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

Anonymous ACL submission

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, 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 (e.g. `https://huggingface.co/`) 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 using generic, UIMA-based (Ferrucci and Lally, 2004) NLP pipelines. 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 PREPROCESSOR[1], a framework for the distributed and scalable processing of NLP tasks using microservices. 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.

2 Related Work

Visualizing and interrelating textual information in a functional, appealing, dynamic, and interactive way is a challenge in various disciplines and projects. A web-based tool to be mentioned in this context is the Genome Context Visualizer (GeCoViz - Botas et al. (2022)), which visualizes the genomic context of prokaryotic genes (see Diesh et al. (2023); Nassar et al. (2023) for related tools). Another example is *REDEN ONLINE* (Frontini et al., 2016) which, in the context 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 the texts of 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 research pur-

[1] Anonymized due to submission policy

poses. 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*[1] project [anonymous]. It uses MediaWikis to represent the information explored by NLP methods, to integrate the associated annotations, and to visualize statistics using e.g. maps. Since the underlying preprocessing software has been further developed in the meantime, this project is no longer usable without adaptations. Another web-based tool is *AnnoPlot* (Fittschen et al., 2024), which allows users to upload annotated data, analyze it, and generate scatter plots, clusterings, and statistics about the texts. 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). In the area of LLM-enhanced annotation, there is the example of ITAKE (Song et al., 2024), which uses generative AI methods to speed up the process of annotating and extracting knowledge from texts. That is, ITAKE focuses on annotating single texts rather than retrieving information from many texts.

The projects mentioned show that the functionalities of a wide-ranging corpus visualization and exploration (such as the dynamic visualization of unstructured data, the integration of ontological information and semantic search based on LLM-enhanced chatbots) are not yet sufficiently available. To close this gap, we developed UCE. The application scenario of UCE can be outlined as follows: a scientist in a specific domain (e.g. chemistry or biology) may have 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 developed as a generic interface that, given any corpus and its extracted UIMA-based annotations, makes the underlying data accessible through various features, including semantic searches and visualization, while integrating the diverse annotations provided through UIMA-based data. Herein, UCE imports the necessary files, establishes an environment consisting of multiple microservices, and dynamically adapts to meet the specific requirements of the corpus and its annotations. Through configuration files, UCE can be customized in terms of thematic appearance (including color schemes and corporate identity), selection of active features, and the integration of different annotations. In this context, UCE can be regarded as the final component of a custom NLP pipeline (see Figure 6), which begins by reading a corpus, annotating it using various tools, and ultimately making the resulting information accessible to the user.

3.1 Microservices

UCE consists of several microservices, each dockerized and utilizing distinct technologies, which is being outlined in the following and through 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 corresponding corpus-configuration JSON file. The importer extracts the raw data and the configured annotations, applying its own post-processing to set up the environment, which includes: **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

As our primary database, we opted for a relational PostgreSQL database, as UCE requires a structured and standardized database scheme that can be extended if necessary. Additionally, its compatibility with the `pgvector` extension (Kane, 2021) enables efficient vector

