^2.5.21",
  "webpack-bundle-analyzer": "^3.3.2"
}
```

Figure 32: List of dependencies from the package.json file.

7 Evaluation and Case Studies

In the final phase of our user interviews an evaluation test takes place, where the users take the control of our corpus analysis tool for around an hour in a consistent control setting with a modern desktop PC and afterwards are asked questions to evaluate their satisfaction with our application. In the following section these answers are displayed on a Likert scale, where 1 represents the lowest (negative) and 5 represents the highest (positive) ratings. 8 users took part in this evaluation round.

7.1 Evaluation

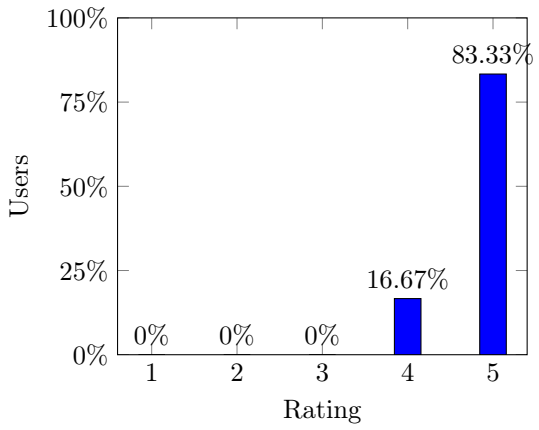

Figure 33: Answers to the question "How would you rate the informativeness of the Corpus Information screen?" on a five point Likert scale (1 = very low, 5 = very high).

The results of this user evaluation for our final implementation presents a high approval from the domain experts. During this round of interviews we discovered some minor bugs and aspects to be improved, and received feature requests by the users for some extended functionalities. Based on this feedback we made even further improvements, however the main structure and functionalities of the tool stayed consistent with our final design decisions and implementation. The interviewees in the evaluation round expressed high interest in using our tool actively for their ongoing projects and with some of them we started collaborations in this regard.

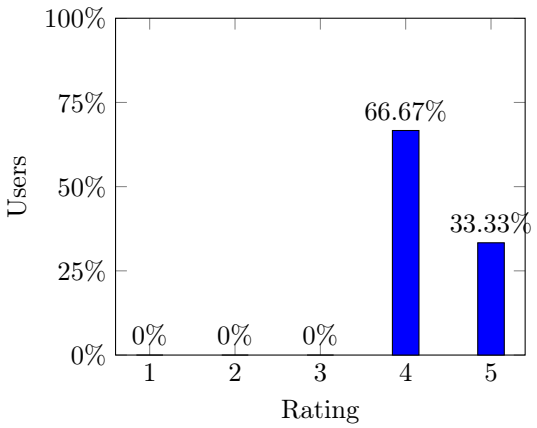

Figure 34: Answers to the question "How much easier is it to make corpus queries compared to the tools you used before?" on a five point Likert scale (1 = much harder, 5 = much easier).

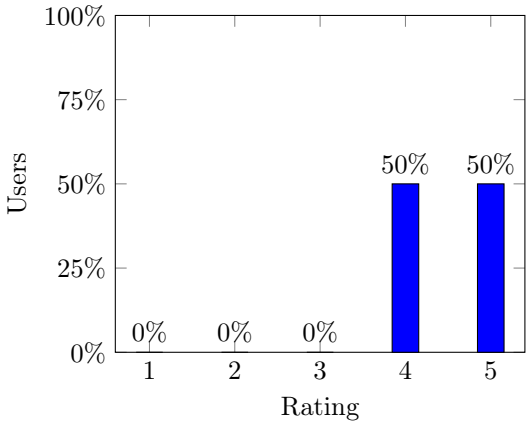

Figure 35: Answers to the question "How would you rate the help this analysis tool could offer to your project?" on a five point Likert scale (1 = very low, 5 = very high).

