

Analyzing Historical Legal Textcorpora

German VET and CVET regulations

Thomas Reiser¹, Jens Dörpinghaus ², and Petra Steiner²

Abstract: The digitization of historical documents has gained particular interest in recent years. The majority of research endeavors aim at digitizing historical documents by extracting text from scanned images. A pipeline that transcribes scanned documents into fully structured texts was utilized to digitize over 900 German VET and CVET regulations. As a preliminary investigation, a basic corpus analysis was conducted to assess the usability of the digitized documents and the necessity for document digitization methods that can generate transcripts that maintain the logical text structure and hierarchy. This paper focuses on the processing of the transcripts created from German VET and CVET regulation images to demonstrate the advantages of fully structured text over plain OCR results and to illustrate that even simple analyses require more information for more comprehensive document understanding.

Keywords: Document digitization, OCR, Legal texts, Corpus analysis

1 Motivation

Document collections are an important part of vocational training research. Vocational education and training (VET) and continuing vocational education and training (CVET) in Germany are regulated since the 1920s [He13]. However, through multiple states, the Third Reich, German Democratic Republic, and Federal Republic Germany, these regulations have evolved over many decades and ideologies. The job archive at the Federal Institute for Vocational Education and Training (Bundesinstitut für Berufsbildung, BIBB) contains many historical VET and CVET regulations that are hardly accessible. In a feasibility analysis in preparation for an effort to digitize this archive, using regulations from 1969 to 2022 that have been published in the Federal Gazette (Bundesanzeiger) and are open available in the internet, a pipeline was developed that transcribes the document images to fully structured texts, with headings, paragraphs, sections, tables, and listings.

The Vocational Training Act (Berufsbildungsgesetz, BBiG) of 1969 established a framework for the majority of these regulations, resulting in only minor differences in their layout and content structure. This consistency makes them suitable for corpus analysis and investigations of the evolution of these regulations over the past fifty years. Our research questions are:

¹ Universität Koblenz, treiser@uni-koblenz.de

² Bundesinstitut für Berufsbildung, jens.doerpinghaus@bibb.de,  <https://orcid.org/0000-0003-0245-7752>;
steiner@bibb.de

1. How does the use of TEI XML facilitate the structure analysis of legal texts, and what information can be gained from such an analysis?
2. How does the structure of the (C)VET documents evolve over time?
3. How much do VET and CVET regulations differ?
4. What are the most frequent words and phrases in the digitized (C)VET regulation text corpus?

2 Related Literature

In recent years, there has been a great deal of research conducted on the subject of document digitization where authors have addressed a number of challenges and goals. In [KKK17], Finnish newspapers are digitized using the Tesseract OCR engine, simply recognizing all text and generating an ALTO XML document. A different approach that addresses the better structured text recognition by modeling the text structure of legal texts in Austria and aligning recognized text to this predefined structure [NF17].

The authors of [Ha19] use OCR to find names of judges in different issues of the Federal Gazette and generate a linked open data set about judges at German federal courts from 1950 to 2019. Similar to this, in [Sa22], invoice papers are digitized by extracting, e.g., product description, quantity, price, and more, into table structured data. Also graph structured data can be extracted from text recognized by OCR. In [KI17], T2KG is presented, an NLP tool that is able to build knowledge graphs from unstructured text. Another way to generate knowledge graphs from text using Open Information Extraction is introduced in [MLR18]. Other approaches include contextual extraction and representation methods based on knowledge graphs, ontologies and taxonomies, see [Dö22a; Dö22b; DS19; DWB22; FD24; FGB22].

There is also several research into web applications that allow interactive OCR, where users can make corrections at each stage of the OCR process. One of the largest efforts towards this is OCR4all [Re19] which allows the usage of different preprocessing steps, segmentation methods and OCR models. However, there are also less extensive tools that also allow the management of digitized document collections, for example in [JMM20; VNP21].