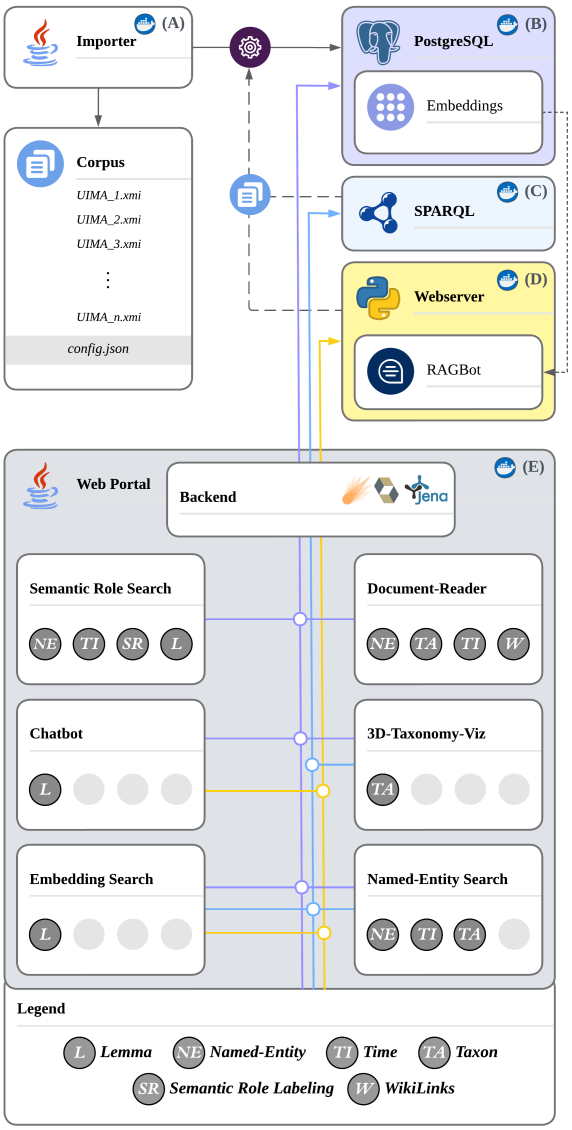

Figure 1: Architectural design of UCE's microservice infrastructure: from top to bottom, the Corpus-Importer **A** reads the UIMA-annotated files of the corpus to be explored along with its configuration and utilizes the graph database **C** to enrich any annotated taxonomies. It then creates high-dimensional 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 III. 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 colored wire connections, leverages different microservices. Each feature uses a distinct set of annotations, represented by shorthand symbols.

operations directly within the database engine. This allows us to store high-dimensional vector embeddings within relational data tables while also enabling fast vector operations and searches.

C Graph Database

In addition to a relational database (**B**), UCE utilizes 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 format. This integration enables the incorporation of domain-specific ontologies (e.g., biological taxonomy, as outlined in Section 4) into the UCE environment, further enriching its search capabilities.

D Python Webserver

Within UCE, we also utilize 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, their 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 REST API and is utilized by 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 for (**B**), providing a variety of search methods, visualization features, and different ways to interact with the underlying information units, as outlined in detail through 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.

I 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.

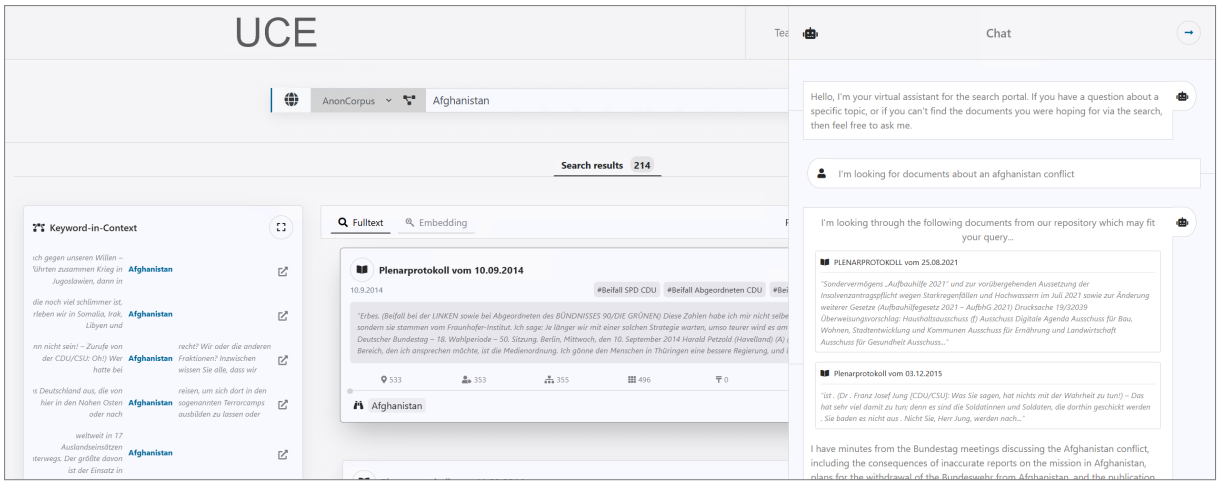

Figure 2: UCE's full-text search using the search term "Afghanistan". Left: a keyword-in-context view is shown, highlighting the occurrences of the search hits in each document. Middle: found documents are listed alongside annotations, including the number of "#"-keywords in the document generated through 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.

