Representing Standard Text Formulations as Directed Graphs

Frieda Josi[1][0000−0001−7124−780X], Christian Wartena[1][0000−0001−5483−1529], and Ulrich Heid[2]

[1] University of Applied Sciences and Arts Hanover, Expo Plaza 12
30539 Hanover, Germany frieda.josi@hs-hannover.de,
christian.wartena@hs-hannover.de, https://im.f3.hs-hannover.de/en/
[2] University of Hildesheim, Lübecker Straße 3, 31141 Hildesheim, Germany
heidul@uni-hildesheim.de,
https://www.uni-hildesheim.de/fb3/institute/iwist/

Abstract. In order to ensure validity in legal texts like contracts and case law, lawyers rely on standardised formulations that are written carefully but also represent a kind of code with a meaning and function known to all legal experts. Using directed (acyclic) graphs to represent standardized text fragments, we are able to capture variations concerning time specifications, slight rephrasings, names, places and also OCR errors. We show how we can find such text fragments by sentence clustering, pattern detection and clustering patterns. To test the proposed methods, we use two corpora of German contracts and court decisions, specially compiled for this purpose. However, the entire process for representing standardised text fragments is language-agnostic. We analyze and compare both corpora and give an quantitative and qualitative analysis of the text fragments found and present a number of examples from both corpora.

Keywords: Graph-based Text Representations · Legal Writings · Standardised formulation

1 Introduction

In legal writings, like contracts or court decisions parts of the text are frequently reused. This type of text reuse is different from plagiarism since this is completely legal; there is no single source of a reused text, but most fragments are used ubiquitous since many years. Reusing smaller or larger text fragments is not only done for efficiency reasons, but the use of standardized phrases and passages is essential for the proper understanding and interpretation of legal writing.

Standardized expressions can vary from short phrases to long passages. In the present work we focus on standardized formulations that consist of several sentences. Surprisingly, hardly any attempt has been undertaken to try to identify such passages automatically. Though the existence and importance of such

The final authenticated version is available online at
https://link.springer.com/chapter/10.1007/978-3-030-86159-9_34.

passages has been noted by various authors there is also no clear definition of a standardized passage. When we try to identify standardized passages we are faced with two problems: (1) when is a passage just a repetition of a few sentences used by one author in similar documents and from which point on can we call it a standardized passage? And (2) when is a passage a variation of a commonly used passage and how much has it to deviate to become an independent formulation? Related to the second issue is also the question which variant among all variants found is the most representative one.

Since we often find many similar passages or variants of the same passage, we propose not to choose one, but to represent standardized passages as directed acyclic graphs (DAG). In an empirical study on two large legal corpora, we show that it is possible to cluster all frequent sequences of sentences in small clusters that are mutually almost disjunct and can be represented in uncluttered directed graphs. We see that many, but not all of these *DAGs* are good candidates for standardized passages. We are still far away from a general definition of a standardized passage, but we think that the investigations on the large legal corpora presented here can contribute to a better understanding of the nature of standardized passages and their role in legal documents.

2 Phrasemes and Standardized Passages in Legal Writing

In the legal domain texts play an important role. The primary function of a legal text is not to inform or to convince the reader, nor to describe something. Rather a legal text or more precisely a *constitutive* text is a declaration in the sense of Searle's illocutionary speech acts [17] and thus creates a reality and shapes its environment [11, 4]. A contract or a court decision is not a model of the reality, but actors are bound to do what the text prescribes. Thus, though these texts are produced en masse, they need to be formulated carefully and their interpretation should be clear and without ambiguities. Thus, often standardized formulations are used to ensure the correct formulation and interpretation (see e.g. [15, p.251]) and the reuse of such passages is an essential trait of legal and normative texts (see e.g. [21]). The standardized formulation often has a specific interpretation and legal consequences both known to the (legally educated) reader and author. Engberg [6, 7] stresses that in many text types, like e.g. court decisions, it is important to use exactly the conventional formulation, since this works as a signal to the (expert) reader, who will be disturbed and might misunderstand the text if the convention is not followed.

