CoUSBi: A Structured and Visualized Legal Corpus of US State Bills

Aikaterini-Lida Kalouli*, Leo Vrana*, Vigile Marie Fabella‡, Luna Bellani‡ and Annette Hautli-Janisz*

*Department of Linguistics ‡Department of Economics
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

This paper reports on an approach to automatically transform semi-structured and public databases of US state-level legislative bills into a structured, legal corpus, namely the Corpus of US Bills (CoUSBi). Our work has resulted in a methodology and a corpus that makes this data usable for natural language processing applications. It thus also lays important groundwork for work in the social sciences, particularly in the fields of political science and economics where there is a growing interest in the relationship between legislative policy-making and economic behavior. Against the backdrop of eventually contributing to a Legal Knowledge Graph, the paper shows that the corpus we provide already fulfills the requirements to be connected to other resources: We automatically extract correspondences between individual state bills and model bills from independent organizations, generating interesting insights into the legislative process. We furthermore use NEREx, a Visual Analytics framework, that allows us to capture important content of the bills at a glance.
Keywords: Resource development, US state bills, model bills, Visual Analytics

1. Introduction

As digitalization becomes increasingly infused throughout all aspects of society, it becomes even more important to make the legal domain and in particular the legislative process more accessible to the public. As a consequence, legislative bodies increasingly make information available online and users are faced with a flood of information, often unconnected to other relevant information and presented in a way that is not conducive to easy reference. This development is a classic case in which natural language processing (NLP) applications can be of great help. By way of structuring information from the legal domain in a certain way we can then design automatic systems that can shed light on aspects of legislative reality, e.g. answer specific queries in question-answering systems or in search engines, summarize legal documents, compare different versions of the same document and link related documents with each other.

Creating these structured resources across languages, legal traditions and types of legislative text involves standardizing information through interchange formats, for example as done with MetaLex (European Committee for Standardization, 2010), an interchange format for sources of law. For a general application of these standards, however, on the one hand a challenge lies in collecting documents from different kinds of sources and converting them into a standardized format. On the other hand, this format must also be common enough to be accessible for the communities involved, in particular NLP. With the aim of eventually creating a Legal Knowledge Graph, any approach faces questions about which key components need to be encoded to make specific types of legal text usefully accessible for further research.

In this paper we report on work that addresses the gap between legal text in the wild and a structured resource which can ultimately serve as input to a Legal Knowledge Graph. To this end, we collected all enacted, education-related bills of the US states North Carolina and New Mexico in the years 2007 to 2015. We automatically extract key information from the bills and convert them to structured documents according to the TEI standard (2017),[1] a common text encoding standard in the NLP community. We then link those bills to model bills of ALEC, the American Legislative Exchange Council,[2] established in 1975 (Hertel-Fernandez, 2014), which is a nonprofit organization in the US that drafts model state-level legislation.[3] In particular, we track down parts of the enacted bills that are also found in the model bills and we mark them accordingly within the structured documents. With this step we aim at detecting potential factors that drive a decision or the passage of a law. We also make a suggestion as to how to present the content of the bill texts: Using NEREx (El-Assady et al., 2017), a framework from Visual Analytics, we shed light on named entities and other important content words relevant in a particular bill.

The paper proceeds as follows: Section 2. presents related work and Section 3. describes our newly developed corpus CoUSBi (Corpus of US Bills) and discusses the challenges related to its development. In Section 4. we show how the standardized encoding of the bills allows us to make reference to other information, in particular the model bills of ALEC. Section 5. presents the visualization that we propose to use for displaying content information in the Legal Knowledge Graph. The paper closes with a discussion in Section 6..

