Towards Unified, Dynamic and Annotation-based Visualisations and Exploration of Annotated Big Data Corpora with the Help of UNIFIED CORPUS EXPLORER

Kevin Bönisch and **Giuseppe Abrami** and **Alexander Mehler**
Goethe University Frankfurt
Frankfurt am Main, Germany

Abstract

The annotation and exploration of large text corpora, both automatic and manual, presents significant challenges across multiple disciplines, including linguistics, digital humanities, biology, and legal science. These challenges are exacerbated by the heterogeneity of processing methods, which complicates corpus visualization, interaction, and integration. To address these issues, we introduce the UNIFIED CORPUS EXPLORER (UCE), a standardized, dockerized, open-source and dynamic Natural Language Processing (NLP) application designed for flexible and scalable corpus navigation. Herein, UCE utilizes the UIMA format for NLP annotations as a standardized input, constructing interfaces and features around those annotations while dynamically adapting to the corpora and their extracted annotations. We evaluate UCE based on a user study and demonstrate its versatility as a corpus explorer based on generative AI.

1 Introduction

The automatic processing and manual annotation of large text corpora poses a general challenge for various projects in different disciplines such as linguistics (e.g. Nguyen et al. (2024); Kyle and Eguchi (2024)) literary science (e.g. Vetulani et al. (2022); Hou and Ma (2023)), digital humanities (e.g. Silvano et al. (2023); Jiménez-Badillo et al. (2020)), biology (e.g. Ayllón-Benítez et al. (2017); Bartley (2022)), chemistry (e.g. Tchechmedjiev et al. (2018); Datta et al. (2019)), biodiversity (e.g. Löffler et al. (2020); Cornwell (2023)) and legal science (e.g. Nazarenko et al. (2018); Kranzlein et al. (2024)), to name a few. In addition to a multitude of processing methods and the underlying representation formats, which are not directly interoperable due to their mutual heterogeneity (Fäth and Chiarcos, 2022), one more challenge arises: the flexible search and interactive visualization of search results extracted from corpora processed by any of these NLP tools. This challenge can be attributed to high computational demands based on corpus size, the absence of general-purpose tools with built-in interaction features, and the need for standardized formats or pipelines.

We address these challenges by means of UNIFIED CORPUS EXPLORER (UCE) (Section 3), which we developed to facilitate corpus navigation and exploration (Section 3.2) using generic, UIMA-based (Ferrucci and Lally, 2004) NLP pipelines. In this context, the "Unstructured Information Management Architecture" (UIMA) ensures data interoperability through its flexible XML Metadata Interchange (XMI) schema-based annotation format. Additionally, UCE is standardized in the sense of being dockerized and reusable (Section 3.1), ensuring broad applicability across various domains (Section 4). The standardized and systematic creation of UCE is accomplished by using the DOCKER UNIFIED UIMA INTERFACE (DUUI – Leonhardt et al. (2023)), a framework for the distributed and scalable processing of NLP tasks (Section 3.3). To evaluate UCE, we conduct a user study (Section 5) that demonstrates its versatility in the context of information retrieval based on parliamentary debate corpus data. Finally, we release UCE as an open-source framework[1] (AGPL-3.0 license) with a public web demonstration[2] and an online screencast[3].

2 Related Work

Visualizing and interrelating textual information in a functional, appealing, dynamic, and interactive way is a challenge in many disciplines and projects. A web-based tool for this is *REDEN ONLINE* (Frontini et al., 2016) which, in the context

[1] https://github.com/texttechnologylab/UCE
[2] http://eval.uce.texttechnologylab.org/
[3] https://www.youtube.com/watch?v=f3kB9pNPjsk

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 522–534

of literary projects, allows the generation and visualization of similarity relations via a web interface using TEI-based editions. Another TEI-based approach is to visualize Shakespeare's plays (Wilhelm et al., 2013). *interHist* (Lyding et al., 2014a) provides explorations of the *PAISÀ corpus* (Lyding et al., 2014b) to support linguistic projects. The corpus can be searched and results can be visualized as charts. A non-browser-based application focusing on visualizing relations within social groups is provided by Bista et al. (2014), who explore emotions to informationally enrich visualizations. Another approach using NLP methods that is similar to UCE, but does not include chatbots is the *DVW* project (Hunziker et al., 2019). It uses MediaWikis to represent the information explored by NLP, to integrate the associated annotations, and to visualize statistics using e.g. maps. Since the underlying preprocessing software has been further developed, this project is no longer usable without adaptations. Another web-based tool is *AnnoPlot* (Fittschen et al., 2024); it allows users to upload annotated data, analyze it, and generate scatter plots, clusterings, and text statistics. This includes embeddings that are reduced by a dimensionality reduction algorithm and used for visualization. With a focus on preprocessing based on NLTK (Bird et al., 2009), *WebNLP* (Burghardt et al., 2014) offers various visualizations (e.g. word clouds) using Voyant (Sinclair and Rockwell, 2016). As an example of LLM-enhanced annotation, there is ITAKE (Song et al., 2024), which uses generative AI to speed up the process of annotating and extracting data from single texts (rather than retrieving information from many texts).

These projects show that the functionalities of comprehensive corpus visualization and exploration (e.g., dynamic visualization of unstructured data, integration of ontological information, and semantic search based on LLM-enhanced chatbots) are not yet sufficiently available. To fill this gap, we present UCE, whose usage scenario can be outlined as follows: a scientist in a specific domain (e.g. chemistry or biology) has the task of gaining a detailed understanding of the content of a large number of texts. To do this, they must be able to skim the texts very quickly on an abstract visual level, as well as to find the smallest semantic units (e.g. at the level of semantic roles or individual arguments). This combination of general corpus-related and detailed text-related information is done by UCE, with the help of LLM-based chatbots.