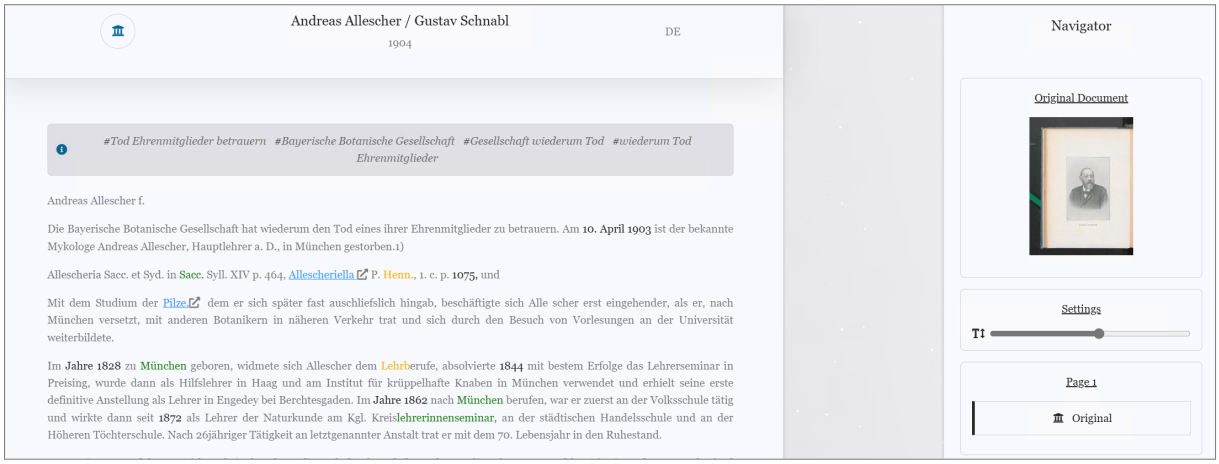

Figure 3: Showcasing an excerpt of the DOCUMENT READER: On the left, the formatted and annotated text is displayed to the user, while the right side provides optional settings and links to the original document—such as, in this case, a link to the original PDF file from which this text was OCR-extracted.

II UCE offers a variety of different search layers, besides traditional fulltext 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 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. 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?*" as outlined by Figure 4.

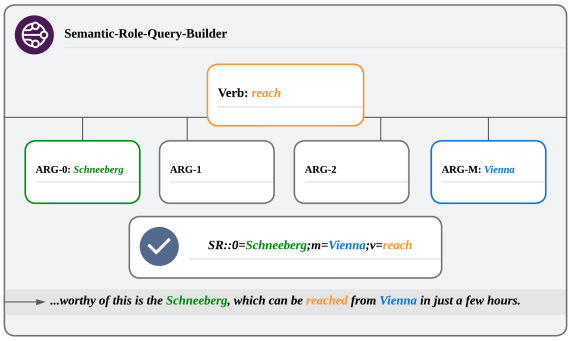

Figure 4: The process of SEMANTIC ROLE SEARCH: The user begins by drag-and-dropping annotations into the respective argument slots *ARG-X* to model the desired query. Within UCE, this information is passed as a query string prefixed by *SR::*, which is then directly parsed into a SQL query. The output consists of document sentences that match the search criteria.

III In addition to various search layers, UCE provides access to the underlying corpora through its custom chatbot. For this, pre-trained LLMs have been successfully 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), as outlined in Figure 7. 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. To address this, we employ our *Chat Context Classification (CCC)-BERT*[2], 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 (see Figure 7). 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.

IV UCE offers a range of visualization features, including 3D-TAXONOMY-VIZ, which provides a three-dimensional visualization of a document's annotated geographical entities (Figure 5). Herein, a traversable three-dimensional globe is instantiated, projecting geographical entities onto it according to their latitude and longitude properties, with common occurrences visually represented as stacked pillars. Every entity is searchable and interactive, allowing users to zoom, rotate, or click on them to retrieve more information about their occurrences and possible cross-reference links.