2. Relevant work

Text mining in the legal domain is as varied as the type of data underlying it, ranging from extracting and analyzing arguments in legal cases (Moens et al., 2007; Wyner et al., 2010) to summarizing legal documents (Farzindar and Lapalme, 2004; Grover et al., 2003; Galgani et al., 2012) and constructing knowledge resources (Francesconi et al., 2010; Ajani et al., 2010, inter alia). Other efforts mine

[1] `http://www.tei-c.org/index.xml`
[2] https://www.alec.org/
[3] Available under `https://www.alec.org/`

legal terms (Pala et al., 2010; Surdeanu et al., 2010) or named entities (Quaresma and Goncalves, 2010; Dozier et al., 2010) in legal text. Another strand of research is concerned with automatic reasoning on legal text, bridging the gap between law and artificial intelligence (Hollatz, 1999; Bench-Capron and Sartor, 2003, among many others).

With respect to standardizing sources of law, the MetaLex initiative (European Committee for Standardization, 2010) has been at the forefront of providing an XML-based interchange format, with for example all Dutch regulations published in this format. In the humanities, the Text Encoding Initiative (TEI) standard is widely used to encode a wide variety of textual data.

With respect to these previous approaches, our work touches upon different aspects. Firstly, we create a structured resource from semi-structured US state bills and discuss the key characteristics that need to be encoded for this type of legislative data. The TEI standard we employ for this effort allows us to link information across different sources e.g. across the model legislation, a prerequisite for eventually contributing information to the knowledge graph. Lastly, the visualization tool can offer us insights in the content and relations of the corpus provided.

3. CoUSBi

3.1. Data collection

For now, CoUSBi consists of all enacted, legislative bills related to education between 2007 and 2015 from two US states, namely North Carolina and New Mexico. Both states offer their bills in a semi-structured and machine-readable format (in contrast to other states which only give the bill text as image or pdf). Creating a corpus of the enacted education-related bills (the rejected bills were irrelevant for the social science aims of the project), turned out to be difficult because such filtering was not catered for. We therefore invested a substantial amount of manual work in extracting the IDs of all enacted bills and then used $crawler4j$, an open source Java web Crawler,[4] to scrape the bills automatically. For now, the creation of the corpus depends on the painstaking task of HTML scraping which can be hard to maintain on a long-term. For the future, a systematic effort could create an API through which the states can directly deliver the bills. This is not meant to be extra work for the states: the present forming of the HTML structure also requires time to split the information of the bill in the corresponding HTML elements. An API could be a more user-friendly way of submitting this information.

The resulting resource consists of a total of 2,599 bills, with the actual text of each bill having an average of 3,257 tokens. We also include the different versions of the bills before their enactment and mark them accordingly in the file name (e.g. 'v1' for version 1). The bills have an average of 3.9 versions, with a maximum of 14 versions. The entire corpus is made available under `https://github.com/kkalouli/CoUSBi`.

[4]Available under `https://github.com/yasserg/crawler4j`

3.2. Encoding in TEI

CoUSBi is encoded in XML according to the TEI standard, a format widely used in the NLP community. Although TEI has not been designed specifically for legal text (in contrast to the aims of the MetaLex initiative, for example), it proves capable of handling the legislative bill data well, both with respect to the metadata and the actual text of the bill. For now we restrict ourselves to encoding the resource in the TEI standard, however a conversion to the MetaLex standard should be unproblematic.

As can be seen in the simplified XML overview in Figure 1, we need the following TEI elements to encode metadata and content structure in TEI (the full XML schema can be found within the resource): The header of the TEI header contains the element `fileDesc` which itself has three mandatory elements (`titleStmt`, `publicationStmt` and `sourceDesc`) and one optional element (`editionStmt`). Each of those elements contains a series of mandatory and optional subelements.