3 UNIFIED CORPUS EXPLORER

We introduce a novel solution for making UIMA-annotated (Ferrucci and Lally, 2004) corpora tangible through our **UNIFIED CORPUS EXPLORER** (UCE). UCE is a generic interface that, given any corpus and its extracted UIMA-based annotations, makes the underlying data accessible through various features, including semantic search and visualization, while integrating various UIMA-based annotations. UCE imports the necessary files, creates a multi-microservice environment, and adapts to the needs of the corpus and its annotations. Configuration files can be used to customize UCE in terms of appearance (including color schemes and corporate identity), selection of active features, and integration of annotations. It also allows the incorporation of multiple corpora into the same instance.

3.1 Microservices

UCE consists of several dockerized microservices as described below and by Figure 1:

A Corpus-Importer

UCE is based on Corpus-Importer, a Java application that reads UIMA annotated documents from a specified path, along with a corpus configuration JSON file. The importer extracts the raw data and the configured annotations and applies its own post-processing to set up the environment, including **text segmentation**, **database indexing**, **keyword extraction**, and the creation of various **embedding spaces**, before finally storing each processed document in a PostgreSQL database (**B**).

B Relational Database

We chose a relational PostgreSQL database as our primary database because UCE requires a structured and standardized database schema that can be extended as needed. Its compatibility with the `pgvector` extension (Kane, 2021) allows efficient vector operations directly in the database engine. This allows us to store high-dimensional vector embeddings in relational data tables while providing fast vector operations and searches.

C Graph Database

In addition to a relational database (**B**), we use an Apache Jena (Carroll et al., 2004) SPARQL database to incorporate basic semantic searches in the *Resource Description Framework (RDF)* (Miller, 1998) and *Web Ontology Language (OWL)* (Antoniou and van Harmelen, 2004) data

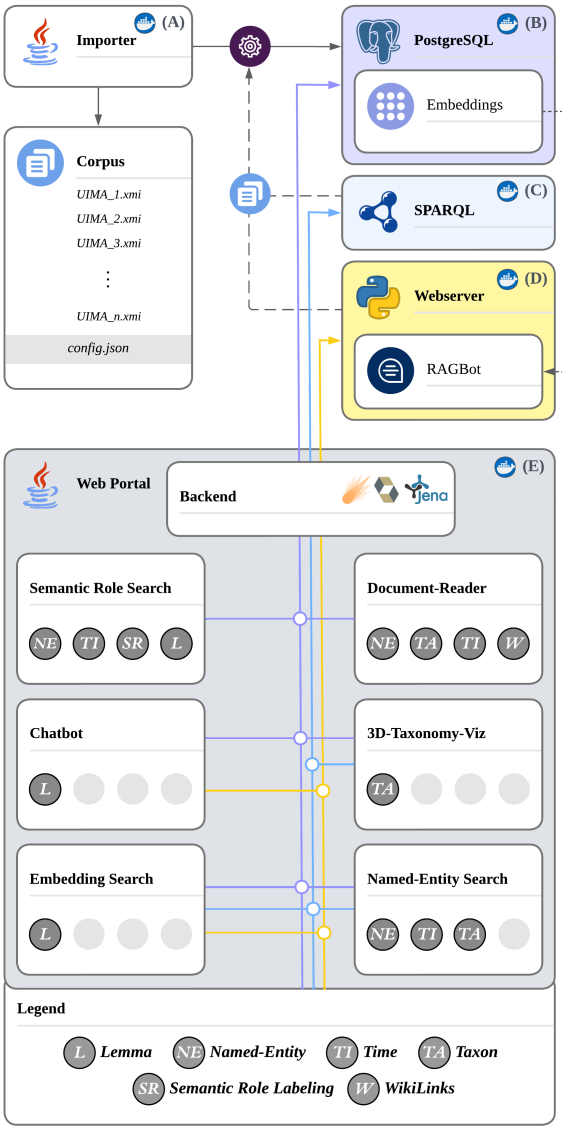

Figure 1: Architectural design of UCE's microservice infrastructure: from top to bottom, the Corpus-Importer **A** reads the UIMA-annotated files of the corpus to be explored along with its configuration and utilizes the graph database **C** to enrich any annotated taxonomies. It then creates high-dimensional embedding spaces on multiple levels through service **D**, which are later, among others, also utilized by the encapsulated *Retrieval-Augmented Generation* pipeline (RAGBot) as outlined in Listing 3. The fully processed document is then stored in the PostgreSQL database **B**. The Web Portal **E** provides user access to the data through a range of features, each of which, indicated by wire connections, leverages different microservices. The wires are colored to indicate their association with the microservice of the same color. Accordingly, the EMBEDDING SEARCH utilizes services **B**, **C**, and **D**, whereas the DOCUMENT READER only utilizes service **B**. Each feature uses a distinct set of annotations, represented by shorthand symbols listed in the legend.

format. This integration enables the incorporation of domain-specific ontologies (e.g., biological taxonomy) into the UCE environment, further enriching its search layers, as shown in Figure 1.

D Python Webserver

Within UCE, we use a Python web service to provide an interface to machine learning and AI models, as these are primarily accessible through Python. In this context, the web server facilitates access to the generation of embedding vectors, dimensionality reduction methods such as *t-SNE* (van der Maaten and Hinton, 2008) and *PCA* (Wold et al., 1987), and the inference of large language models. The web server is accessible via a REST API and is used by the services (**A**) and (**E**).

E UCE Web Portal

The user interacts with UCE and all of its features through a web portal implemented in Java. This service communicates with all other services except (**A**). It provides a variety of search and visualization methods and different ways to interact with the underlying information units (see Section 3.2).

3.2 Features

In this section, we outline the various features showcased in Figure 1, that are built on top of the previously outlined microservices.

1. The DOCUMENT READER (Figure 3) enables users to view every document imported into UCE, with all annotations highlighted within the text and interactive features for further exploration. Using keyword extraction such as RAKE (Rose et al., 2010) and YAKE (Campos et al., 2018), each page is tagged by its keywords. The Document Reader also supports OCR-extracted texts and properties, including the reconstitution of paragraphs, indentations, headers, lines, and blocks.