Our aim is to digitize the text corpus of the job archive in a way that each document is encoded in fully structured text in the TEI XML format such that logical units as paragraphs, sections, lists, tables, and figures are recognized, giving more meaning to the text. Although some approaches to do so exist, e.g., in [Ca22] where a tree structured document hierarchy is generated with the HELD model, most of this literature is from around the 2000s and requires some document-specific knowledge by the users [Be07; KK01; TA92]. Due to time constraints, the first version of our digitization workflow still uses predefined rules which we aim to recognize automatically using layout and text features from the text in future

research. More recent approaches utilize object detection methods to fine-tune pre-trained machine learning models with image information [BM19]. As we started our work with only few knowledge about the existing text corpus, except for few manual investigations, the heuristic approach was considered to be reasonable enough for this first analysis.

There is no solution for such a task that is universally applicable for any document and some form of knowledge about the layout is always required, either automatically recognized or pre-defined. On the selected subset of VET and CVET regulations that are open available in the Federal Gazette, a feasibility analysis of such a pipeline and a usability analysis of the results has been conducted. The findings are discussed in this paper.

3 Methodology

We created a document digitization pipeline that takes images of scanned documents, recognizes layout and text by applying OCR, detects important layout elements and utilizes them to structure the text. Because legal texts are usually well structured, this enables the detection of the text hierarchy, consisting of paragraphs, sections, and listings. While paragraphs and their sections contain information on what part of the training program is regulated, lists often contain the taught skills and knowledge. Our corpus analysis is used to show that the transcripts are well processable and the more structured data contains more information than the plain OCR result. As a proof of concept, different aspects have been investigated.

To first understand which paragraph headings are most frequent and relevant throughout time, they are counted. However, some errors are expected in the character recognition process as no OCR model is perfect. Same headings are clustered using DBSCAN with Levenshtein distance. Because there are many longer compound words like they are often found in German language in the paragraph headlines, where spell checkers cannot always find the correct word, this clustering approach has been found more suitable. The headings are extracted from the document image transcripts, clustered, and the relative frequency of each heading is computed. This is done for each VET and CVET regulations separately. For better readability, only the 25 most frequent headings are selected.

To analyze the evolution of paragraph lengths, the ten most frequent paragraphs are selected and their average length in each year is plotted to see which parts became longer and shorter during the selected period. In our case, paragraph length is measured in number of words.

For a better insight into the document lengths, the number of pages as well as the number of paragraphs in each document is plotted in a histogram.

In a last step, the regulation wording and its evolution over time is analyzed. The regulations are grouped by year, then, for all regulations of a year, each word is spell checked to remove most OCR errors, lemmatized to turn them into their base form, and afterwards, counted.

The ten most frequent words of all time are then selected and their relative frequency in each year is plotted.

4 Evaluation

In the first step, similar headlines are clustered using DBSCAN with a maximum Levenshtein distance of 4 between elements in the same cluster and a minimum cluster size of 10 to keep only more relevant paragraph headings. The 25 largest clusters for VET regulations are shown in Figure 1, labeled with the most frequent element in each cluster, i.e., the mode.

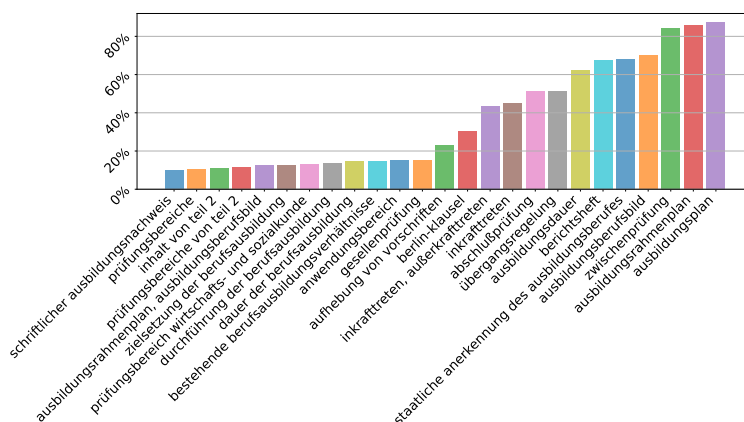


Fig. 1: The 25 most frequent headlines of paragraphs in the selected VET regulations.