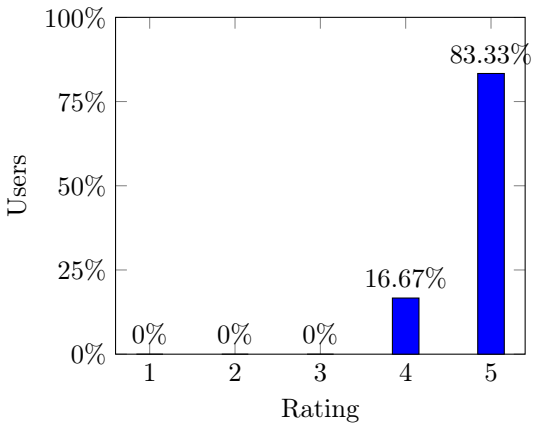

Figure 36: Answers to the question "Do you see yourself and your team members using this tool actively?" on a five point Likert scale (1 = not at all, 5 = very actively).

7.2 Case Studies

In this section we inspect the visual analysis functionalities of our tool in detail and conduct a qualitative evaluation with the representative case study examples, which we gather while the researchers use the software for their respective purposes and needs.

7.2.1 Case Study 1

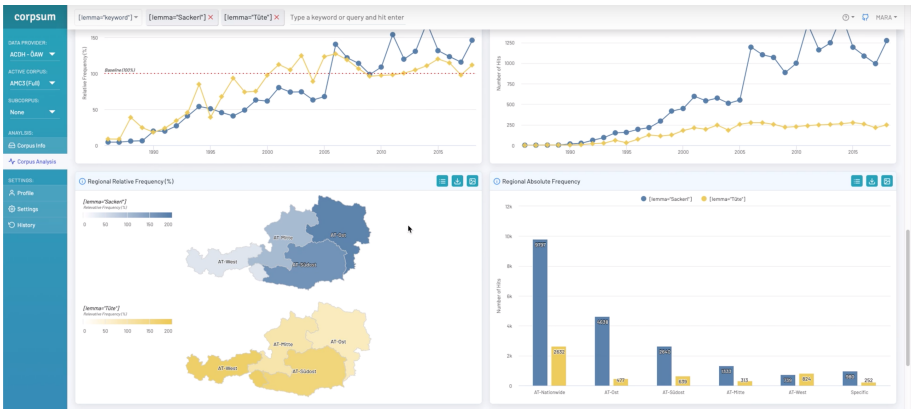

Figure 37: User1 is inspecting the multi-query results on regional and temporal dimensions

In this case study we focus on User1's experience using our tool for her research. User1 has a master's degree in Translation Studies and works mainly with German and Spanish language. In her research project she has a list of 30 pairs of sentences / words, in which each item has a corresponding German and Austrian variety. This list was previously generated by conducting a survey with children and adults living in different regions of Germany and Austria. In her project she queries these terms in the Austrian Media Corpus to inspect the similarities and differences between the survey and corpus data. In this project the main metadata dimensions of interests are regional and temporal. In our case study she queries some of the words, which are from the predefined search list, such as "Tüte" vs. "Sackerl" and "Mütze" vs. "Haube".

The summary of User1's main usage steps with the tool is as follows:

- She starts with the "Corpus Info" screen and focuses on the distribution of the documents per region. This allows her to have a clear understanding of the characteristics of the corpus in this dimension.
- On the "Corpus Analysis" screen she makes two queries using the query builder input area. Since she wants to compare two related terms in all possible word forms, she selects the "lemma" option and types in the search terms. At this point she mentions that making two queries simultaneously and this easily was not possible with the tools she has used before.
- After inspecting the statistics about total relative and absolute frequencies, she focuses on the temporal and regional distributions. For her search terms "Mütze" and "Haube", which are representative for the regional differences in German and Austrian languages, she points out that the choropleth map view delivers exactly the results her hypothesis has suggested. She expresses that presenting the regional results in such a map was not possible in her previous workflow, and the ability to export and use such a visualization would have been an great added value for her publication.