For encoding the body of the bill document, we use the `body` element with different subelements that specify the particular structure of the document. All in all, the following elements are included in each bill document:

- the short title of the bill (element: `<title>`)
- the authors of the bill: the representatives who took part in the writing process (element: `<author>`)
- the edition of the bill (element: `<edition>`)
- the publication place: the state the bill was presented in (element: `<pubPlace>`)
- the ID number: a combination of the state, year and bill number of the bill, e.g. NM-2013-S039 stands for the Senate Bill (S) with the number 039 of the state New Mexico (NM) and the year 2013 (element: `<idno>`)
- the source link: the url from which the bill originates (element: `<bibl>`)
- a short abstract which gives information on what the bill is about (element: `<head>` of the element `<div1 type="abstract">`)
- the text of the bill itself, separated into sections and paragraphs, according to the original sections and paragraphs
- further highlighting for underlined and strike-through parts of the bills: some bills have various versions (editions) and therefore some parts of them are either underlined to represent new parts or struck-through to represent parts that were removed from a later version. Since this information is important for the history of a bill, it is preserved and encoded in the TEI format.
- extra annotation whether a specific part of the bill is identical to a passage from a model bill (element: `<cit>` with subelements `<quote>` to hold the identical passage and `<bibl>` to hold the ID and section of the model bill it is identical to).

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader xml:lang="en">
    <fileDesc>
      <titleStmt>
        <title xml:id="NC-2007-H15">Txtbks Assignmts on Short-Term Suspension.</title>
        <author>Representatives Glazier, E. Warren, Parmon, and  Johnson (Primary Sponsors). </author>
      </titleStmt>
      <editionStmt>
        <edition>v0</edition>
      </editionStmt>
      <publicationStmt>
        <pubPlace>North Carolina</pubPlace>
        <idno>NC-2007-H15</idno>
        <date>2007</date>
      </publicationStmt>
      <sourceDesc>
        <bibl>www.ncleg.net\Sessions\2007\Bills\House\HTML\H15v0.html</bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text xml:lang="en">
    <front>
      <div1 type="abstract">
        <head>A bill to be entitled an act to implement a recommendation of the house select committee on the education of
        students with disabilities to allow students placed on short-term suspensions to take their textbooks home for the
        duration of the short-term suspension and to have access to homework assignments.</head>
      </div1>
    </front>
    <body>
      <div1>
        <opener>
          <salute>The <term type="NE" subtype="ORGANIZATION"> General Assembly of North Carolina </term> enacts:</salute>
        </opener>
        <div2 type="section" n="1">
          <head>SECTION 1. G.S.</head>
          <p>(b)The principal of a school, or his delegate, shall have authority to suspend for a period of <term type="NE" subtype="DURATION">
          10 days </term> or less any student who willfully violates policies of conduct established by the local board of
          <hi rend="strikethrough">education: Provided, that a</hi><hi rend="underlined">education.</hi>student suspended
          <hi rend="strikethrough">pursuant to</hi><hi rend="underlined">under</hi>this subsection shall be provided<hi rend="strikethrough">