Standardized formulations are quite flexible and show a considerable degree of variation. Names of persons, dates etc. can be exchanged but also words or phrases can be added or deleted. A standardized formulation can consist of just a few words or of several sentences. Phrases or short pieces of text that are frequently used and have a fixed meaning are well known from other disciplines and from general language and are called idiomatic expressions, phrasemes, phraseologisms or routine formulae or routine expressions [2, 16]. All these terms are used as synonyms or with slightly different meanings but always for short phrases.

For longer passages Lindroos [13] found the following terms: text patterns (*Textmuster*), formulaic or schematic texts (*formelhafte oder schematisch gestaltete Texte*), stereotyped texts (*stereotypisierte Texte*) or preformed structures (*vorgeformten Strukturen*). Wozniak [21] uses the term pragmatic phraseologisms for all types of recurrent phrases and text and calls longer recurrent passages *Kleintexte* (Small texts), formulaic (short) texts or textual phraseologisms (*textwertige Phraseologismen*). Płomińska [15] uses the terms micro and macro routine expressions for recurrent phrases and recurrent units that consist of one or more sentences, respectively. Płomińska further classifies routine expressions found in court decisions according to their function.

According to Wozniak [21] there is no consensus whether standardized passages can have the status of phraseologisms. However, she notes that such passages frequently occur in certain text types, like court decisions or contracts. Often these textual phraseologisms have a fixed syntactic structure with a variable lexical filling. The number of slots that need to be filled differs for each text type and the part of the text the passage belongs to. E.g. the final provisions in contract show less lexical variable slots than the legal instructions in a notification.

3 Related Work

There is abundant research on the possibilities to automatically find collocations and short phrases with an idiomatic meaning or that are typical for a discipline or text type. However, not much work was done on the automatic recognition of longer routine expressions. Wahl and Gries [19] is an exception, but they still focus on the phrase level and units shorter than complete sentences. Finding reused text passages and sentences is often important for the analysis of documents. Kliche et al. [12] developed a tool suite where users, can upload the texts to be analyzed and then define patterns to find reused texts in the newsletter corpus. Another research project on text reuse in newspaper articles is described by Clough et al. [5]. The similarity of the documents is calculated using the n-gram overlap, with n-gram lengths of 1 to 10 words. For the overlaps between the sentences of these similar documents the substring matching algorithm Greedy-String-Tiling [20] is used. Recently, a number of papers have described the reuse of texts in legal documents. Burgess et al. [3] present a tool to discover the reuse of text passages from laws for new bills in order to make it more difficult for lobbyists to influence legislation. They assume that lobbyists include similar text formulations in the bills. To detect this text reuse, they use Elasticsearch to find relevant documents on the Internet. Then, they use the Smith-Waterman algorithm for aligning text passages to identify related text passages in the law and in the bill. Finally, they calculate the similarity of these two text passages using the Jaccard coefficient. Graph-based representations have been used before to capture variation in texts. Filippova [8] identifies word changes in a set of similar sentences by transforming the words in the sentences into a graph. Based on the shortest sentence in the set, the paths in the sentence are verified. In this way,

Table 1. Data overview of both corpora.

	Case law corpus	***Contract corpus***
Documents	4,250	2,167
Sentences	308,832	751,281
Tokens	7,202,106	10,433,943

longer and more complex sentences can be succinctly reduced to the shortest variant (multi-sentence compression). Ma et al. [14] propose paraphrasing sentences using a graph-based method. The sentences used are from one domain, so the authors could ensure that the thematic scope of the sentences is quite similar. In the first step, they use the Word aligner by Sultan et al. [18] to align pairs of sentences. They then generate *DAGs* for each group of sentences to identify paraphrases.

4 Legal Corpora

We have compiled two corpora with two different types of legal texts. The first corpus consists of court decisions and the second of contract texts.[3]

Case Law Corpus The first corpus contains anonymized court decisions from German criminal procedures from the years 2015 to the beginning of 2020. The court decisions are published by the Federal Court of Justice (*BGH*). The court decisions were crawled directly from the website of the case law database and are available in HTML format for further work. The source for the documents in the case law corpus is given in the appendix under A.1. An overview of the number of documents, tokens, sentences and sentence clusters for all used corpora is shown in Table 1.