- As she inspects the word combinations provided by the sankey diagrams and the results in the keyword-in-context view, she quickly recognizes that the term "Haube" is also used in another meaning, in the context of an automobile's motor. For the accuracy of her regional variety comparison the hits related to this meaning should be filtered out. In order to achieve this she uses the filtering option on the KWIC view and finds all the articles from the "auto" newspaper section. Moreover she makes some regular text searches in the KWIC view for automobile-related terms. After deselecting these false positives she saves the results as a new subcorpus and continues her research using only the documents from this newly created collection. She mentions that this process of filtering and creating

new subcorpora is much easier and time-efficient than the tools she has used before.

- Her workflow requires multiple iterations of the previous step to be repeated while close-reading the articles and making sure that no false positives are remaining. According to her feedback, using our tool helps her a lot with recognizing outliers, since it gives a complete overview of the results from different metadata dimensions.

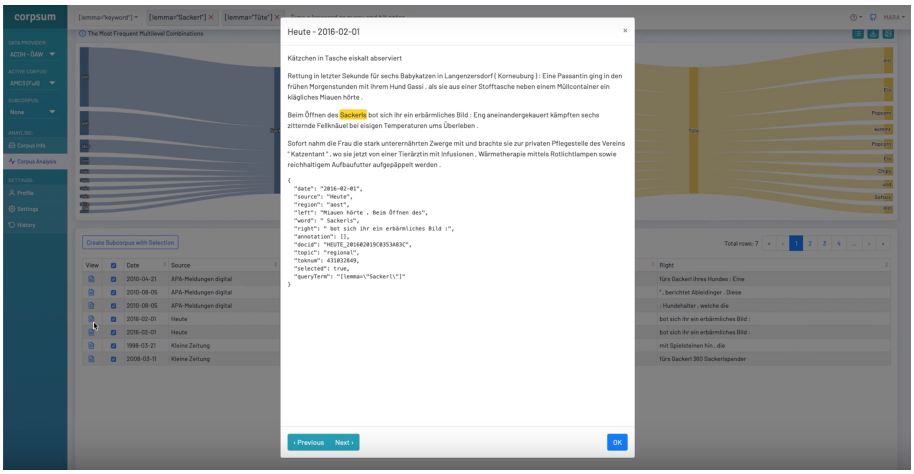

Figure 38: User1 is filtering and sorting the results directly on the KWIC view to create subcorpora with include / exclude buttons.

The regular workflow for User1 was to query the search terms in AMC using the software SketchEngine and often using the KWIC view with words frequencies to have a first impression about the context and the results. However, getting the results in the desired context was a challenge since it is difficult to filter out the duplicates caused by regional copies of the newspapers, as a solution she developed a post-query script to filter out the undesired results.

After getting the desired results she exported the output and worked in Microsoft Excel to analyze and possibly visualize the data. However, it took a lot of time because of the iterative workflow of querying, exporting, analyzing and often finding false positives, which resulted in additional queries. As she is interested in the temporal and regional dimensions she worked on making some graphs: line charts for temporal trends and for regions she used bar charts. She wanted to create maps as well, however this would have taken too much time, thus it was not done.

After using our tool for her research question, User1 stated that such a tool would have made her research much faster, easier and the visual presentation

of her findings much more informative. Since the user interface offers include / exclude buttons for building subcorpora, filtering out the undesired results in her queries and saving this set of documents, these functions would have made her workflow much more efficient. She stated that the ability to export results as visualizations from the user interface is a great advantage, because exporting data from SketchEngine (this task had to be done by another colleague, who has the technical expertise) and importing these into Microsoft Excel sometimes caused undesired results and repeating the same workflow was time consuming. In her opinion, SketchEngine offers only basic visualizations, which didn't help her discovering useful information.