The generation, annotation and aggregation of the data required for UCE is conducted in a standardised and automated process using PREPROCESSOR [anonymous][3], which also enables the processing of large corpora in a reasonable time. PREPROCESSOR is illustrated in Figure 6: It encapsulates the definition of a pipeline before the required documents – whether or not they have been preprocessed – are processed through the pipeline, ensuring high scalability within a cluster. Finally the UCE-writer uses Service **A** to import the documents enriched with annotationsinto the database system (Service **B** and **C**) provided in UCE-Docker image. A complete list of all annotations currently supported by UCE is provided in Table 1 in the appendix.

[2]anonymous
[3]Published at EMNLP Findings 2023

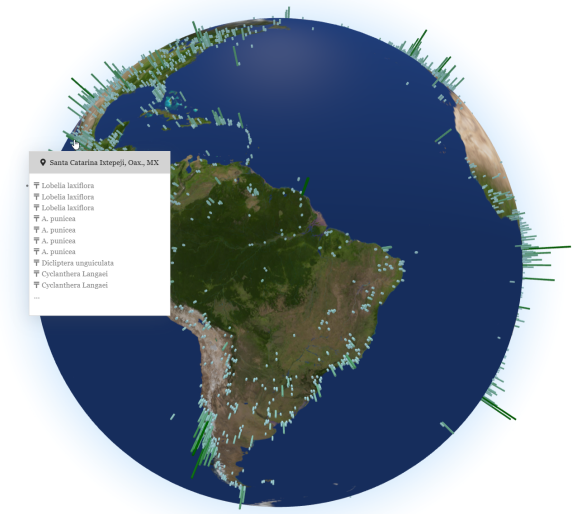

Figure 5: Demonstrating 3D-TAXONOMY VIZ on biodiversity taxonomies (Section 4), annotated by their GBIF occurrences and projected onto a 3D world representation using `three.js` (Danchilla, 2012) and `globe.gl` (Asturiano, 2019).

4 Use Cases

We consider two application areas for the use of UCE: biology and political science.

Biodiversity

In the field of biodiversity, the loss of biodiversity in the Anthropocene is a fundamental research topic in both the geosciences and the life sciences (Hallmann et al., 2017). In this context, data on the distribution of species, their occurrence and their adaptation to environmental changes are crucial for identifying emerging developments and threats. However, legacy literature remains frequently neglected, primarily due to its existence in printed form, limiting accessibility for comprehensive analysis. To address this challenge, semantic search portals have been developed [anonymous][4], incorporating NLP tools to analyze legacy biodiversity data. However, these portals are neither standardized nor designed for cross-domain applicability. In this context, UCE can contribute significantly, as it is capable of processing OCR-extracted data and can be tailored to fit the specific characteristics of the underlying corpus and its annotations. By

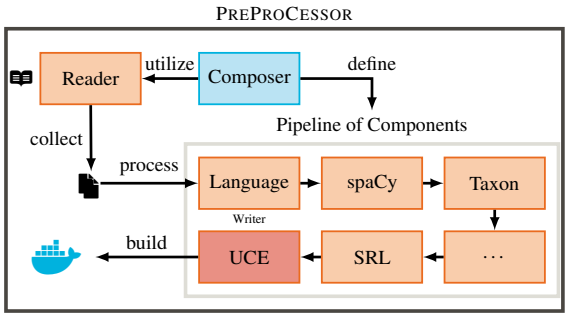

Figure 6: The NLP-pipeline implemented by PREPROCESSOR. In the context of UCE, the most important unit is the UCE-Component defined as a writer. It collects all annotated documents processed with PREPROCESSOR, executes the functions of UCE (see Section 3) and generates a Docker image to host UCE.