The three most frequent paragraph headlines are training plan (Ausbildungsplan), general training plan (Ausbildungsrahmenplan), and intermediate examination (Zwischenprüfung). These three headlines occur about 500 times across all regulations. Although almost every regulation starts with the first paragraph, state recognition of the training occupation (Staatliche Anerkennung des Ausbildungsberufes), this headline is only on fifth position in this plot, caused by known errors in the OCR process that need to be fixed in future work. Issues like this show the necessity of machine learning models that are fine-tuned on our target data set to recognize and classify text areas. These will be part of future research.

For CVET regulations, the same clustering of paragraph headlines as for VET regulations was done. The resulting 25 largest clusters are shown in Figure 2. The top three most frequent headlines for CVET regulations are transitional provision (Übergangsvorschrift), entry into force (Inkrafttreten), and master examination paper (Meisterprüfungsarbeit).

As some parts of the regulations have moved, became less relevant or got added later, the frequency of the ten most frequent paragraphs in each year was considered. The plot in Figure 3 shows this evolution for VET regulations. If one of the values in the lower plot is larger than one, it means that a paragraph occurred multiple times within one

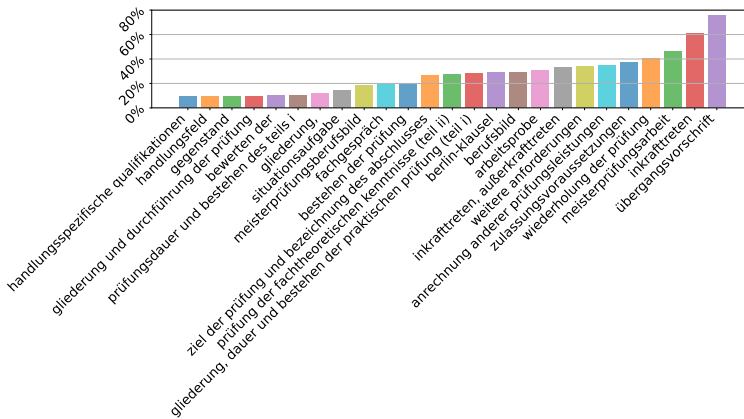


Fig. 2: The 25 most frequent headlines of paragraphs in the selected CVET regulations.

regulation, e.g., in regulations for an industrial sector. For example, in 1999, the regulation on vocational training in the construction industry (Verordnung über die Berufsausbildung in der Bauwirtschaft), contains regulations for eighteen specializations, each of them having its own paragraph for report book (Berichtsheft), causing the number of occurrences in this year to be almost 250%. The mean length of the general training plan (Ausbildungsrahmenplan) is much longer in regulations from 1971 to 1973 than for most of the rest of the years where this paragraph was included. In these years, the general training plan was described in the respective paragraph. Starting from 1974, the general training plan still has its own paragraph, but all details are explained in the appendix, making this paragraph much shorter. Most of the other paragraphs have the same length, except for the final examination (Abschlussprüfung), which has fluctuating values. This is also due to the fact that a new examination structure was introduced in 2002, the *final examination in two parts that fall apart in time* (Abschlussprüfung in zwei zeitlich auseinander fallenden Teilen, § 37 BBiG) or *journeyman's examination in two parts that fall apart in time* (Gesellenprüfung in zwei zeitlich auseinander fallenden Teilen, § 31 HwO).

This form of final examination is better known as the *extended examination* (Gestreckte Prüfung), extended final examination (Gestreckte Abschlussprüfung, GAP), *extended journeyman's examination* (Gestreckte Gesellenprüfung, abbreviated GGP). This examination structure is mapped differently to the traditional examination structure of intermediate examination (Zwischenprüfung) and final examination (Abschlussprüfung), which is reflected in shorter paragraphs.