Contract Corpus The second corpus contains contracts of the Hamburg City Administration and the Bremen City Administration. Some cooperation agreements between universities are also included. Among these contracts are several contracts that universities have concluded with external service providers. All contracts are available in PDF format. The contract texts had to be extracted, cleaned and prepared for further processing. The quality of the scanned contracts from the city administrations is not so good. There are documents with a lower scanning resolution, pages have been scanned at different angles, and so on. Moreover, all information that represents personal data has been blacked out. The contracts in this corpus are from the years 2014 to 2019 and are publicly available under the *Data License Germany Attribution 2.0* or

[3] The sources for the documents compiled for both corpora will be published on our website: `http://textmining.wp.hs-hannover.de/juver.html`. Likewise, we publish the developed methods and also the document collections on our project page.

Data License Germany Null Version 2.0 license. The corpus is available under `http://textmining.wp.hs-hannover.de/juver.html`. The sources used in this corpus are listed in the appendix under A.2.

5 Method

5.1 Recurrent Sentences

In order to abstract away from these small variations, we cluster all sentences into clusters with very similar sentences. We use the minimum link (agglomerative) clustering algorithm and the Jaccard coefficient between the sets of character trigrams extracted from each sentence as a similarity measure. In the experiments described below, we require a minimal Jaccard index of 0.75 between two sentences in order to be considered for merging their clusters. Given the large number of sentences we cannot compute the similarity between each pair of sentences. Using a word index, for each sentence we retrieve for each sentence the first 100 sentences with the highest number of common words (excluding stop words). Only for these sentences we compute the trigram similarity.

In order to put sentences in which a number or date is changes into the same cluster, we remove stop words and replace numbers and dates by a general token.

As we will see below, in almost all cases the results are quite intuitive. There are, however, a few cases in which the result is not as we would like it to be. In the first place errors in sentence segmentation obviously cannot be undone by the clustering. In the second place, if there are too many (small) OCR-errors, the trigram overlap between two sentences is often large enough, but sometimes the number of common tokens is too small and the sentence pair is not considered as a candidate for computing the exact similarity.

5.2 Standardized Passages

Standardized legal formulations often go beyond the sentence level and can extend to long paragraphs. Not all frequent sequences of sentences that we find in our corpora are routine formulations. We also have cases where almost the same contract is concluded with various partners and only a few names and amounts and dates are changed in a standard contract. For the moment, we will not try to draw a line between standardized contracts and standardized passages as the distinction is not clear cut and we find all kind of intermediate cases.

Frequent Sequences of Sentences We start our search for standardized passages by selecting all sequences that exceed a given minimum frequency. We start counting all pairs of two sentences and then proceed step wise to longer sequences. For efficiency we use the fact that a sequence $\langle s_1, s_2 \dots s_n \rangle$ only can be frequent if both $\langle s_2 \dots s_n \rangle$ and $\langle s_2 \dots s_{n-1} \rangle$ are, very similar to the procedure followed in the Apriori algorithm for pattern detection in databases [1].

Clustering of Sequences and Representation as Directed Graphs Once we collected all frequent patterns, there are many overlapping ones. Patterns can either be a proper subpattern of another pattern, or we have partially overlapping patterns.

We now consider each pattern as a directed graph in which each sentence is a vertex (node) and in which there is an edge from s_1 to s_2 if s_2 follows s_1 in the detected pattern. Now we add start and end nodes to each graph and cluster them. To do so, we need a similarity measure between graphs. Various similarity measures based on the number of common edges or common can be used. We decided to simply use the number of common edges as a similarity measure and to use again the minimum link (agglomerative) clustering algorithm. We do not use any stopping criterion and thus two graphs that share an edge are guaranteed to be in the same cluster. This has the advantage that, when analyzing the corpus, there is no ambiguity and there are no conflicting overlapping sequences in the corpus. Finally, we merge all graphs from a cluster into one new graph.