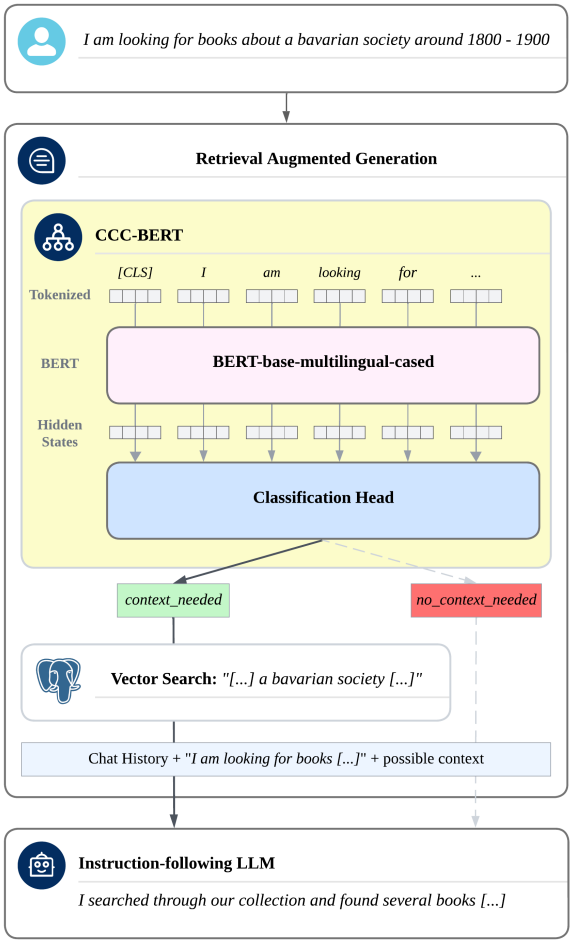

Figure 7: Outlining the workflow of UCE's RAG pipeline from top to bottom: the user inputs a question concerning the underlying corpus. Within the RAG pipeline, we first consult our custom-trained *CCC-BERT* (see Feature III) model to determine whether additional context is needed. If so, we fetch relevant 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.

[4]Published at BISS, 2021

using RAG, it is possible to detect unknown developments, to uncover overlooked connections and relationships between species and locations, and possibly to recognize trends. This is achieved by the user interacting intensively with the underlying corpora, by increasingly refining their semantic searches or, conversely, by expanding them to include ever larger amounts of text.

Parliamentary documents

The wide range of parliamentary documents, be they protocols, inquiries, votes, legislative initiatives or amendments, generate a vast amount of information that is hard to be adequately conveyed in fast-moving times (c.f. [anonymous][5]). In addition, prior knowledge or additional information (social, economic, geopolitical, and cultural events, among other things) places the topics, discussions and facts in a particular context. Regardless of the fact that there is already much work on making heterogeneous parliamentary corpora available and providing annotations for them (c.f. [anonymous][6]), as well as a feature-rich application for analysing and searching plenary protocols [anonymous][7], it makes sense to adapt such socially and scientifically diverse information through UCE in order to reflect its contextual diversity. Each parliament can then be addressed individually, in alternative combinations or all at once. Through unified annotation and visualization, a deeper but also more focused look at current topics would then be facilitated, with a broad geographical filter to simplify the possibility of participation for everyone.

5 Evaluation

For the evaluation of UCE, the use case from the research area of parliamentary debates was chosen by selecting a sample from AnonCorpus [1] [anonymous][6], the largest German-language collection of protocols of plenary debates, which were already annotated with spaCy as well as sentiment annotations. In order to perform evaluations in UCE for the following scenarios, the existing annotations were extended by Semantic Role Labeling using a PreProcessor-Component based on [anonymous][6], using PreProcessor as well. The primary objective of this evaluation is to assess the usability and user experience of the system, and secondly, to evaluate its effectiveness for conducting research within the underlying corpus. We subsequently formulated the following research question for the evaluators to address using UCE: *"What topics were discussed, decisions made, and resolutions passed in the German Parliament concerning*

[5]Published at LREC 2020
[6]Published at LREC-COLING 2024
[7]Published at JURIX 2023

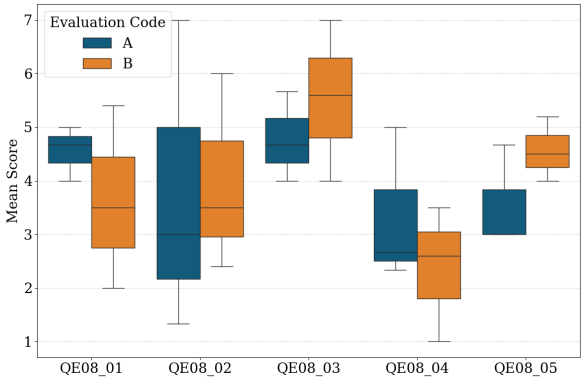

Figure 8: Results of the UMUX questions (Finstad, 2010) as a boxplot. The y-axis represents a 7-step Likert scale, where 1 corresponds to *"Absolutely disagree"* and 7 corresponds to *"Absolutely agree"*. The questions on the x-axis are to be read as follows from left to right: 1: "The functionality 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."