          an opportunity to take any <term type="NE" subtype="SET"> quarterly </term>, semester or grading period examinations missed during the
          suspension period.</hi><hi rend="underlined">all of the following:</hi></p>
          <p>
            <hi rend="underlined">(1)</hi>
            <hi rend="underlined">The opportunity to take textbooks home for the duration of the suspension.</hi>
          </p>
          <p>
            <hi rend="underlined">(2)</hi>
            <hi rend="underlined">The right to inquire about homework assignments for the duration of the suspension.</hi>
          </p>
          <p>
            <hi rend="underlined">(3)</hi>
            <hi rend="underlined">The opportunity to take any <term type="NE" subtype="SET"> quarterly </term>, semester, or grading period examinations missed during the suspension period.</hi>
          </p>
        </div2>
        <div2 type="section" n="2">
          <head>SECTION 2. This act is effective when it becomes law.</head>
        </div2>
      </div1>
    </body>
  </text>
</TEI>
```

Figure 1: A simplified structure of the TEI-formatted bill of North Carolina H15 (version 0).

3.3. Automatic conversion

The conversion from HTML into the TEI XML format is done automatically with a rule-based system designed on the basis of the scraped HTML pages. The structure of the original webpages differs markedly across states, therefore each state requires its own conversion script. Concerning metainformation, we benefit from the fact that the file name of each bill consistently encodes the date, the bill number, the bill version and the state. Those are straightforwardly converted into their corresponding TEI elements. For the other TEI header elements, namely the representatives of each bill, the short title and the abstract, we use patterns in the bill text to extract the information. For example, in the North Carolina bills the representatives' names can be extracted by looking for the lexical pattern *Representative* or *Sponsor*, while in the New Mexico bills the information is encoded via the pattern *Introduced by*.

The body of each TEI document corresponds to the main body of the original bill. For transforming the paragraph elements of the HTML files, we apply straightforward conversion of the HTML element to the TEI element. So, excerpt (1) from a North Carolina HTML is converted to the TEI paragraph in (2) by moving the `<p>` element one-to-one and converting the `<s>` and `<u>` elements to the TEI elements `<hi rend="strikethrough">` and `<hi rend="underlined">`, respectively.

In order to include information whether parts of the text correspond to model bills of ALEC (for more information see Section 4.) we add the `<cit>` element that captures the matching parts.

(1)
```
<p class=amargin1>(b) The principal of a school, or his delegate, shall have authority to suspend for a period of 10 days or less any student who willfully violates policies of conduct established by the local board of <s>education: Provided, that a </s><u>education. A </u>student suspended <s>pursuant to </s><u>under </u>this subsection shall be provided <s>an opportunity to take any quarterly, semester or grading period examinations missed during the suspension period.</s><u>all of the following:</u></p>
```

(2)
```
<p>(b)The principal of a school, or his delegate, shall have authority to suspend for a period of <term type="NE" subtype="DURATION"> 10 days </term> or less any student who willfully violates policies of conduct established by the local board of<hi rend="strikethrough">education: Provided, that a</hi><hi rend="underlined">education. </hi>student suspended<hi
```