The paragraph length evolution of the ten most frequent CVET headlines is shown analogously in Figure 4. Although it is among the ten most frequent headlines, the job profile (Berufsbild) appears only from 1971 to 1999 with some gaps where it is not found in any regulation. It is also by far the longest paragraph when considering the average word

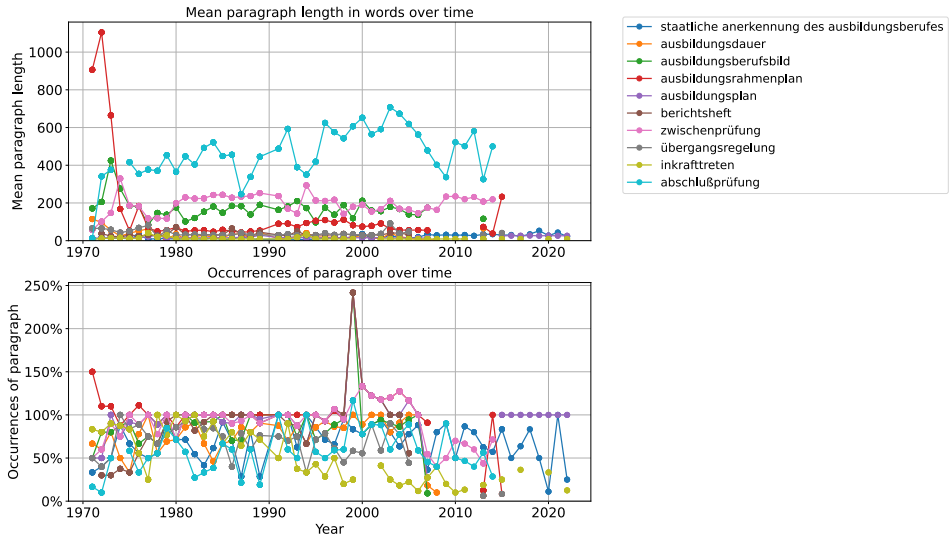


Fig. 3: The upper plot shows the evolution of mean paragraph length in number of words for the ten most frequent paragraphs in VET regulations. The lower plot illustrates the percentage of regulations that contain that particular paragraph in each year. Values greater than 100% indicate that a regulation was for multiple different specializations or professions.

count. The entry into force (Inkrafttreten) occurs in nearly all regulations almost every year from 1975 to 1999 and has short but fluctuating length during this period. Starting from 2000, it has almost the same length in each year but is only in 10 – 65% of all regulations. The length of transitional provision (Übergangsvorschrift) paragraph fluctuates over the years but becomes longer in average. In general, most of the paragraphs have inconsistent word counts, do not occur in all years, and it seems that there is not really a correlation between the occurrences of different paragraphs.

Even though the page numbers can also be taken from the input images and OCR result, this information is still kept by the generated transcripts. To get a first insight into document lengths, the distribution of page counts for VET and CVET regulations is plotted in Figure 5. While about 550 of the regulations have at most forty pages, there are also eighteen regulations with up to 120 pages. These longer regulations are again mostly for industrial sectors or multiple professions in one document which makes them more extensive.

Overall, a large proportion of the continuing training regulations are shorter than the regulations for training occupations. The brevity compared to training regulations mirrors that advanced vocational training regulations are de facto advanced training examination regulations. The content focuses on individual aspects of the examination procedure, while the other typical components of training regulations are missing. The longest of these is about 25 pages long. The biggest regulation that gives the outlier with 26 pages in 2009 is