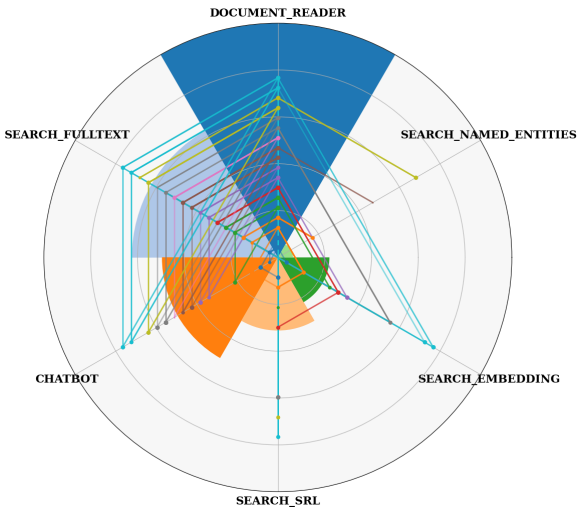

Figure 9: A radar chart visualizing the usage of different features within the evaluation. The six circular bars represent the total accumulated usage of these features, while the individual colored lines depict a comprehensive timeline, illustrating the flow of actions during a single evaluation run 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.

Afghanistan during the 18th and 19th legislative period? Specifically, what votes were held, and what were the outcomes?" The evaluators were allocated a maximum of 30 minutes to conduct their research before documenting their findings in the questionnaire. Unbeknownst to the evaluators, we conducted **A/B** tests by dividing the pool of participants into two groups, with each group using a different set of UCE's features. **Group A** was given exclusive access to the SR-SEARCH and NE-SEARCH, while **Group B** had access to the EMBEDDING SEARCH and the CHATBOT. Both groups had access to the standard full-text search and the DOCUMENT READER for reviewing all documents. This approach allowed us to assess UCE comprehensively in terms of user experience and usability, while also critically evaluating the individual features for their usefulness and effectiveness in facilitating the achievement of the research objectives. After the evaluators had documented their research findings, they were presented with questions regarding the user experience of UCE, following the *Usability Metric for User Experience* (UMUX) (Finstad, 2010). The results of this questionnaire are displayed in Figure 8. We also recorded the frequency of interactions with each feature and illustrate their popularity in Figure 9. Finally, participants were also given the opportunity to provide open-ended feedback on features they wished were available, which features they used most and least, and whether they would choose to use UCE again in a similar context. The evaluation was conducted with 15 evaluators.

5.1 Results

The results indicate that the majority of evaluators found UCE to be an acceptable tool for the task, describing it as easy to use, with tolerable loading times, and no significant errors encountered (Figure 8). In contrast, a more significant portion of evaluators reported frustration while using UCE. Upon examining the open-ended feedback, this frustration may be primarily attributed to insufficient explanations of key tools such as SR-SEARCH, NE-SEARCH, and EMBEDDING-SEARCH. This lack of clarity likely contributed to these features being the least used (see Figure 9) as well. Evaluators with no prior experience using NLP tools like UCE felt overwhelmed and frequently reverted to the familiar full-text search, which ultimately diminished the quality of their overall experience.

Additionally, many evaluators expressed a desire for more distilled and 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 retrieved through searches, combined with its presentation within the DOCUMENT READER, was often perceived as overwhelming. On the other hand, the CHATBOT proved successful, as it was easy to understand even without prior explanation. It consistently provided helpful information and relevant documents according to the evaluators, making it the most positively highlighted feature in the evaluation results. In general, it can be observed that evaluators in group B provided more favorable feedback than those in group A. This discrepancy may be attributed to the availability of more user-friendly tools, such as the CHATBOT and EMBEDDING SEARCH, for group B. Finally, 86% of the evaluators indicated that they would use UCE again for the given task.