```
rend="strikethrough">pursuant
to</hi> <hi rend="underlined">under
</hi>this subsection
shall be provided<hi
rend="strikethrough">an
opportunity to take any <term
type="NE" subtype="SET"> quarterly
</term>, semester or grading
period examinations missed during
the suspension period.</hi><hi
rend="underlined">all of the
following:</hi></p>
```

3.4. Named entity annotation

For tagging named entities, we use the Stanford Named Entity Recognizer (Finkel et al., 2005), and conclude that further model training is not necessary. The text of each bill section was fed into the recognizer, and a script parsed the output to add XML tags around the entities identified. Tags for *NUMBER* were disregarded, as their extremely high frequency led to so many labels in section headings as to be no longer of use. In excerpt (2) above we can see how the named-entity *10 days* is marked with the element <term> of the type *NE* (for Named-Entity) and the subtype *DURATION*.

3.5. Challenges

The main challenge in creating CoUSBi was to convert inconsistent information of the source into a standardized format. There were several types of inconsistencies. Firstly, there were formatting inconsistencies in the HTML encoding of bills within one state. For example, some bills would be encoded in CSS, while other bills contained a mixture of CSS and HTML, but encoding the same information.

A second group of inconsistencies was the incomplete information in the source. This means that not all bills within one state encoded the same kind of information, i.e. they did not include the same document elements. This inconsistency is tightly bound to a third one, namely the misplacement of some of the relevant information. Many of the bill documents contained the relevant information but not always in the same position within the document. This meant, for example, that although in most of the bills the date information was found after the title, in some of them we had to look for the date elsewhere. We also observed that many smaller details relevant for the task were not consistent, e.g. the use of capital or small letters for specific elements.

As a consequence of these issues and also because some of them were so profound, we have so far not converted the bills of West Virginia — an additional state on which we had started working — into the TEI format, as this requires further extensive manual effort. As it is, conversion of the bills from the other two states has already been very time-consuming. Although the issues mentioned can be solved with relatively simple, additional rules, e.g. for the inconsistency of small-capital letters we can use case-insensitive rules, it is tedious to make sure that all such inconsistencies have been traced and handled. Although we cannot provide a formal evaluation as to the completeness and accuracy of the resource, we believe that our repeated attempts to detect and handle all inconsistencies have paid off in a way that there is no missing information.

4. Linking information

One of the defining characteristics of knowledge graphs is that they link information from different sources. Despite a comparatively preliminary size and coverage of CoUSBi, we are nevertheless able to make interesting connections and comparisons and show that even preliminary investigations shed light on legislative processes as a whole.

In the US, many different organizations produce and distribute draft legislation which can be adopted and adapted by the concerned authorities. One such organization is the American Legislative Exchange Council (ALEC), with members including politicians as well as corporate representatives. Together they produce model bills that can then be directly introduced for debate in state legislatures (Hertel-Fernandez, 2014). Some states such as Arizona, Wisconsin, Colorado, Michigan, New Hampshire, and Maine make heavy use of the ALEC model bills (Rizzo, 2012). Approximately 200 ALEC bills become law each year (Greenblatt, 2003). As part of our work on CoUSBi, we investigated whether any bills introduced in state legislatures include influences from ALEC bills.

Preparing the ALEC bills All education model bills were scraped from the ALEC website, where they are freely available. There were 74 education bills in total posted on the ALEC website at the time of access, with dates ranging from 1995 to 2017. Some model bills did not include a date. It is not immediately clear in every case whether these dates reflect the online publishing date, or the date they were originally drafted. We also accept that this time frame both predates and extends beyond the dates of the bills collected in our corpus. However, model legislation that is several years old is by no means past its shelf life, and legislation introduced by a state may be then copied and distributed as model legislation, a relationship which would also be of interest.

Once scraped, each piece of model legislation was converted to the TEI standard consistent with the elements listed above for state legislation. Although most of this information is not available for the model bills, keeping these elements consistent will facilitate future analysis. Further, automatic annotation of bill sections was possible, providing important reference points for comparison with state bills. In all, this process produced a second smaller TEI corpus of ALEC education bills.