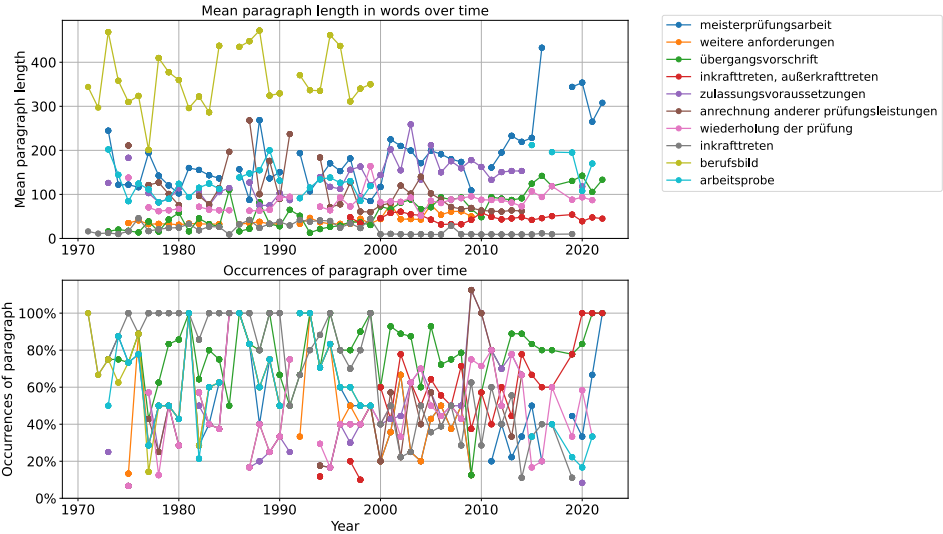


Fig. 4: The upper plot shows the evolution of mean paragraph length in number of words for the ten most frequent paragraphs in CVET regulations. The lower plot illustrates the percentage of regulations that contain that particular paragraph in each year. Values greater than 100% indicate that a regulation was for multiple different specializations or professions.

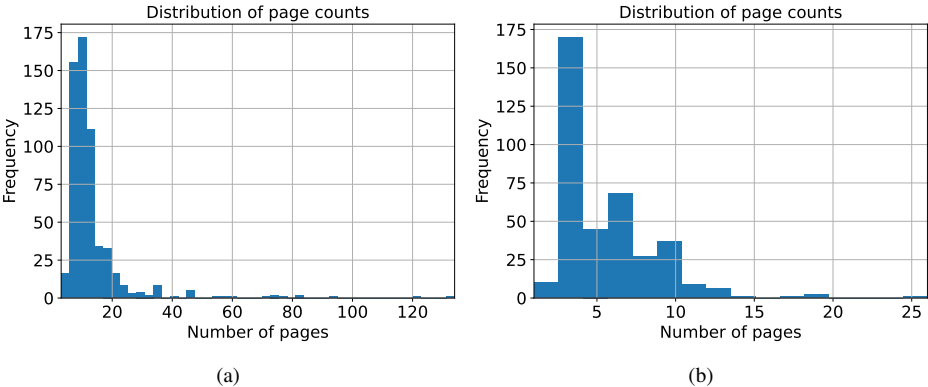


Fig. 5: Distribution of page counts for VET (left) and CVET (right) regulations.

the Regulation on the examination for recognized advanced training qualifications in the media industry (Verordnung über die Prüfung zu anerkannten Fortbildungsabschlüssen in der Medienwirtschaft). Besides these observations, there are no other irregularities or unusual features in the plot.

The more structured data set allows to count the number of paragraphs in each regulation. The distribution of regulation lengths in number of paragraphs for VET and CVET regulations is displayed in Figure 6. Figure 6a shows that most VET regulations consist of eight to twenty paragraphs. There are a few regulations with over thirty paragraphs but these are rare. However, there are some regulations with over ninety paragraphs for an industrial sector such as the regulation on vocational training in the construction industry from 1999 which contains the regulations for eighteen specializations where the same paragraph can be within each specialization. For the CVET regulations, as shown in Figure 6b, most regulations are between seven and twenty pages long. The longest regulation is 32 paragraphs long and again in an industrial sector, the regulation on continuing vocational training in the field of information and telecommunications technology from 2002 with two specializations.

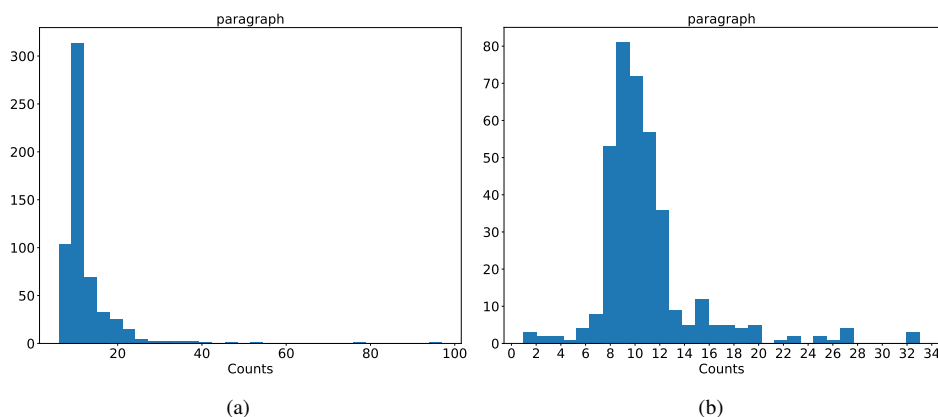


Fig. 6: Distribution of paragraph counts for VET (left) and CVET (right) regulations.