6 Conclusion

We have presented the UNIFIED CORPUS EXPLORER (UCE): a generic NLP application designed for exploring UIMA-annotated corpora. In doing so, our objective is to unify the heterogeneous landscape of NLP by delivering a generic framework that adheres to the standardized UIMA format. This framework facilitates the comprehensive collection, annotation, extraction, and visualization of large text corpora within a cohesive pipeline, which can be tailored to implement the specific requirements of diverse research. In this context, we detailed its microservice architecture, which integrates various technologies within Docker containers, and showcased the diverse features built atop these services. Additionally, we highlighted the range of annotations currently supported by UCE, emphasizing its dynamic adaptability and its applicability across multiple domains and research areas. Finally, the evaluation has demonstrated that UCE, in its current form, provides a foundational platform applicable to research questions within extensive NLP corpora. However, the evaluation also highlighted the need for further development and acknowledged a variety of current limitations, which are discussed in the Limitations section.

Limitations

While UCE already provides a wide array of useful features and has been successfully applied to different corpora, we wish to address its current limitations and outline future directions for development. As UCE is designed as a generic corpus explorer without static underlying data, it is essential to further emphasize its dynamic adaptation and processing capabilities. Currently, domain-specific tools may provide more tailored features and filtering options. To bring UCE on par with these tools while maintaining its versatility across multiple use cases, we must focus on reconstructing and structuring documents in a generic manner that aligns with the specific corpus, which was also one of the key observations from the evaluation. Enhancing its filtering, visualization, and reading functionalities remains one of our primary objectives. Additionally, when handling large corpora, we must ensure that UCE, despite offering a broad array of features, is capable of scaling effectively — an aspect that has not yet been thoroughly evaluated. This necessitates further stress-testing of each service, reevaluating caching and indexing mechanisms, and expanding the infrastructure only with careful consideration. Additionally, the evaluation has shown that UCE's CHATBOT has emerged as a popular and reliable source for guidance and information. We aim to further expand into the realms of generative AI to enhance the capabilities of UCE's tools. Finally, we acknowledge the necessity for improved explanation of UCE within its own environment. One of the primary weaknesses identified during the evaluation was the lack of a proper introduction to the features and the tool itself.

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

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.

Jorge Botas, Álvaro Rodríguez del Río, Joaquín Giner-Lamia, and Jaime Huerta-Cepas. 2022. GeCoViz: genomic context visualisation of prokaryotic genes from a functional and evolutionary perspective. *Nucleic Acids Research*, 50(W1):W352–W357.

Manuel Burghardt, Julian Pörsch, Bianca Tirlea, and Christian Wolff. 2014. Webnlp – an integrated web-interface for python nltk and voyant. 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.

Colin Diesh, Garrett J. Stevens, Peter Xie, Teresa De Jesus Martinez, Elliot A. Hershberg, Angel Leung, Emma Guo, Shihab Dider, Junjun Zhang, Caroline Bridge, Gregory Hogue, Andrew Duncan, Matthew Morgan, Tia Flores, Benjamin N. Bimber, Robin Haw, Scott Cain, Robert M. Buels, Lincoln D. Stein, and Ian H. Holmes. 2023. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biology*, 24(1):74.

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.

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. Online manuscript released August 20, 2024.

Andrew Kane. 2021. pgvector.

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.

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.

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.

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

Luis R Nassar, Galt P Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Brian T Lee, Christopher M Lee, Pranav Muthuraman, Beagan Nguy, Tiana Pereira, Parisa Nejad, Gerardo Perez, Brian J Raney, Daniel Schmelter, Matthew L Speir, Brittney D Wick, Ann S Zweig, David Haussler, Robert M Kuhn, Maximilian Haeussler, and W James Kent. 2023. The ucsc genome browser database: 2023 update. *Nucleic Acids Research*, 51(D1):D1188–D1195.

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*.

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.

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.

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.	[anonymous]
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.	[anonymous]

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 PREPROCESSOR pipeline (Figure 6) 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 PREPROCESSOR, owing to the standardized UIMA format.