2. UCE offers a variety of different search layers, besides traditional full-text searches. The NAMED-ENTITY (NE) SEARCH leverages the extracted NE annotations from the text to create its own dedicated search space. Instead of searching through the full text of each document, it concatenates the NE units into a single text. Using GIN indexing, character tri-grams are generated from it, and the given search term is applied using regular expressions to this newly created search space.
The EMBEDDING SEARCH facilitates searching within spaces based on high-dimensional vector

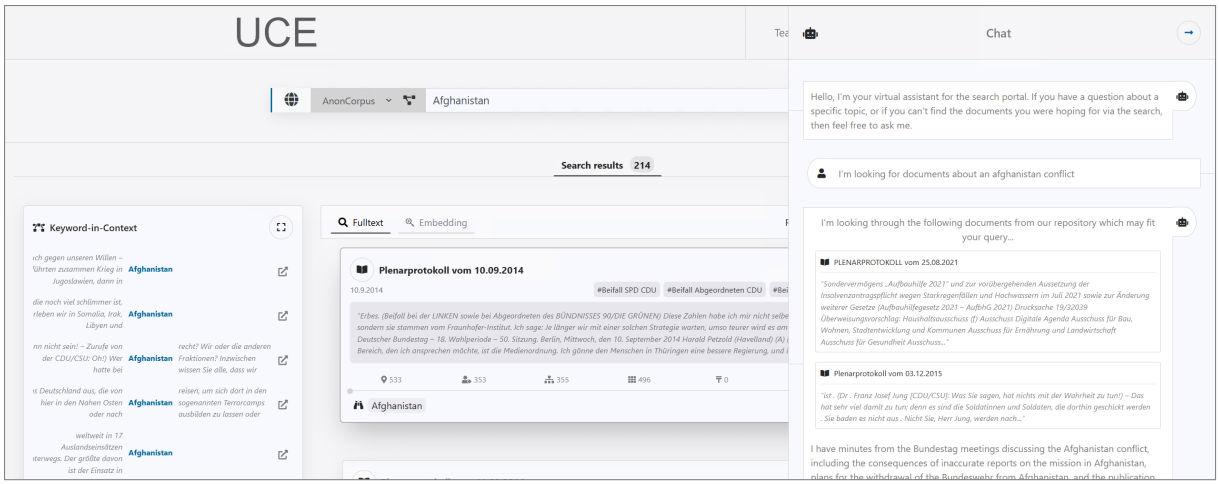

Figure 2: UCE's full-text search using the search term "Afghanistan". Left: a keyword-in-context view is shown, highlighting the occurrences of the search hits in each document. Middle: found documents are listed alongside annotations, including the number of "#"-keywords in the document generated through keyword extraction. Displayed are the counts of locations, people, organizations, and miscellaneous entities (NER), along with taxonomies and time. Right side: the user interacted with UCE's custom chatbot, which suggested additional documents.

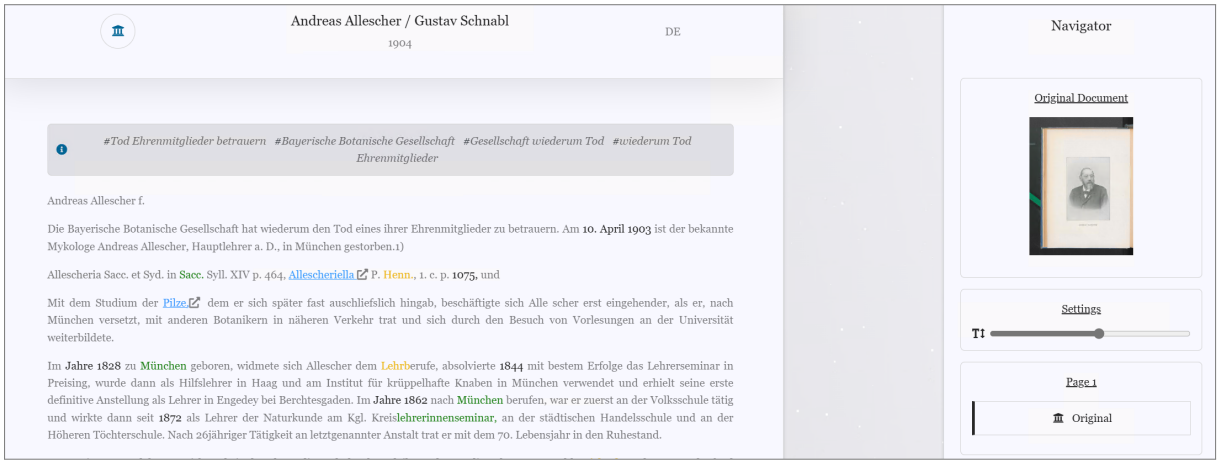

Figure 3: Showcasing an excerpt of the DOCUMENT READER: On the left, the formatted and annotated text is displayed to the user, while the right side provides optional settings and links to the original document—such as, in this case, a link to the original PDF file from which this text was OCR-extracted.

representations of the text at both the paragraph and document level. Upon import, each document is segmented into fixed-length chunks. For each chunk, an embedding is generated using a `SentenceTransformer` (Reimers and Gurevych, 2019, 2020) from the *MixedBread* (Lee et al., 2024) transformer family. However, owing to the open-source microservice architecture, the specific model can be dynamically interchanged. Upon user input, the search query is embedded using service **D**, which initiates a similarity search in service **B** and ultimately outputs the contextually closest text to the user. This search is particularly applicable to users that are unsure of a specific keyword and may only partially (or even incorrectly) recall a segment. Finally, UCE utilizes Semantic Role Labeling (SRL) (Gildea and Jurafsky, 2002) annotations. By doing so, the SEMANTIC ROLE (SR) SEARCH enables users to specifically model their search queries in response to the question "*Who did what to whom, and where?*".