The final step is the wording analysis using the relative frequency of the ten most frequent words in each year. As shown in Figure 7, the word knowledge (Kenntnis) occurs almost twice as often from 1971 to 1976 as in later regulations. Due to the tabular representation of some information in later regulations, this word is required less often as it may only be used in a table header once per page instead of being used in multiple sentences. Noticeable is also that the shapes of the lines for temporal (zeitlich), guideline (Richtwert), week (Woche) look very similar. General training plans (Ausbildungsrahmenplan) often have a time column, named temporal guidelines in weeks (zeitliche Richtwerte in Wochen). Also, until 1998, *training year* (Ausbildungsjahr) runs close to these lines. The line gets closer to zero starting from there since in the appendix tables, columns like time guidelines in weeks in training year (Zeitliche Richtwerte in Wochen im Ausbildungsjahr) are either less present in later regulations or replaced by columns like time reference values in weeks in the training

month (Zeitliche Richtwerte in Wochen im Ausbildungsmonat). Starting from 1977, similar behavior is observed for the words knowledge (Kenntnis) and skill (Fertigkeit). Again, these two words can be found in multiple training plans with a column named skills, knowledge and abilities to be taught (Zu vermittelnde Fertigkeiten, Kenntnisse und Fähigkeiten).

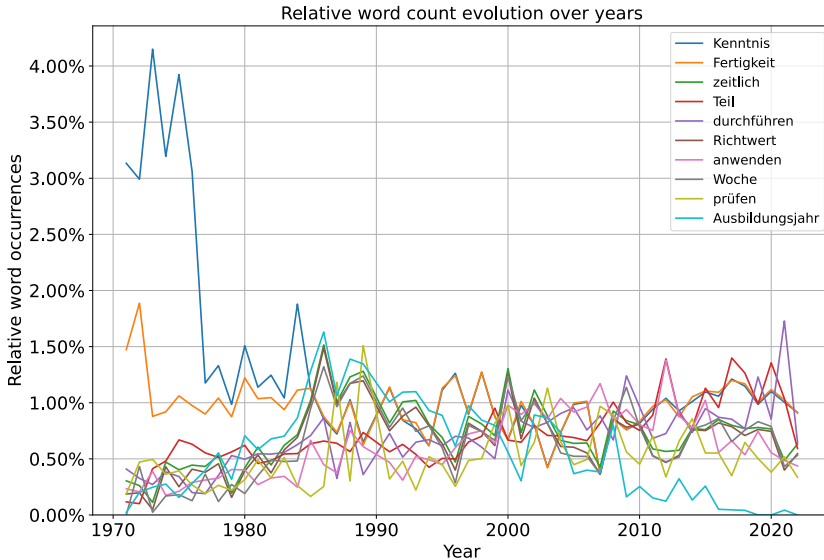


Fig. 7: Relative frequency of the ten most frequent words across all VET regulations in each year.

The result of this wording analysis for the CVET regulations is shown in Figure 8. Reflecting that CVET regulations are advanced training examination regulations, the word examination (Prüfung) is the most frequent word in almost all years. The word examination (Prüfung) is the most frequent word in almost all years. While knowledge (Kenntnis) also has the highest frequency in some years, starting from 2001, it is hardly used. Although it hardly occurs in early years, *assess* (bewerten) appears more than twice as often as before and in 2022 even as the most frequent word. Section (Absatz) becomes significantly more frequent in 2010 and has a small drop afterwards. In this year, many regulations had references to other paragraphs where specific sections have been referenced. Besides that, there are many fluctuations in the curves and hardly any correlations among the ten selected words.