Linking ALEC bills and state bills In order to determine sections of ALEC bills which provided relevant matches to passages of bills introduced in the state legislatures, simple 15-grams from the text of each piece of model legislation were searched for in the text coming from the bills. In order to aid matching, the text of the bills was stripped of punctuation, and case was ignored. Further matching was able to determine which passages of the bills matched the passages in model legislation, which could then be automatically annotated. The annotation uses the <cit> element of the TEI format, which features the

North Carolina	ALEC
The Founding Principles Act, H588v0 & v1 (2015)	The Civic Literacy Act (No date)
Whereas, the adoption of the Declaration of Independence in 1776 and the signing of the United States Constitution in 1787 were seminal events in the history of the United States, the Declaration of Independence providing the philosophical foundation on which the nation rests, and the Constitution of the United States providing its structure of government; and Whereas, the Federalist Papers embody the most eloquent and forceful argument made in support of the adoption of our republican form of government; and Whereas, these documents, along with the writings of the Founders, stand as the foundation of our form of democracy, providing at the same time the **touchstone** of our national identity and the vehicle for orderly growth and change; and Whereas, these Founding Documents established a set of principles, known as the Founders' Principles, which are the heart and soul of a government for a free society; and Whereas, these principles enabled a group of 13 colonies to become the greatest and most powerful nation on earth in a relatively short period of time; and Whereas, most Americans do not know about nor understand the timely and timeless importance of these principles to our form of government and to their current lives; and Whereas, the survival of the republic requires that our nation's children, the future guardians of its heritage and participants in its governance, have a **clear understanding** of these principles and the importance of their preservation;	(A) The adoption of the Declaration of Independence in 1776 and the signing of the United States Constitution were principal events in the history of the United States, the Declaration of Independence providing the philosophical foundation on which this nation rests and the Constitution of the United States providing its structure of government. (B) The Federalist Papers embody the most eloquent and forceful argument made in support of the adoption of our republican form of government. (C) These documents stand as the foundation of our form of democracy providing at the same time the **basis** of our national identity and the vehicle for orderly growth and change. (D) Many Americans lack even the most basic knowledge and understanding of the history of our nation and the principles set forth in the Declaration of Independence, codified in the Constitution and defended in the Federalist Papers. (E) The survival of the Republic requires that our nation's children, the future guardians of its heritage and participants in its governance, have a **firm knowledge and understanding** of its principles and history.

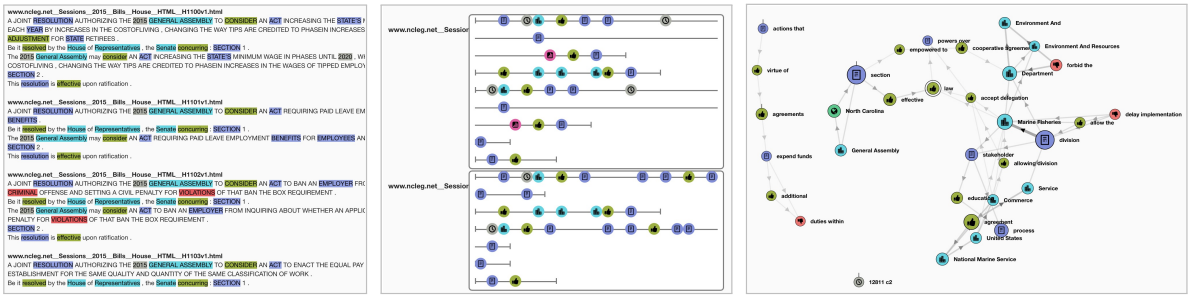
Figure 2: In the example above, green highlighted sections are identical, and closer reading reveals other sections, highlighted in teal, that are highly similar. Bold words indicate differences in otherwise similar passages.

subelements <quote> and <bibl>. The <quote> element contains the passage that is taken from the model bill and the <bibl> element specifies the ID and the exact section of the matching model bill.

Results The results of this analysis found that 13 non-overlapping verbatim spans of 15 words or longer from model bills were also found in North Carolina state education bills during this time period, and 10 portions were found in New Mexico's state bills. These spans ranged from 17 to 36 words in North Carolina's bills and 16 to 36 words in New Mexico's bills. The length of the n-gram threshold helped to ensure the retrieval of relevant similarities. Some passages seemed to offer formulaic speech for a definition and contained no resemblance in language or structure in the surrounding context. Other passages revealed that the sections preceding and/or following the verbatim passage contained multiple slight alterations that did not alter the meaning of the text, but which did prevent our method from identifying it as an extended block of verbatim text.