3. In addition to various search layers, UCE provides access to the underlying corpora through its chatbot. For this, pre-trained LLMs have been fine-tuned for instruction-following tasks to support contextual, multi-turn question-answer interactions with users (Taori et al., 2023; Iyer et al., 2023; Team et al., 2024). While these models are capable of handling dialogues, their knowledge is inherently limited by their training data (Cheng et al.,

2024). In an environment like UCE, where the underlying information units are both unknown and dynamic, we opted for an approach that combines the instruction-following behavior of LLMs with mechanisms to fetch additional, domain-specific context at runtime, utilizing *Retrieval-Augmented Generation* (RAG) techniques (Gao et al., 2024; Lewis et al., 2021). A known challenge in RAG is retrieving accurate context for queries that lack relevant information to search for (Anantha et al., 2023). Specifically, a user might reference a previous dialogue turn ("*You've said earlier that...*"), request clarification on earlier context ("*What does that mean?*"), or provide a query lacking any appropriate contextual information ("*Can you print out the book in PDF format?*"). In such cases, retrieving context from within the corpus is unnecessary and may lead to confusion for the LLM when prompted, while also hindering natural multi-turn conversations. To address this, we employ our *Chat Context Classification (CCC)-BERT*, a BERT model enhanced with a classification head (composed of a feedforward layer) that has been fine-tuned to determine whether a user query requires additional context (refere to Figure 7 in the appendix for a more detailed outline of this). CCC-BERT was fine-tuned on 35 000 synthetic multi-turn chats generated with OpenAI's GPT-3.5 Turbo, following the principle of model alignment through self-instruction as proposed by Wang et al. (2023). This enables us to manage the prompting more effectively, as we can either inject additional context or instruct the LLM to consider prior context and dialogue turns explicitly within the prompt. The resulting chatbot has access to all documents and can answer questions about them, follow instructions (e.g., *summarize*), and suggest relevant documents and their snippets to the user.

4. UCE provides several visualization features, including 3D-TAXONOMY-VIZ, which visualizes a document's annotated geographic entities (Figure 4). It instantiates a traversable 3D globe on which geographic entities are projected according to their latitude and longitude properties, with common occurrences visually represented as stacked columns. Each entity is searchable and interactive, allowing users to zoom, rotate, or click on them to get more information about their occurrences and cross-reference links.

UCE's data is generated in a standardized DUUI (Leonhardt et al., 2023) process that can handle large corpora (see Section 3.3). Any UIMA-annotated corpus, regardless of its source, can be imported into UCE. Table 1 (Appendix) shows a list of all annotations currently supported by UCE.

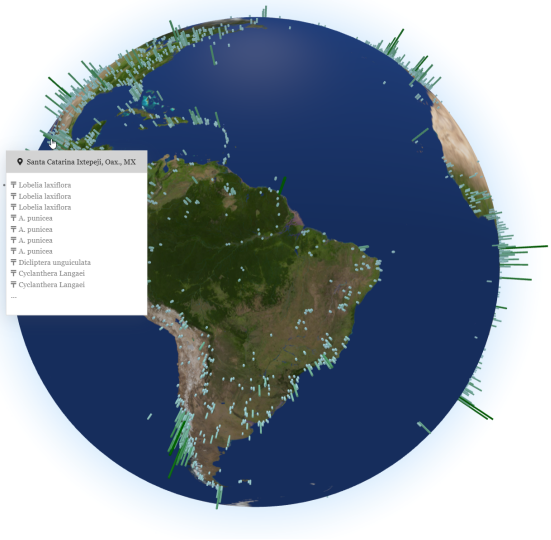

Figure 4: 3D-TAXONOMY VIZ on biodiversity taxonomies, annotated by their GBIF occurrences and projected onto a 3D world representation using `three.js` (Danchilla, 2012) and `globe.gl` (Asturiano, 2019).

3.3 DUUI Pipeline

To use UCE, the generation, annotation, and aggregation of the required data should be done in a standardized and automated way that allows even large corpora (e.g. Abrami et al. (2024)) to be processed efficiently. To this end, UCE uses DUUI (Leonhardt et al., 2023), a software framework for distributed processing of heterogeneous UIMA-based NLP tools, based on horizontal (*multiple nodes*) and vertical (*multiple instances*) scaling. DUUI is preferred because a) it currently provides the best performance compared to existing alternatives, b) it is based on UIMA as its annotation format, and c) it can be easily extended with custom modules based on microservices (Abrami and Mehler, 2024). Using DUUI involves defining a pipeline by selecting the desired annotations to be applied. Then, the required documents (pre-processed or not) are processed by the pipeline, ensuring high scalability within a cluster (Abrami et al., 2025). Each document is annotated by individual annotation steps defined as DUUI components. Each pipeline ends with different evaluator or writer routines that serialize the preprocessing results.

4 Use Cases

Research on Biodiversity

The loss of biodiversity is a central research topic in the geosciences and the life sciences (Hallmann et al., 2017). Data on the spatial distribution of species, their interactions and adaptation to changes are essential to identify environmental processes. Older literature provides a largely unused dataset for this purpose, which allows us to look far back into the past. To exploit this source, search portals have been developed (Pachzelt et al., 2021) that use OCR and NLP (Lücking et al., 2021) to extract biodiversity data from this literature. However, these portals are neither standardized nor designed for cross-domain applicability. This is where UCE can make a contribution, as it can process OCR-extracted data and be tailored to the specifics of the underlying corpora. Using RAG, it can help discover unknown processes and relationships between entities and identify trends. As UCE users indicate a need to uncover the underlying NLP annotations to get an overview of what has been annotated, the addition of wiki-related functionality as provided by Wikidition (Mehler et al., 2016) is required. That is, users should be able to switch from corpus searches to wiki pages that summarize all the information UCE has about the respective item (species, interaction type, location, time, etc.).

Political Science

Parliamentary documents (minutes, votes, legislative initiatives, etc.) generate a huge amount of information that is contextualized by background information, such as social, economic, geopolitical and cultural events (Abrami et al., 2022). Although there is much work on providing annotated parliamentary corpora (Abrami et al., 2024) and applications for analyzing and searching plenary documents (Bönisch et al., 2023), UCE can contribute by making searchable additional information sources. UCE allows each parliament's documents to be addressed individually, in alternative combinations, all at once, or in the context of additional corpora (e.g., newspapers). This allows for a deeper look at political issues and events, with a broad geographic filter to make participation easier.