5 Summary and Outlook

We presented results of a corpus analysis that was conducted on automatically generated document transcripts of over 900 German VET and CVET regulations to assess the usability of the digitized documents and to show the necessity for document digitization methods

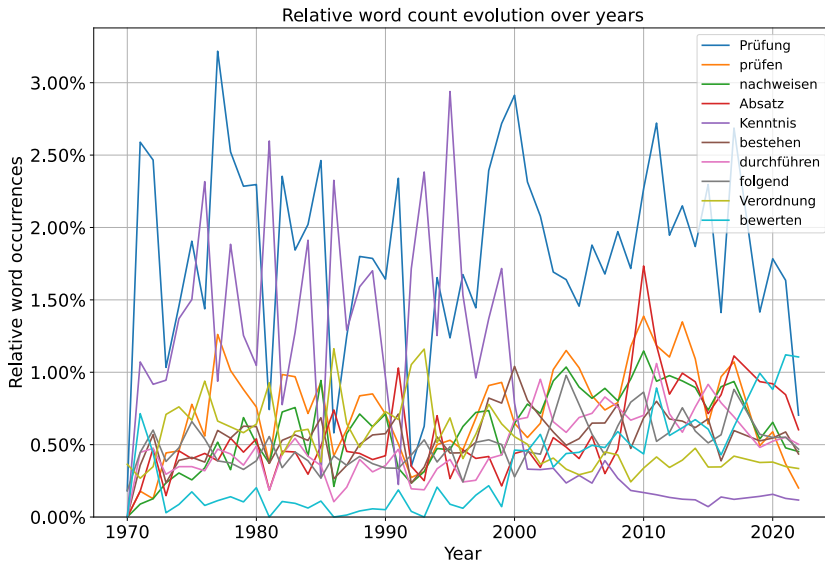


Fig. 8: Relative frequency of the ten most frequent words across all CVET regulations in each year.

that maintain the logical text structure and hierarchy. This paper focused on the transcript processing to demonstrate the advantages of fully structured text and illustrates that even simple analyses require more information for better document understanding.

The used pipeline will currently only work on text with a specific structure. Although this has not been tested, the pipeline should work on most texts that have been published in the Federal Gazette. While machine learning approaches allow for more flexibility, they still need to be fine-tuned for our data set. For better accuracy, we will label training data and apply object detection methods in a future project to improve OCR quality through better page segmentation and better structure recognition by classification of text areas into different categories. While the corpus analysis gives first insights and shows possible investigations on an introductory level, other metrics such as correlation between words should be considered for a deeper insight. Once the pipeline produces high quality image transcripts, the creation of linked data can make our approach interoperable for more labor market research. This includes social media analysis [Dö22b; DT23], historical and longitudinal analysis [DWB22], ontology and knowledge graph driven approaches [Dö23], job advertisement or other labor market data analysis [DD23; DSH23; FDF23].

While our approach is both working and provides substantial results, the corpus analysis on the open available VET and CVET regulations is only a first step, showing that even simpler text processing methods depend on structured text. For a more comprehensive

research on training programs in the past century, we aim to create transcripts of all VET and CVET regulations in the BIBB job archive. This will not only enable structure analyzes of each document but also make text mining easier to extract structured information from all documents. Currently, the developed pipeline for automated document transcription is being advanced and the paper documents are being scanned. Once the archive is available in the target data format, text mining methods can be used to create structured data sets in a standardized format.