The graphs build in this way can be cyclic. For some purposes it is advantageous if we have acyclic graphs. For this purpose, we use a number of heuristics to remove edges. First, we find a number of cycles by searching the shortest path from each vertex to itself. Now we remove the edge that is on the largest number of cycles. If there is a tie we remove the edge going to the vertex with the highest in-degree. If there is still a tie, we remove the edge starting in the vertex with the highest out-degree. If there are still several possibilities we remove the edge with the lowest count (number of occurrences, see below). This process is repeated until the graph is acyclic. In the examples below, removed edges will be displayed in red.

6 Corpus Analysis

The graphs found are based on the patterns counted in the corpus. Here we used some minimum frequency. When merging the graphs, new possibilities arise to traverse the graph. Some of them might occur in the corpus but with a frequency below the initial threshold. E.g. consider the situation in which we have a threshold of 10 and sequences $\langle s_1 s_2 s_3 \rangle$ and $\langle s_2 s_3 s_4 \rangle$ that both occur 15 times. These sequences will be merged into one graph that has a path $\langle s_1 s_2 s_3 s_4 \rangle$ corresponding to a sequence that eventually occurs in the corpus bit less than 10 times. Thus, after building our graphs we count all instances of the possible patterns again and add weights to the edges of the graph, indicating how often each pair of two sentences connected by the edges occurs in the corpus.

Both corpora consist of a number of documents. We divide each document into sections, defined as a heading or a series of headings followed by normal text. For the case law corpus, we can extract the required information for this easily from the HTML structure. For the contract corpus, we rely on our classification of text elements (see [9]). For the case law corpus we use a threshold of 10 occurrences for a pattern to be considered. For the contract corpus we take 50 occurrences as a lower boundary.

Table 2. Sentences, Patterns and Graphs

	Case law corpus	*Contract corpus*
Sentences	308,832	751,281
Unique Sentences	197,811	448,288
Sentence Clusters	178,579	381,895
Patterns	161	1605
Pattern Graphs	94	227
Pattern/Graph Occurrences	7,569	26,954

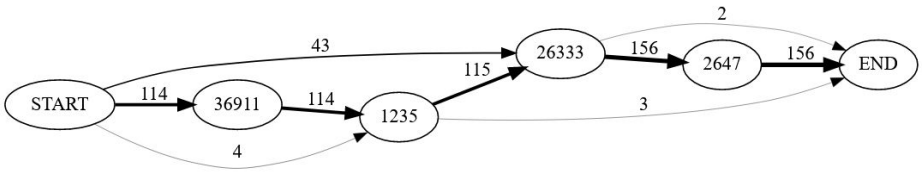

Fig. 1. Example of a typical linear sequence of sentences represented as a DAG with various options to start and end from the contract corpus. Sentences are represented by their IDs.

6.1 Statistics

In the case law corpus, we find 161 patterns that can be clustered into 94 graphs that have in total 7,569 instances in the corpus. For the contract corpus we find 227 graphs with 26,954 instances. Details are given in Table 2.

Let us have a closer look at the graphs. The majority of the graphs consists just of two nodes. In case law corpus 79 out of 94 graphs consist just of two sentences while the largest graph comprises 7 sentences. For the contract corpus 124 graphs have only two sentences. Here, the largest graph has 17 sentence nodes. In most graphs all sentences are on the longest path. This means that there are various options where to start and stop, but there are no options where one sentence or another sentence can be chosen. In most cases there seem to be optional parts at the start and the end of the standardized passage. In some cases, also the beginning or the continuation was not properly recognized, because of too much variation, OCR errors or segmentation errors. Fig. 1 shows an example of such a "linear" graph. For the case law corpus only 3 graphs do not have such a linear structure. In the contract corpus 12 graphs have a nonlinear structure.