Educational Sciences

Critical online learning is a process in which students explore multiple documents by searching the Web for relevant information, evaluating it, and reflecting on their responses based on that evaluation (Zlatkin-Troitschanskaia et al., 2021). This process can be studied by looking at which segments of documents students actually look at, at what times, in what order, with what intensity, and with what intermediate results (e.g., in the form of notes). This provides two research perspectives: an in-depth analysis of the behavior of individual students and a broad analysis of the behavior of groups of them. UCE provides opportunities to support both of these perspectives: it makes all documents retrievable, allowing researchers to search for consulted texts that manifest certain linguistic patterns (e.g., semantic roles, certain lexical items) or metadata (e.g., authorship, genre). This allows textual data to be retrieved in collections that would otherwise be difficult to identify. Of interest, e.g., are AI chatbot texts that are made accessible by subjecting all consulted documents to an NLP test (Verma et al., 2024) to check whether they are artificial. This annotation can then be used as a search criterion to identify such texts used by students.

New Data Spaces in the Social Sciences

So far, we considered three application scenarios based on text corpora. It is not only in the social sciences that we need to analyze multimedia content, e.g. from online social networks (OSN), microblogging platforms or video platforms. Especially in the context of surveys, it is increasingly common for people to be reluctant to participate or to give socially desirable answers. Therefore, we need alternative data sources that can be used to study people's attitudes, such as the data traces they leave in OSNs. UCE also offers development opportunities in this respect, namely by extending its corpus concept to relational, non-primary textual data, but also to video and image data, so that images, videos and behavioral data can be searched and queried with RAG in addition to text.

5 Evaluation

For evaluating UCE, the use case from the area of parliamentary debates was chosen by selecting a sample from GERPARCOR (Abrami et al., 2024), the largest collection of transcripts of German plenary debates annotated with spaCy and sentiments. In order to perform evaluations in UCE for the following scenarios, the existing annotations were extended by Semantic Role Labeling using a DUUI-Component based on (Konca et al.,

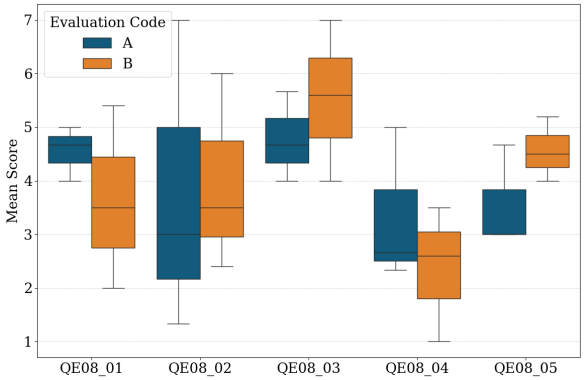

Figure 5: Boxplot of the UMUX questions (Finstad, 2010). The y-axis represents a 7-step Likert scale, where 1 corresponds to *"Absolutely disagree"* and 7 corresponds to *"Absolutely agree"*. A and B denote the split evaluator groups. The questions on the x-axis are to be read from left to right: 1: *"The functionality of UCE meets the requirements of the task at hand."*, 2: *"Using UCE is a frustrating experience."*, 3: *"UCE is easy to use."*, 4: *"I had to spend too much time correcting issues in UCE."* 5: *"The loading times were short."*

2024). The evaluation aims to assess the usability and user experience of the system and its effectiveness for research within the corpus. We subsequently formulated the following research question for the evaluators using UCE: *"What topics were discussed, decisions made, and resolutions passed in the German Parliament concerning Afghanistan during the 18th and 19th legislative period? Specifically, what votes were held, and what were the outcomes?"* Evaluators had 30 minutes to conduct their research and document findings in the questionnaire. We conducted **A/B** tests, dividing participants into two groups, each using different UCE features. **Group A** accessed the SR-SEARCH and NE-SEARCH, while **Group B** used the EMBEDDING SEARCH and the CHATBOT. Both groups could use the full-text search and the DOCUMENT READER for document review. After documenting findings, evaluators answered *Usability Metric for User Experience* (UMUX) questions (Finstad, 2010), with results shown in Figure 5. Participants also provided feedback on desired features, feature usage, and their likelihood of using UCE again. The evaluation was conducted with 15 evaluators.

5.1 Results

The results show that most evaluators found UCE acceptable for the task, calling it easy to use with tolerable load times and no major errors (Figure 5). However, a significant number reported frustration, mainly due to insufficient explanations of tools like SR-SEARCH, NE-SEARCH, and EMBEDDING-SEARCH, leading to low usage of these features (see also Figure 6 in the appendix). Evaluators unfamiliar with NLP tools felt overwhelmed and often reverted to full-text search, reducing their overall experience quality. Additionally, many evaluators requested more concise information representation. In the specific case of German Parliament minutes, visualizing elements such as comments, polls, speeches, and agenda items emerged as a primary expectation. The information presented, especially in the DOCUMENT READER, was often seen as overwhelming. Conversely, the CHATBOT was well-received for its clarity, ease of use without prior explanation, and helpfulness in providing relevant information, making it the most positively rated feature. In general, group B gave more favorable feedback, likely due to more user-friendly tools like the CHATBOT and EMBEDDING SEARCH. Finally, 86% of the evaluators indicated that they would use UCE again for the given task.