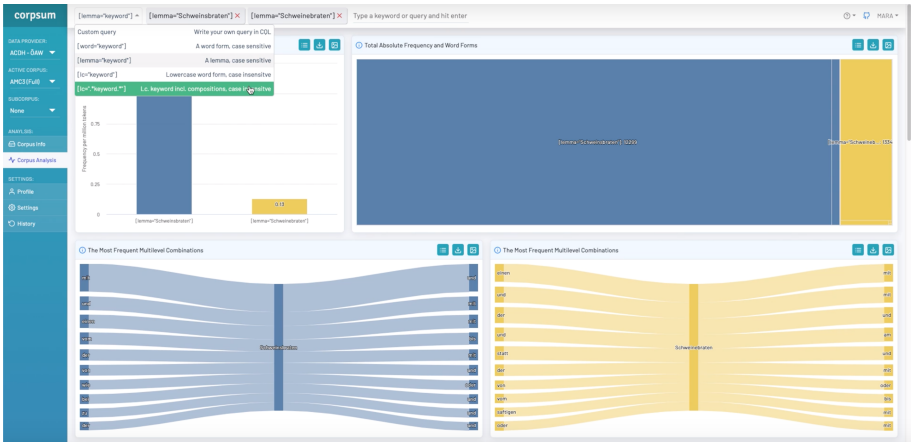

Figure 39: User1 is trying out different query combinations with the help of the query input guide component.

After trying out our tool, User1 states that she likes the overview and the layout of the application a lot, which is very helpful for a first time user. She points out that for the people with less technical expertise it would be a very easy-to-use graphical user interface. In terms of metadata dimensions she states that the dashboard offers all the dimensions she needs.

User1 states that it is much easier to make queries with our tool, because with the query input guide the user does not have to know the notation for the common search types. The possibility to make multiple queries at the same time is a great advantage, which was not possible before with the tools she has used.

According to User1, she would use this tool in her research projects actively since it does all the things that she has to do manually but much more efficiently. In her opinion the corpus linguistics community such as researchers and students would profit from this tool as well since they would get a much better overview, explore the corpus themselves and export their results easily. This speeds up

the hypothesis creation and validation process.

7.2.2 Case Study 2

In this case study we focus on the experience of User2 using our tool for her research. User2 has a master's degree in German Language and is currently conducting her research in a topic related to corpus analysis for gender terms in German. She uses mainly the APA-Database (Austrian Press Agency) at the University of Vienna's online library. Other than this she could not find a helpful online corpus tool and mainly uses Microsoft Excel to clean and analyze the results.

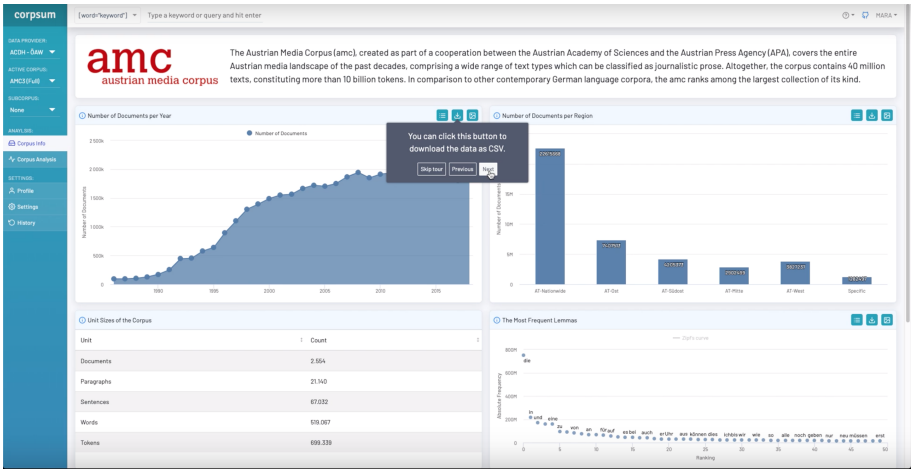

Figure 40: User2 is using the built-in tutorial to learn the functionalities of the tool.

In her view, the APA-Database tool has the advantages of being fast to deliver results and showing the article images next to text. However it crashes a lot and shows a maximum 100 results at a time. One of the main drawbacks of this tool is that it lacks the function to export the results. Thus she has to copy and paste manually from the browser. No metadata filtering or complex query possibilities are offered and it functions basically only as a free text search tool.