References

- [Be07] Beusekom, J. v. et al.: Example-Based Logical Labeling of Document Title Page Images. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Vol. 2, pp. 919–923, 2007.
- [BM19] Binmakhashen, G. M.; Mahmoud, S. A.: Document Layout Analysis: A Comprehensive Survey. *ACM Comput. Surv.* 52 (6), 2019.
- [Ca22] Cao, R.-Y. et al.: Extracting Variable-Depth Logical Document Hierarchy from Long Documents: Method, Evaluation, and Application. *Journal of Computer Science and Technology* 37 (3), pp. 699–718, 2022.
- [DD23] Derksen, F.; Dörpinghaus, J.: Digitalization and Sustainability in German Continuing Education. In: *INFORMATIK 2023 - Designing Futures: Zukünfte gestalten*. Gesellschaft für Informatik e.V., Bonn, pp. 1945–1953, 2023.
- [Dö22a] Dörpinghaus, J. et al.: Context mining and graph queries on giant biomedical knowledge graphs. *Knowledge and Information Systems* 64 (5), pp. 1239–1262, 2022.
- [Dö22b] Dörpinghaus, J. et al.: From social networks to knowledge graphs: A plea for interdisciplinary approaches. *Social Sciences & Humanities Open* 6 (1), p. 100337, 2022.
- [Dö23] Dörpinghaus, J. et al.: Towards a German labor market ontology: Challenges and applications. *Applied Ontology* (18(4)), pp. 1–23, 2023.
- [DS19] Dörpinghaus, J.; Stefan, A.: Knowledge Extraction and Applications utilizing Context Data in Knowledge Graphs. In: *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, pp. 265–272, 2019.
- [DSH23] Dörpinghaus, J.; Samray, D.; Helmrich, R.: Challenges of Automated Identification of Access to Education and Training in Germany. *Information* 14 (10), p. 524, 2023.
- [DT23] Dörpinghaus, J.; Tiemann, M.: Vocational Education and Training Data in Twitter: Making German Twitter Data Interoperable. *Proceedings of the Association for Information Science and Technology* 60 (1), pp. 946–948, 2023.
- [DWB22] Dörpinghaus, J.; Weil, V.; Binnewitt, J.: Towards the Analysis of Longitudinal Data in Knowledge Graphs on Job Ads. In: *The Workshop on Computational Optimization*. Springer, pp. 52–70, 2022.
- [FD24] Fischer, A.; Dörpinghaus, J.: Web Mining of Online Resources for German Labor Market Research and Education: Finding the Ground Truth? *Knowledge* 4 (1), pp. 51–67, 2024.
- [FDF23] Fechner, R.; Dörpinghaus, J.; Firll, A.: Classifying Industrial Sectors from German Textual Data with a Domain Adapted Transformer. In: *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, pp. 463–470, 2023.

- [FGB22] Fettach, Y.; Ghogho, M.; Benatallah, B.: Knowledge graphs in education and employability: A survey on applications and techniques. *IEEE Access* 10, pp. 80174–80183, 2022.
- [Ha19] Hamann, H.: The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data. *Journal of empirical legal studies* 16 (3), pp. 671–688, 2019.
- [He13] Herkner, V.: Grundzüge der Genese und Entwicklung einer korporatistischen Ordnung von Ausbildungsberufen. *Berufsbildung in Wissenschaft und Praxis-BWP* 42 (3), pp. 16–19, 2013.
- [JMM20] Jayoma, J. M.; Moyon, E. S.; Morales, E. M. O.: OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines. In: 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). Pp. 1–6, 2020.
- [KI17] Kertkeidkachorn, N.; Ichise, R.: T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. In: *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
- [KK01] Klink, S.; Kieninger, T.: Rule-based document structure understanding with a fuzzy combination of layout and textual features. *International Journal on Document Analysis and Recognition* 4 (1), pp. 18–26, 2001.
- [KKK17] Koistinen, M.; Kettunen, K.; Kervinen, J.: How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine. *Proc. of LTC*, pp. 279–283, 2017.
- [MLR18] Martinez-Rodriguez, J. L.; Lopez-Arevalo, I.; Rios-Alvarado, A. B.: OpenIE-based approach for Knowledge Graph construction from text. *Expert Systems with Applications* 113, pp. 339–355, 2018.
- [NF17] Nabizai, A.; Fill, H.-G.: Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetzestexten. *Jusletter IT*, 2017.
- [Re19] Reul, C. et al.: OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences* 9 (22), p. 4853, 2019.
- [Sa22] Sai Rakesh Kamisetty, V. N. et al.: Digitization of Data from Invoice using OCR. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). Pp. 1–10, 2022.
- [TA92] Tsujimoto, S.; Asada, H.: Major components of a complete text reading system. *Proceedings of the IEEE* 80 (7), pp. 1133–1149, 1992.
- [VNP21] Van Nguyen, S.; Nguyen, D. A.; Pham, L. S. Q.: Digitalization of Administrative Documents A Digital Transformation Step in Practice. In: 2021 8th NAFOSTED Conference on Information and Computer Science (NICS). Pp. 519–524, 2021.