6 Conclusion

We introduced the UNIFIED CORPUS EXPLORER (UCE), a generic NLP system for exploring UIMA-annotated corpora. UCE unifies the heterogeneous NLP landscape by providing a standardized framework that facilitates the collection, annotation, extraction, and visualization of large corpora in a customizable pipeline. Its microservices-based architecture, implemented in Docker containers, integrates multiple technologies and supports a wide range of annotations. This makes UCE adaptable to various domains and research areas. Our evaluation shows that UCE provides a platform for addressing research questions using large corpora, but also identifies areas for improvement. Future work includes improved filtering and reading capabilities, and better alignment of document structures with corpus specifics. Scalability for large datasets requires stress testing of UCE, while systematic evaluations of its tools (e.g. *CCC-BERT*) are needed to assess the impact of UCE on research projects. As its CHATBOT has proven to be a reliable tool, generative AI offers opportunities to improve UCE. From a DH perspective, integrating Wikidition technologies is also a future prospect. Finally, explanatory and onboarding materials are needed to address usability issues identified in the evaluation.

Acknowledgements

We gratefully acknowledge the financial support provided by the German Research Foundation (DFG) for the projects "Critical Online Reasoning in Higher Education" (FOR 5404 (DFG: 462702138)), for CRC "Negation in Language and Beyond" (SFB 1629 NegLaB (DFG: 509468465)), for "New Data Spaces for the Social Sciences" (SPP 2431) - "ENTAILab - Forschungsinfrastruktur und Innovationslabor" (DFG: 539634240) as well as for the project "Ausbau und Konsolidierung des Fachinformationsdienstes Biodiversitätsforschung[4] (BIOfid)" (DFG: 326061700).

Ethical Aspects

This work has been developed with the ACL Code of Ethics in consideration. With our contribution, we would like to provide an innovation in the systematic, dynamic and annotation-based visualisation of large text corpora. Therefore, due to the subject matter, our contribution does not entail any ethical issues. Regardless of this, we cannot prevent, in the long run, texts that are hurtful or disturbing, or even legally prohibited, from being processed with our application. The authors are aware of this situation, but we also respect free research.

References

Giuseppe Abrami, Mevlüt Bagci, and Alexander Mehler. 2024. German parliamentary corpus (GerParCor) reloaded. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7707–7716, Torino, Italy. ELRA and ICCL.

Giuseppe Abrami, Mevlüt Bagci, Leon Hammerla, and Alexander Mehler. 2022. German Parliamentary Corpus (GerParCor). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1900–1906, Marseille, France. European Language Resources Association.

Giuseppe Abrami, Markos Genios, Filip Fitzermann, Daniel Baumartz, and Alexander Mehler. 2025. Docker Unified UIMA Interface: New perspectives for NLP on big data. *SoftwareX*, 29:102033.

Giuseppe Abrami and Alexander Mehler. 2024. Efficient, uniform and scalable parallel NLP preprocessing with DUUI: Perspectives and Best Practice for the Digital Humanities. In *Digital Humanities Conference 2024 - Book of Abstracts (DH 2024)*, DH, pages 15–18. Zenodo.

[4]Expansion and consolidation of the specialized information service for biodiversity research

Raviteja Anantha, Tharun Bethi, Danil Vodianik, and Srinivas Chappidi. 2023. Context tuning for retrieval augmented generation. *Preprint*, arXiv:2312.05708.

Grigoris Antoniou and Frank van Harmelen. 2004. *Web Ontology Language: OWL*, pages 67–92. Springer Berlin Heidelberg, Berlin, Heidelberg.

Vasco Asturiano. 2019. Globe.gl.

Aarón Ayllón-Benítez, Patricia Thébault, Jesualdo Tomás Fernández-Breis, Manuel Quesada-Martínez, Fleur Mougin, and Romain Bourqui. 2017. Deciphering gene sets annotations with ontology based visualization. In *2017 21st International Conference Information Visualisation (IV)*, pages 170–175.

Bryan A. Bartley. 2022. Tyto: A python tool enabling better annotation practices for synthetic biology data-sharing. *ACS Synthetic Biology*, 11(3):1373–1376. PMID: 35226470.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with python*. O'Reilly Media Inc.

Sanat Kumar Bista, Surya Nepal, and Cécile Paris. 2014. Multifaceted visualisation of annotated social media data. In *2014 IEEE International Congress on Big Data*, pages 699–706.

Kevin Bönisch, Giuseppe Abrami, Sabine Wehnert, and Alexander Mehler. 2023. Bundestags-Mine: Natural Language Processing for Extracting Key Information from Government Documents. In *Legal Knowledge and Information Systems*. IOS Press.

Manuel Burghardt, Julian Pörsch, Bianca Tirlea, and Christian Wolff. 2014. Webnlp – an integrated web-interface for python nltk and voyant. In *Proceedings of the 12th edition of the KONVENS conference*, page KONVENS 2014.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval*, pages 806–810, Cham. Springer International Publishing.

Jeremy J. Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. 2004. Jena: implementing the semantic web recommendations. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, WWW Alt. '04, page 74–83, New York, NY, USA. Association for Computing Machinery.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *ArXiv*, abs/2403.12958.

Peter Cornwell. 2023. Progress with repository-based annotation infrastructure for biodiversity applications. *Biodiversity Information Science and Standards*, 7:e112707.

Brian Danchilla. 2012. *Three.js Framework*, pages 173–203. Apress, Berkeley, CA.

Surabhi Datta, Elmer V. Bernstam, and Kirk Roberts. 2019. A frame semantic overview of nlp-based information extraction for cancer-related ehr notes. *Journal of Biomedical Informatics*, 100:103301.

Christian Fäth and Christian Chiarcos. 2022. Spicy salmon: Converting between 50+ annotation formats with fintan, pepper, salt and powla. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 61–68, Marseille, France. European Language Resources Association.

David Ferrucci and Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Kraig Finstad. 2010. The usability metric for user experience. *Interacting with Computers*, 22(5):323–327. Modelling user experience - An agenda for research and practice.

Elisabeth Fittschen, Tim Fischer, Daniel Brühl, Julia Spahr, Yuliia Lysa, and Phuoc Thang Le. 2024. AnnoPlot: Interactive visualizations of text annotations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 106–114, St. Julians, Malta. Association for Computational Linguistics.

Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. 2016. REDEN ONLINE: disambiguation, linking and visualisation of references in TEI digital editions. In *11th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2016, Krakow, Poland, July 11-16, 2016, Conference Abstracts*, pages 193–197. Alliance of Digital Humanities Organizations (ADHO).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Caspar A. Hallmann, Martin Sorg, Eelke Jongejans, Henk Siepel, Nick Hofland, Heinz Schwan, Werner Stenmans, Andreas Müller, Hubert Sumser, Thomas Hörren, Dave Goulson, and Hans de Kroon. 2017. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLOS ONE*, 12(10):1–21.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2015. spaCy: Industrial-strength Natural Language Processing in Python.

Yanwen Hou and Yurong Ma. 2023. Cross-lingual sentiment analysis in literary translation: A case study of the novel crystal boys. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6.

Alex Hunziker, Hasanagha Mammadov, Wahed Hemati, and Alexander Mehler. 2019. Corpus2wiki: A mediawiki-based tool for automatically generating wikiditions in digital humanities. In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*, pages 143–149. Gesellschaft für Informatik e.V., Bonn.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *Preprint*, arXiv:2212.12017.

Diego Jiménez-Badillo, Patricia Murrieta-Flores, Bruno Martins, Ian Gregory, Mariana Favila-Vázquez, and Raquel Liceras-Garrido. 2020. Developing geographically oriented nlp approaches to sixteenth–century historical documents: Digging into early colonial mexico. *Digital Humanities Quarterly*, 14(4).

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Pearson. Online manuscript released August 20, 2024.

Andrew Kane. 2021. pgvector.

Maxim Konca, Andy Luecking, and Alexander Mehler. 2024. German SRL: Corpus construction and model training. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7717–7727, Torino, Italia. ELRA and ICCL.

Michael Kranzlein, Nathan Schneider, and Kevin Tobia. 2024. CuRIAM: Corpus re interpretation and metalanguage in U.S. Supreme Court opinions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4247–4258, Torino, Italia. ELRA and ICCL.

Kristopher Kyle and Masaki Eguchi. 2024. Evaluating nlp models with written and spoken l2 samples. *Research Methods in Applied Linguistics*, 3(2):100120.

Nicolas Le Guillarme and Wilfried Thuiller. 2022. Taxonerd: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Alexander Leonhardt, Giuseppe Abrami, Daniel Baumartz, and Alexander Mehler. 2023. Unlocking the heterogeneous landscape of big data NLP with DUUI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 385–399, Singapore. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Felicitas Löffler, Nora Abdelmageed, Samira Babalou, Pawandeep Kaur, and Birgitta König-Ries. 2020. Tag me if you can! semantic annotation of biodiversity metadata with the QEMP corpus and the BiodivTagger. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4557–4564, Marseille, France. European Language Resources Association.

Andy Lücking, Christine Driller, Manuel Stoeckel, Giuseppe Abrami, Adrian Pachzelt, and Alexander Mehler. 2021. Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Language Resources and Evaluation*, 56(3):807–855.

Verena Lyding, Lionel Nicolas, and Egon Stemle. 2014a. 'interHist' - an interactive visual interface for corpus exploration. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 635–641, Reykjavik, Iceland. European Language Resources Association (ELRA).

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014b. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden. Association for Computational Linguistics.

Alexander Mehler, Rüdiger Gleim, Tim Vor Der Brück, Wahed Hemati, Tolga Uslu, and Steffen Eger. 2016. Wikidition: Automatic lexiconization and linkification of text corpora. *it-Information Technology*, 58(2):70–79.

Alexander Mehler, Wahed Hemati, Pascal Welke, Maxim Konca, and Tolga Uslu. 2020. Multiple texts as a limiting factor in online learning: Quantifying (dis-)similarities of knowledge networks. *Frontiers in Education*, 5:206.

Eric Miller. 1998. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19.

Adeline Nazarenko, François Levy, and Adam Wyner. 2018. An annotation language for semantic search of legal sources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Li Nguyen, Shiva Taslimipoor, and Zheng Yuan. 2024. What can nlp do for linguistics? towards using grammatical error analysis to document non-standard english features. *Linguistics Vanguard*.

Adrian Pachzelt, Gerwin Kasperek, Andy Lücking, Giuseppe Abrami, and Christine Driller. 2021. Semantic search in legacy biodiversity literature: Integrating data from different data infrastructures. *Biodiversity Information Science and Standards*, 5:e74251.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, chapter 1. John Wiley & Sons, Ltd.

Maria da Purificação Silvano, Evelin Amorim, António Leal, Inês Cantante, Maria de Fátima Henriques da Silva, Alípio Jorge, Ricardo Campos, and Sérgio Sobral Nunes. 2023. Annotation and visualisation of reporting events in textual narratives. In *Proceedings of Text2Story 2023: Sixth Workshop on Narrative Extraction From Texts*.

Stéfan Sinclair and Geoffrey Rockwell. 2016. Voyant tools.

Jiahe Song, Hongxin Ding, Zhiyuan Wang, Yongxin Xu, Yasha Wang, and Junfeng Zhao. 2024. ITAKE: Interactive unstructured text annotation and knowledge extraction system with LLMs and ModelOps. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 326–334, Bangkok, Thailand. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2015. A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 541–547, Lisbon, Portugal. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio, and Clement Jonquet. 2018. Sifr annotator: ontology-based semantic annotation of french biomedical text and clinical notes. *BMC Bioinformatics*, 19(1).

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.

Zygmunt Vetulani, Marta Witkowska, and Marek Kubis. 2022. *NLP Tools for Lexical Structure Studies of the Literary Output of a Writer. Case Study: Literary Works of Tadeusz Boy-Żeleński and Julia Hartwig*, page 259–276. Springer International Publishing.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Thomas Wilhelm, Manuel Burghardt, and Christian Wolff. 2013. "to see or not to see" - an interactive tool for the visualization and analysis of shakespeare plays. In Regina Franken-Wendelstorf, Elisabeth Lindinger, and Jürgen Sieck, editors, *Kultur und Informatik: Visual Worlds & Interactive Spaces*, pages 175–185. Verlag Werner Hülsbusch, Glückstadt.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4212–4227, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Olga Zlatkin-Troitschanskaia, Johannes Hartig, Frank Goldhammer, and Jan Krstev. 2021. Students' online information use and learning progress in higher education–a critical literature review. *Studies in Higher Education*, 46(10):1996–2021.