User2 describes her project as a comparative linguistics study about discourse on the usage of gender terms in Austria. The first step in her work is to search for articles which have the topic of "gender-equal language", and to look into how the topic has been covered. This step is mainly a qualitative work, which results in defining a list of search terms. Afterwards the quantitative analysis begins, which includes the analysis of frequencies over different dimensions. The focus moves onto the articles which include a "meta discus-

sion" about gender specific language, however this criterion brings a lot of false positives, which should be filtered out.

The relevant metadata dimensions in her research are the media sources and newspaper sections. Until now she has collected around 700 articles and based on these she would like to look deeper into the meta discussion such as comparing discourse in individual articles and clustering these results. She states that the current workflow takes a lot of time and a lot of manual work because there is no single tool she can rely on in terms of querying, exporting, and analyzing data.

The summary of User2's main usage steps with the tool is as follows:

- User2 starts by carefully following the steps of the getting-started guide of our tool. She states that after finishing this tutorial she feels much more comfortable using the tool further and this would be an advantage for the new users of the tool.
- Afterwards she inspects the total statistics of the corpus in terms of media sources and newspaper sections on the "Corpus Info" screen. She comments that the visualizations for these distributions are highly useful compared to the tabular listings which are provided in the tools she has used before.
- As she moves on to the "Corpus Analysis" section, she switches the query selector option to "custom query", because she has some experience with the CQL (Corpus Query Language) and wants to make a query using regular expressions. Since her research question is about "gender-equal language" and the term "gender" can be used in various word combinations in German, she builds a query which would match almost all possible combinations with this term. She points out that the query input area is very helpful for selecting the most commonly used query types with clear descriptions on the side.
- Her query brings interesting results on the temporal dimension and she starts inspecting the years with distinct increases or decreases in relative frequency. Focusing on some time intervals, she dives deeper into the article content using the document viewer. She mentions that having the connection between frequency statistics and the KWIC view increases the efficiency of her research workflow.
- While inspecting the word-forms in tree-map view, she realizes that the results have some unwanted hits such as the word "Gendarmerie", which is not relevant to her research question. She uses the filtering and subcorpus creation functions to remove these hits. At this point she mentions the ease of recognizing outliers and filtering out undesired results with this tool, which is a big improvement on the her existing workflow.

- Using the views for media sources and newspaper sections, she defines some focus points such as the newspaper "Der Standard" and the section "Lesebriefe". These data points have higher relative frequency compared to the baseline and should be inspected further. She filters out the results for these criteria and saves the articles as a subcorpus with a desired title. Then she continues with the close reading and further filtering / clustering based on these results. She mentions that the functionalities of our tool allows her to conduct a qualitative discourse analysis much more systematically than in her previous workflow.

User2 points out that an annotation function was something, which her workflow was lacking, and thus creating subcorpora was a difficult task. For her it is highly important to get the full texts after searches and a KWIC view with subcorpus creation buttons would be very helpful. In her work and study community there are a lot of people who have similar context-related questions and similar needs for a software tool.

Figure 41: User2 is analyzing the media sources and newspaper sections on a query with using the zooming and filtering interactions.

User2 mentions having remote access to a corpus tool as a web application as a big advantage since many of her colleagues would be able to work from the University. Exporting the findings as visualizations is an ongoing problem because there is no integrated solution in her workflow for this.

After using our tool User2 mentions that she finds the user interface very well structured and likes the calm colors in the application's theme. She suggests that it would be good if the user interface is offered in different languages, for example in German as well. This is a feature request we add to our list for future improvements.