Since the clustering is based on the number of common edges, it is possible that a sentence is part of several graphs. For the case law corpus, we find 142 sentences that are part of at least one graph; 29 of them show up in at least two graphs, the most frequent one in 16 graphs. This most frequent sentence turns out to be the heading *Gründe* (Reasons). The most volatile real sentence is part of 8 graphs and reads *Gegen dieses Urteil wendet sich der Angeklagte mit seiner auf die Verletzung formellen und materiellen Rechts gestützten Revision.*

(The defendant opposes this judgment with his appeal based on the violation of formal and substantive law.)

In the contract corpus, we found 715 sentences that are part of a longer frequent sequence, 78 of which occur in at least two graphs. The most frequent one, however, is only a part of two graphs, and again is just a single word: *Einzelpreis [EUR]* (Unit price [EUR]). The sentence showing up in the largest number of different graphs, three to be precise, reads: *Die einzelnen Aufgaben und die Verteilung der Zuständigkeiten sind wie folgt geregelt:* (The individual tasks and the distribution of responsibilities are settled as follows:).

6.2 Examples of Sentence Clusters

We have discussed sentence clustering in detail elsewhere, see [10] (to appear) and briefly described in section 5.1. Here we will just give an impression of the results.

It turns out that the sentence clustering is quite essential in the whole approach. If we do not cluster at all, we miss many interesting sequences, since several occurrences of a sentence have changed names, dates or other small variations. If too many similar sentences end up in a cluster, we find different continuations that in fact correspond to different sentences in the same cluster. We found that too low trigram similarity requirement in combination with single link clustering in some cases leads to long chains of sentences that are increasingly different.

A typical example of a cluster is given by the following four sentences:

- *Die Gefährlichkeitsprognose begegnet ebenfalls durchgreifenden rechtlichen Bedenken.*
- *Auch die Gefährlichkeitsprognose begegnet durchgreifenden rechtlichen Bedenken.*
- *7 c) Auch die Gefährlichkeitsprognose begegnet durchgreifenden rechtlichen Bedenken.*
- *11 c) Zuletzt begegnet auch die Gefährlichkeitsprognose durchgreifenden rechtlichen Bedenken.*
- Translation: "The danger prognosis also encounters sweeping legal concerns."

6.3 Examples and Analysis of the Found Passages

Since it is unclear what a standardized passage exactly is and what standardized passages occur in our corpora, it is impossible to give a numeric evaluation of our approach. Instead, let us have a look at the sequences that are found.

The results found are of course highly dependent on the minimum support required for each sequence. If we lower that requirement much more sequences would be found, but we would expect them to be less general.

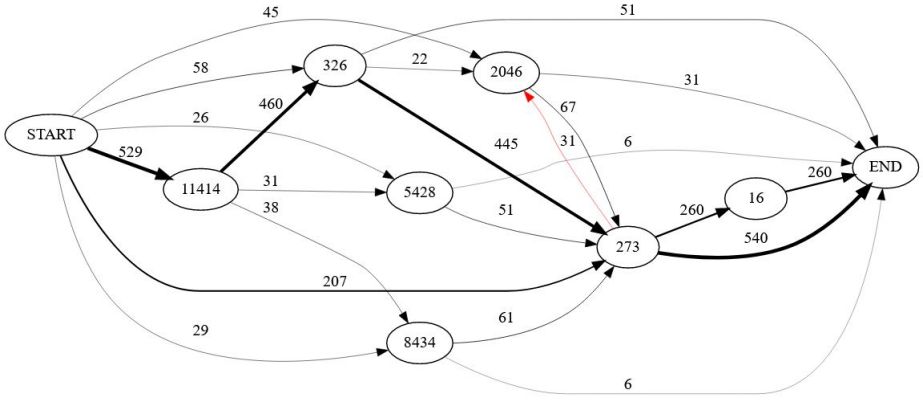

Fig. 2. Example of a complex graph from the case law corpus.