A Appendix

Annotation	Description	Model
SENTENCE	Divides the documents into their respective sentences.	spaCy (Honnibal et al., 2015)
NAMED-ENTITY	Extracts named entities from a document, categorizing them into four types: organization (ORG), person (PER), location (LOC), and miscellaneous (MISC) (Grishman and Sundheim, 1996).	
LEMMA	Lemmatization reduces inflected words to their root form. Within UCE, searches are enhanced by considering these root forms.	
SEMANTIC ROLE LABELS (SRL)	SRL identifies semantic relations between the lexical constituents of a sentence (Jurafsky and Martin, 2024), assigning labels to words or phrases that indicate their semantic roles, such as agent, goal, or result.	TreeCRF (Zhang et al., 2022)
TIME	Extracts temporal expressions, including time and date formats, from a document, analogous to Named-Entity Recognition tasks.	HeidelTime (Strötgen and Gertz, 2015)
TAXON	The recognition of unambiguous names of biological entities is referred to as a taxon.	TaxoNERD (Le Guillarme and Thuiller, 2022)
WIKILINKS	Maps potential words and phrases to their corresponding Wikidata URLs, facilitating the retrieval and access of additional information.	(Mehler et al., 2020)
OCR	Since much of the literature has yet to be digitized, UCE provides support for corpora containing documents that have undergone Optical Character Recognition (OCR) extraction. These annotations assist in reconstructing the physical layout of the pages within UCE.	(Lücking et al., 2021)

Table 1: The corpus exploration capabilities within UCE not only allow for interaction with the corpus in its raw form but primarily leverage various annotations extracted through the DUUI pipeline to enhance its extensive features. This table enumerates all annotations currently supported within UCE, accompanied by a concise description and a list of models employed to achieve each task. These models are designed to be dynamic and flexible, as UCE is agnostic to the specific methodologies employed in the generation of annotations within DUUI, owing to the standardized UIMA format.

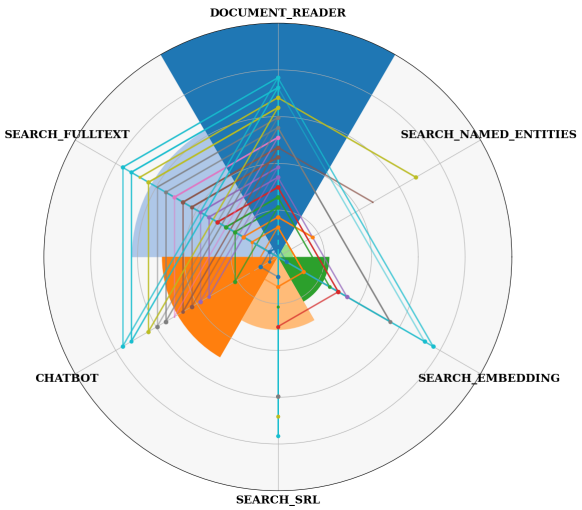

Figure 6: A radar chart visualizing the usage of different features within the evaluation. The six circular bars represent the total accumulated usage of these features, while the individual colored lines depict a comprehensive timeline, illustrating the flow of actions during a single evaluation run within UCE. It can be observed that, among the various searches, the full-text search was utilized the most (primarily because both evaluation groups had access to it), followed by the CHATBOT. Conversely, the NE-SEARCH and EMBEDDING-SEARCH experienced the least usage.

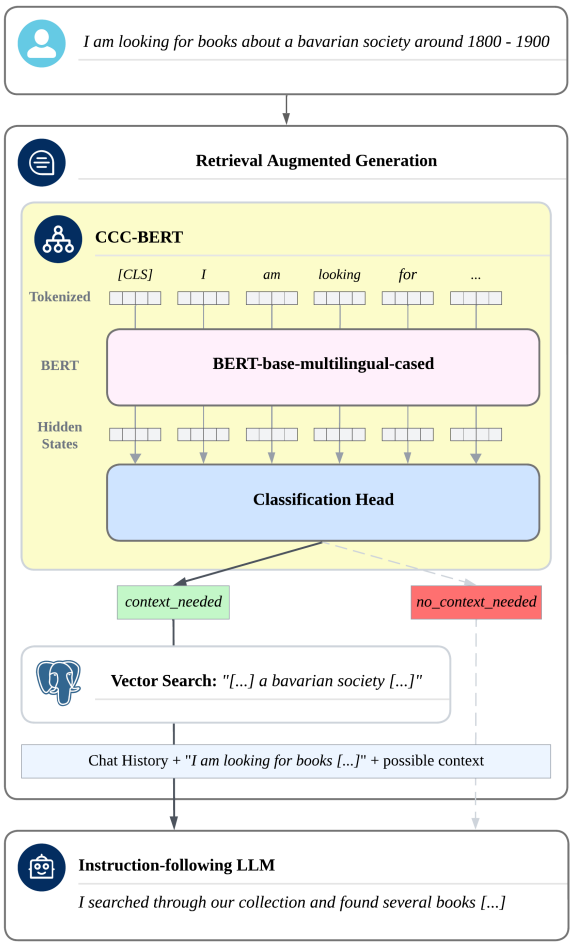

Figure 7: Outlining the workflow of UCE's RAG pipeline from top to bottom: the user inputs a question concerning the underlying corpus. Within the RAG pipeline, we first consult our custom-trained *CCC-BERT* (see Feature 3) model to determine whether additional context is needed. If so, we fetch relevant context by searching through the high-dimensional embedding space of the corpus in service **B** for contextually similar text passages and documents. Finally, we prompt an instruction-following LLM, utilizing either the OpenAI API or `huggingface` models, as specified in the corpus configuration file, to generate the chatbot's response.