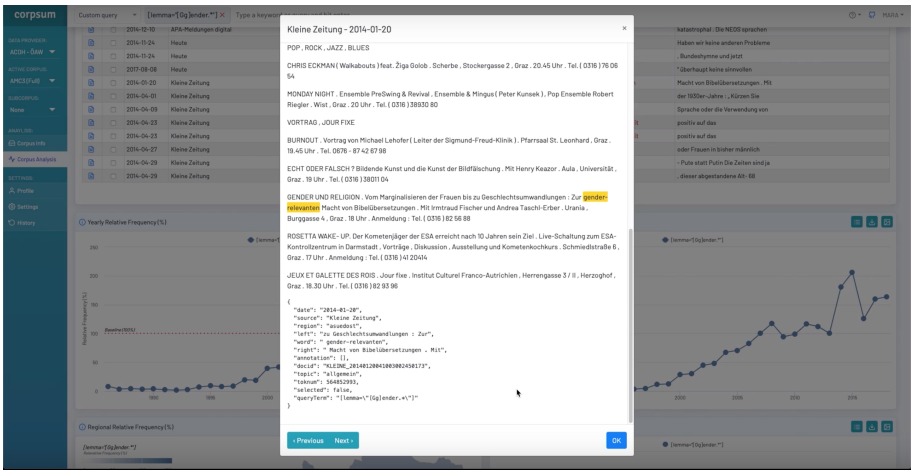

Figure 42: User2 is using the document viewer component, which allows her to easily navigate between documents and inspect the context of the results.

User2 finds the "Corpus Info" screen very informative and states that it gives a great overview for a user who does not know much about the corpus. Moreover she finds the information popups, which are activated by clicking the info button on any visualization component, very useful for the new users of this tool.

In her opinion this tool would help her and many colleagues' research because it makes the subcorpus creation, querying and exporting the results much easier, which is the main issue with her current time consuming workflow. We talk about the possibility of offering our tool to User2 for long time use and we stay in contact in this regard.

8 Lessons Learned

One of the main lessons learned during the project is the importance of having useful real-life data sources, so that the prototypes can build on and visualize these dataset from the start of the project. Having this opportunity provided two main advantages for us: the first one being the consistency between the low- and high-fidelity prototypes in terms of data types, dimensions and components, and the second one being the access to the already established user network around the data provider institute, which curates and maintains the corpora. This allowed us to easily contact researchers who are in close contact with the service providers of AMC and CORPES corpora.

During the prototyping process we learned the important lesson of paying a lot of attention to the low-fidelity prototypes. It is highly valuable to spending a fair amount of time at this stage, because the decisions made at this phase change the course of the further design and development dramatically. We were able to recognize the importance of this step and gained very valuable feedback, which changed our initial design ideas in a positive way as described in the "Design Decisions" section.

Another important lesson has been the importance of surveying the existing tools in detail before starting the interview process. Since many users have already experience with some of the commonly used tools in the given domain, this allows the visualization researchers to ask precise questions about pros and contras of the tools the interviewee has used. This knowledge opens up opportunities to build a deeper and more insightful conversation with the user. In our experience it was essential to build a solid relationship with the interviewees in order to understand their needs and priorities better, and construct a rewarding feedback loop for multiple rounds of interviews and evaluation rounds.

9 Conclusion and Future Work

In this project we focused on designing and developing a visual analysis tool for researchers who work with contemporary media corpora. Since corpus linguistics researchers have significantly different tasks and research questions, we categorized the most important tasks and the functionalities to be implemented in the tool, which would serve the needs of a fairly broad user group.

As described in the "Users" section we defined the common profile of our users as humanities scholars with lower levels of technical expertise and designed our prototypes to be suitable for this target group. Afterwards we evaluated and improved these prototypes based on the feedback we received from our domain experts. As datasets we have used two large contemporary media corpora, Austrian Media Corpus and CORPES, which served our case studies perfectly and allowed us to collaborate with domain experts who work with these corpora.

We see one of the main contributions of this project being a detailed survey of the existing and commonly used corpus analysis tools. This survey has presented the fact that user centered design and evaluation methodologies were almost never used in the design and development of the existing solutions, which may be a possible explanation of the dissatisfaction of the users we interviewed with the existing tools. Based on this insight we decided to apply the design study methodology with multiple evaluation rounds in the process of developing a tool in this domain, which is another main contribution of this project.

Another contribution of our work is the analysis and abstraction of tasks of the corpus linguistics researchers and students as described in the "Tasks" section. We base our analysis and definitions of the tasks on the multi-level typology of abstract visualization tasks by Brehmer and Munzner [14], which perceives the tasks from the perspective of the users in the main questions blocks of "Why?", "How?" and "What?", and bridges the gap between the low-level and the high-level tasks in this domain.