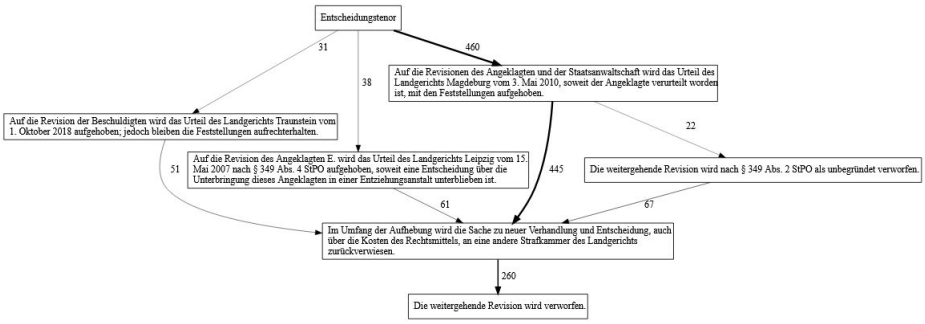

Fig. 3. Alternative representation of the graph from Fig. 2

Case Law Corpus As mentioned above most graphs are very simple. Let us nevertheless have a look at one of the most complex graphs found, as it shows the potential of this type of representation. Fig. 2 shows a graph with some sentences on the revision of a decision in its full complexity. If we remove the start and end node (end thus the information on possible subsequences) and also remove the cycle in the graph, the structures become quite clear and it is even possible to add the text to the nodes (Fig. 3). The structure now becomes quite clear. First we have a heading ('Decision Tenor') followed by three variants of the statement that the decision of a regional court is overturned on the appeal of the accused. In the first case, the findings remain, in the third variant, the findings are also overturned, the second variant refers to a more special case. Then a sentence about the further procedure and the costs follow. Finally, there are two variants of the sentence stating that a further revision is not possible that can appear at different positions in the passage.

If we take a closer look at the large number of simple graphs, we see many cases, that consist of a heading (like 'Decision' or 'Reasons') followed by a typical

opening sentence for that section. Furthermore, we have a lot of sequences in which the first part gives the decision of the court, especially on an appeal, followed by one or two sentences on the court costs. A typical example is the following sequence:

1. *Die weitergehende Revision wird als unbegründet verworfen.* (The further revision is rejected as unfounded.)
2. *Im Umfang der Aufhebung wird die Sache zu neuer Verhandlung und Entscheidung, auch über die Kosten des Rechtsmittels, an eine andere Strafkammer des Landgerichts zurückverwiesen.* (To the extent of the annulment, the matter will be referred back to another criminal division of the regional court for a new hearing and decision, including the costs of the appeal.)

Two further frequent types of sequences are constituted by passages about the role and power of the court and by definitions of certain facts, e.g.:

1. *Bedingten Tötungsvorsatz hat, wer den Eintritt des Todes als mögliche Folge seines Handelns erkennt (Wissenselement) und billigend in Kauf nimmt (Willenselement).* (Those who recognize the occurrence of death as a possible consequence of their actions (element of knowledge) and approve of it (element of will) have a conditional killing intention.)
2. *Beide Elemente müssen durch tatsächliche Feststellungen belegt werden.*(Both elements must be substantiated by factual findings.)

Contract Corpus In the contract corpus, we find a different situation. This is partially due to the bad PDF-quality of the downloaded contracts and the OCR errors made. Here we often find long sequences extending over a section heading. We only search sequences within sections, but sometimes the headings are not recognized.

As an example, consider the following sentence cluster. This cluster has 26 variants, but if we ignore the variants caused by OCR-errors, 2 versions remain:

1. *3.1 Infrastruktur Die Leistung des Auftragnehmers erfolgt ausschließlich auf unterstützten Plattformen, die durch Hersteller freigegeben sind.*
2. *Die Leistung des Auftragnehmers erfolgt ausschließlich auf unterstützten Plattformen, die durch Hersteller freigegeben sind.*
3. Translation: [3.1 Infrastructure] The service of the contractor is carried out exclusively on supported platforms that have been approved by the manufacturer.

All variants together are found 154 times in the corpus, 88 times preceded by the sentence *3.1 Infrastructure* and 52 time preceded by each time exactly the same sentence with system configuration. Similar as in this example variants in the graphs are often due to segmentation errors and so to say repaired by the clustering of the sequences.