The web-based visual analysis application is a practical outcome and contribution of this project, which allows corpus linguists to easily query corpora and conduct analysis with the help of an interactive multi-query dashboard. As described in the "Evaluation and Case Studies" section our final tool received high approval ratings and very positive feedback from the users, who often stated they would like to use this solution as their main research tool actively.

We offered some of these researchers ongoing access to our tool and after long-term usage we will conduct further interviews to continuously improve the application, which is an evaluation process based on the Multi-dimensional In-depth Long-term Case studies (MILCs) methodology suggested by Shneiderman and Plaisant [30].

Based on our survey of existing tools we think that our tool fills an important need for corpus researchers and has the potential to be a useful analysis tool for a broad range of users. With this motivation and our experience from many iterations of user interviews, we plan to keep working and improving on this software solution.

10 Acknowledgements

Thanks to my supervisor Univ. Prof. Torsten Möller, PhD and my co-supervisor Dr. Thomas Torsney-Weir, MSc for their great support at every stage of my project, as they took countless hours of valuable time and effort to guide me and lead my project to the most successful outcome as possible.

Thanks to two main corpus experts at the Austrian Centre for Digital Humanities and Real Academia Española, Mag. Hannes Pirker and Jordi Porta-Zamorano, PhD, who spent a lot of time to make it possible for me to use the corpus data sources in the best way possible.

Thanks to many colleagues at the Austrian Centre for Digital Humanities, who supported this project with their valuable input and participation in the user interviews.

Thanks to all users who took part in multiple rounds of interviews and gave highly useful feedback for the improvement of the software tool.

This work was partly supported by a Travel Grant of the ELEXIS Project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

[1] BNCweb. http://corpora.lancs.ac.uk/BNCweb/index.html. Accessed: 2019-10-28.

[2] CQPweb - Research Portal | Lancaster University. http://www.research.lancs.ac.uk/portal/en/upmprojects/cqpweb(855e22e7-97fb-4863-9101-9b245729ccbd).html. Accessed: 2019-10-28.

[3] CQPweb Main Page. https://cqpweb.lancs.ac.uk/. Accessed: 2019-10-28.

[4] Getting Started - Voyant Tools Help. https://voyant-tools.org/docs/#!/guide/start. Accessed: 2019-10-29.

[5] LancsBox: Lancaster University corpus toolbox. http://corpora.lancs.ac.uk/lancsbox/index.php. Accessed: 2019-10-28.

[6] MonoConc Image. http://www.athel.com/context.GIF. Accessed: 2019-10-16.

[7] MONOCONC: Text Searching Software. https://www.monoconc.com/. Accessed: 2019-10-28.

[8] NoSketch Engine. https://nlp.fi.muni.cz/trac/noske. Accessed: 2019-10-28.

[9] Wmatrix corpus analysis and comparison tool. http://ucrel.lancs.ac.uk/wmatrix/. Accessed: 2019-10-29.

[10] Anthony, L. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.* (2005), IEEE, pp. 729–737.

[11] Anthony, L. Developing AntConc for a new generation of corpus linguists. In *Proceedings of the Corpus Linguistics Conference (CL 2013)* (2013).

[12] Anthony, L. AntConc, 2019.

[13] Barlow, M. MonoConc Pro Manual, 2003.

[14] Brehmer, M., and Munzner, T. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19* (12 2013), 2376–85.

[15] Brezina, V. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press, Cambridge ; New York, 2018.

[16] Chandler, B. *Longman Mini Concordancer*. Harlow: Longman, 1989.

[17] Davies, M. English Corpora: Most widely used online corpora. https://www.english-corpora.org/faq.asp. Accessed: 2019-10-28.

[18] Davies, M. Mark Davies, Professor of (Corpus) Linguistics, Brigham Young University (BYU). http://davies-linguistics.byu.edu/personal/. Accessed: 2019-10-28.

[19] Hockey, S. *Micro-OCP (OCP Version 2)*. Oxford University Press, 1988.