The last example also shows a further characteristic of the contract corpus. This sentence does not look like a routine formulation but seems to be part of

a long section (the above example was from a graph with 11 sentences and 3 headings) with general conditions for IT-systems that is copied or appended to many contracts.

Again, we also find many section headings followed by a typical first sentence, but now also we find many headings followed by a subheading.

In our corpus there are many contracts on IT-services. Since we only consider sequences occurring at least 50 times, we have a lot of very specific passages on the availability of systems, back-ups, etc.

Nevertheless, also typical general contract formulations show up:

1. *Mit diesem Vertrag wird eine etwaige Vorvereinbarung abgelöst.* (This contract replaces any preliminary agreement.)
2. *Rechte und Pflichten der Vertragsparteien bestimmen sich ab dem Zeitpunkt seines Wirksamwerdens ausschließlich nach diesem Vertrag.* (Rights and obligations of the contracting parties are exclusively determined by this contract from the time it becomes effective.)

Comparison of the Corpora The contract corpus poses much more challenges to extract proper sequences, as the texts are more structured with headings, subheadings, tables, lists, appendices, etc. Moreover, the contracts are only available as a scanned PDF. This combination makes it hard to extract proper sequences of sentences.

If we compare the sequences found a general trend seems to be that in court decisions, we find many short sequences that can be seen as instances of what are called routine expressions in the literature discussed above. In the contracts we also find this type of formulations, but not as many. Here, the page-wise copying of terms and conditions from one contract to another seems to be the main source for recurring sentence sequences.

7 Conclusion and Future Work

Routine expressions consisting of several sentences have been observed and discussed in the literature. We are not aware of any previous attempt to detect such schematic text fragments automatically. One of the challenges is the great flexibility and many variations these formulations have. To overcome this, we comprise similar sentences in clusters and represent routine expressions as directed graphs of the sentence clusters. We have shown that we effectively can find many such longer formulations by a common pattern detection algorithm and subsequent clustering of the patterns found.

The present work is a first exploration of the topic of automatic detecting standardized formulations and can be extended in many directions. One of the following steps we want to do is to get more insight in the variations that are possible and the aspects of a formulation that have to be constant. This also can lead to a first application that readers can point to remarkable deviations from a standard text.

References

1. Agrawal, Rakesh und Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. pp. 487–499. VLDB '94, Morgan Kaufmann Publishers Inc. (1994)
2. Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N.R.: Phraseologie: Objektbereich, Terminologie und Forschungsschwerpunkte. In: Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N.R. (eds.) Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung., pp. 1–10. Mouton de Gruyter, Berlin/New York (2007)
3. Burgess, M., Giraudy, E., Katz-Samuels, J., Walsh, J., Willis, D., Haynes, L., Ghani, R.: The Legislative Influence Detector: Finding Text Reuse in State Legislation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. pp. 57–66. ACM Press (2016). https://doi.org/10.1145/2939672.2939697
4. Busse, D.: Sprache und Recht, pp. 383–393. J.B. Metzler, Stuttgart (2018). https://doi.org/10.1007/978-3-476-04624-6_37
5. Clough, P., Gaizauskas, R., Piao, S.S.L., Wilks, Y.: METER: MEasuring TExt reuse. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002), `http://dx.doi.org/10.3115/1073083.1073110`, conference Name: ACL-02 Library Catalog: eprints.whiterose.ac.uk Meeting Name: ACL-02 Pages: 152-159 Place: Philadelphia Publisher: ACL
6. Engberg, J.: Signalfunktion und Kodierungsgrad von sprachlichen Merkmalen in Gerichtsurteilen. HERMES - Journal of Language and Communication in Business pp. 65–82 (1992). https://doi.org/10.7146/hjlcb.v5i9.21506
7. Engberg, J.: Does routine formulation change meaning? - The impact of genre on word semantics in the legal domain, pp. 31–48. De Gruyter Mouton (2000), `https://www.degruyter.com/view/book/9783110826005/10.1515/9783110826005.31.xml`
8. Filippova, K.: Multi-sentence Compression: Finding Shortest Paths in Word Graphs. In: Proceedings of the 23rd Int. Conference on Computational Linguistics. pp. 322–330. COLING '10, Association for Computational Linguistics (2010)
9. Josi, F., Wartena, C.: Structural Analysis of Contract Renewals. In: Proceedings of the ACM CIKM 2018 Workshops. Turin (2018)
10. Josi, F., Wartena, C., Ulrich, H.: Identifizierung von häufig vorkommenden Textabschnitten in juristischen Korpora. In: 56th Linguistics Colloquium. vol. 56. Peter Lang (2021), to appear
11. Kjær, A.L.: On the structure of legal knowledge: The importance of knowing legal rules for understanding legal texts. Language, Text, and Knowledge. Mental Models of Expert Communication pp. 127–161 (2000)
12. Kliche, F., Blessing, A., Heid, U., Sonntag, J.: The eIdentity text ExplorationWorkbench. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA) (2014)
13. Lindroos, E.: Dissertation: Im Namen des Gesetzes. Eine vergleichende rechtslinguistische Untersuchung zur Formelhaftigkeit in deutschen und finnischen Strafurteilen. Fachsprache **37**(3), 218–222 (2015). https://doi.org/10.24989/fs.v37i3-4.1293
14. Ma, D., Chen, C., Golshan, B., Tan, W.C.: Essentia: Mining domain-specific paraphrases with word-alignment graphs. In: Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing

(TextGraphs-13). pp. 52–57. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-5307
15. Płomińska, M.: Routine expressions in german legal texts – an attempt at typology. Colloquia Germanica Stetinensia **29**, 239–253 (2020). https://doi.org/10.18276/cgs.2020.29-13
16. Sailer, M.: Idiom and phraseology. In: Aronoff, M. (ed.) Oxford Bibliographies in Linguistics. Oxford University Press, New York (2013). https://doi.org/10.1093/obo/9780199772810-0137
17. Searle, J.R.: A taxonomy of illocutionary acts. Language, mind, and knowledge **07** (1975), `http://conservancy.umn.edu/handle/11299/185220`, accepted: 2017-03-16T18:32:14Z Publisher: University of Minnesota Press, Minneapolis
18. Sultan, M.A., Bethard, S., Sumner, T.: Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. Transactions of the Association for Computational Linguistics **2**, 219–230 (2014). https://doi.org/10.1162/tacl_a_00178
19. Wahl, A., Gries, S.T.: Computational extraction of formulaic sequences from corpora. Computational Phraseology **24**, 83 (2020)
20. Wise, M.J.: Neweyes: a system for comparing biological sequences using the running karp-rabin greedy string-tiling algorithm. Proceedings. International Conference on Intelligent Systems for Molecular Biology **3**, 393–401 (1995)
21. Woźniak, J.: Pragmatische Phraseologismen in ausgewählten Rechtstexten–ein Systematisierungsversuch. Lingwistyka Stosowana/Applied Linguistics/Angewandte Linguistik pp. 149–162 (2017)

A Appendices

A.1 Sources for *Case law corpus*

1. Bundesgerichtshof (BGH) – Decisions from criminal law:
`https://www.hrr-strafrecht.de/hrr/db/abfrage.php?type=erweitert&sortieren=relevanz&sortrichtung=ab&gericht=BGH&aktenzeichen=&datvon=&datbis=&volltext=&kurzbeschreibung=&norm=StGB&medium=-&verknuepfung=und&sz=2`

A.2 Sources for *Contract corpus*

1. Stadtverwaltung Hansestadt Hamburg – City administration of Hamburg:
`http://suche.transparenz.hamburg.de/dataset?q=vertrag&esq_title=&check_all_`
2. Stadtverwaltung Bremen – City administration of Bremen:
`https://www.transparenz.bremen.de`, Keyword: *Vertrag*
3. Cooperation contracts between universities and also between universities and service providers: We searched specifically for contract files on university websites and added them to *Contract corpus*.