[20] Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., and Möller, T. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19* (October 2013), 2818–2827.

[21] Johansson, S., L. G., and Goodluck, H. *Manual of Information to Accompany the Lancaster–Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo, 1978.

[22] Johansson, S. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In *Corpus Linguistics: An International Handbook*, no. Bd. -29.2 in Handbooks of Linguistics and Communication Science. Walter de Gruyter, Berlin ; New York, 2009. OCLC: ocn259716120.

[23] Karlsson, F. Origin and history of corpus linguistics - corpus linguistics vis-a-vis other disciplines. In *Corpus Linguistics: An International Handbook*, no. Bd. -29.2 in Handbooks of Linguistics and Communication Science. Walter de Gruyter, Berlin ; New York, 2009. OCLC: ocn259716120.

[24] Kaye, G. A corpus-builder and real time concordance browser for an ibm pc. In *Theory and Practice in Corpus Linguistics*, J. Aarts and W. Meijs, Eds. Amsterdam: Rodopi, 1990.

[25] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. The Sketch Engine: Ten years on. *Lexicography 1* (July 2014), 7–36.

[26] Kilgarriff, A., R. P. S. P., and Tugwell, D. The sketch engine. In *Proceedings of the Eleventh International Congress of Euralex 2004*, G. Williams and S. Vessier, Eds. Bretagne, France: Universite de Bretagne-Sud, 2004.

[27] Kocincová, L., Jakubíček, M., Kovář, V., and Baisa, V. Interactive visualizations of corpus data in sketch engine. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015* (July 2015), 17–22.

[28] Lam, H., Tory, M., and Munzner, T. Bridging from goals to tasks with design study analysis reports. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan 2018), 435–445.

[29] Lüdeling, A., and Kytö, M. Introduction. In *Corpus Linguistics: An International Handbook*, no. Bd. -29.2 in Handbooks of Linguistics and Communication Science. Walter de Gruyter, Berlin ; New York, 2009. OCLC: ocn259716120.

[30] McEnery, T., and Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge ; New York, 2012. OCLC: ocn732967848.

[31] Munzner, T. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15* (Jan. 2010), 921 – 928.

[32] Ransmayr, J., Mörth, K., and Ďurčo, M. Linguistic Variation in the Austrian Media Corpus. Dealing with the Challenges of Large Amounts of Data. *Procedia - Social and Behavioral Sciences 95* (Oct. 2013), 111–115.

[33] Reed, A. *CLOC User Guide*. University of Birmingham, Computer Centre, 1978.

[34] Scott, M. Step-by-step guide to WordSmith. `https://lexically.net/wordsmith/step_by_step_English7/index.html`. Accessed: 2019-10-29.

[35] Scott, M. WordSmith Tools Manual. `https://lexically.net/downloads/version7/HTML/index.html`. Accessed: 2019-10-29.

[36] Scott, M. *WordSmith Tools*. Oxford University Press, 1996.

[37] Sedlmair, M., Meyer, M., and Munzner, T. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (Dec. 2012), 2431–2440.

[38] Shneiderman, B., and Plaisant, C. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (2006), ACM, pp. 1–7.

[39] Sinclair, S. Sgsinclair/Voyant, Oct. 2019.

[40] Teubert, W. My version of corpus linguistics. *International Journal of Corpus Linguistics 10*, 1 (2005), 1–13.

[41] Tory, M., and Möller, T. Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics 10* (Jan. 2004), 72–84.

[42] Tory, M., and Möller, T. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications 25* (Sept. 2005), 8–11.

[43] Tribble, C. Teaching and Language Corpora: Quo Vadis? In *10th Teaching and Language Corpora Conference (TALC)* (Warsaw, 2012).

[44] Valiati, E., Freitas, C., and Pimenta, M. Using multi-dimensional in-depth long-term case studies for information visualization evaluation. In *Proceedings of the BELIV 2008 - BEyond time and errors: novel evaLuation methods for Information Visualization* (Jan. 2008), p. 9.