Such methods "will never replace careful and close reading of texts" (Grimmer and Stewart, 2013), and indeed this method can be of most utility in flagging sections for further examination, an example of which is shown in Figure 1. Examining the dates associated with the documents shows that only a small proportion of the matching state bills were introduced after the model legislation's reported date on ALEC's website (4 of 23 passages). However, this does not preclude the possibility that the passages in the state legislation were written by ALEC, or another organization. Previous study into ALEC has depended on internally leaked documents, as the group strives to keep a low profile (Hertel-Fernandez, 2014), and it is known to operate by distributing legislation to lawmakers without making this process public. Furthermore, the extent to which this language is mirrored between sources suggests that there is some link between the source of these passages.

This type of analysis can be helpful in analyzing legislation to identify similarities between states and other entities, and how influence may manifest itself in shorter passages, even if the full bill does not completely adhere to the goals of model legislation. Other methods such as fuzzy string matching and Levenshtein edit distance calculations could be employed in order to find matches with subtle differences, and to gauge whether the similarities are meaningful or coincidental.



(a) Text-Level View (b) Entity-Level View (c) Entity Graph

Figure 3: Named-Entity Relationship Explorer

5. Visualization

One of the core aims of encoding information in a knowledge graph is to present this information at a glance. In the case of a large amount of bills, a crucial component is to give the user an idea of the content of the bill, i.e. provide an intuitive overview of important concepts and entities that are covered in the bill. To this end we employ NEREx (El-Assady et al., 2017), a Visual Analytics framework for the analysis of different concepts and their relation in the utterances. Using Visual Analytics for this task is motivated by the challenge of dealing with large amounts of data, while at the same time providing the user with an interactive and exploratory access to the data (Keim et al., 2008).

The data is uploaded through a web interface and relevant named-entities and concepts from the text are categorized into ten classes: Persons, Geo-Locations, Organizations, Date-Time, Measuring Units, Measures and Context-Keywords. Using a perceptually preattentive visual encoding for these categories, the text is abstracted from the Text-Level View (Figure 3a) to the Entity-Level View (Figure 3b) to allow a high-level overview of the entity distribution across utterances.

For extracting relations in the text, the framework uses a tailored distance-restricted entity-relationship model, which relates two entities if they are present in the same sentence within a small distance window defined by a user-selected threshold. The concept map of the conversations can then be explored in the Entity Graph (Figure 3c). All views support a rich set of interactions, e.g., linking, brushing, selection, querying and interactive parameter adjustment.

The visualization supports the analyst in two ways: First, the *content of the bill* can be displayed with increasing abstraction, catering for different demands of the analyst (from close reading to distant reading). The Entity Graph gives an overview over highly relevant terms: The more saturated the colors of the arcs, the more frequent the nodes (i.e. entities/concepts), with the direction of the arc showing the order of the items.

For illustrative purposes, we use NEREx to display the content of only one bill, namely the 2015 bill S140 from the North Carolina Senate, which authorizes the town of Lake Santeetlah to levy an occupancy tax. Figure 4a shows the Entity Graph for the complete bill, the subgraphs in Figures 4b and 4c zoom in on the upper and lower middle part of the overall graph, respectively. In subgraph 1, the terms ‘Tourism’ and ‘Authority’ are at the center, with the former connected to the bigrams ‘Town ,’ and ‘Town .’, which in turn co-occur with ‘attract tourists’ and ‘expenditures’. Consulting the Text-Level View shows that the bill specifies the way the Tourism Authority spends the tax revenue: By funding tourism-related expenditures and support tourism-related activities. Subgraph 2 shows all cities belonging to the Lake Santeetlah Council that levy an occupancy tax – these cities are plotted on the canvas according to their geolocations. These examples show that an analyst can use the Entity Graph to get an immediate overview of the content of the bill, with the Text-Level View allowing for a more detailed investigation of the actual text.

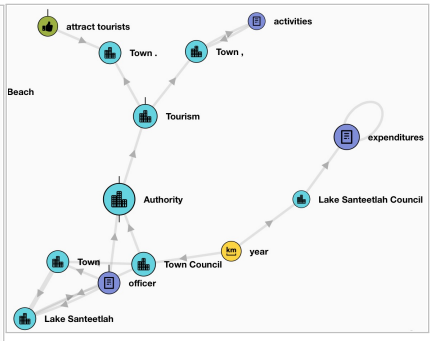
The visualization is also important for *resource validation*: As we were going through the visualization of individual bills, the Entity-Level View showed that in a small number of cases the bill content was repeated, based on inconsistencies in the source. In other cases, the bill text was blank, also due to erroneous HTML encoding in the source. These errors were then manually corrected in the corpus.

6. Conclusion and future work

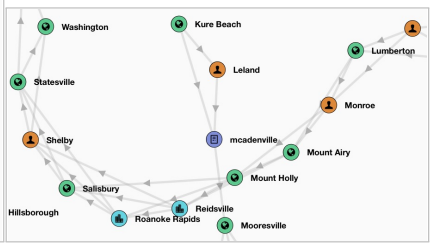
This paper reported on an approach of mining legal data in the wild and the requirements, challenges and potentials that go with it. Such a resource can be part of the Legal Knowledge Graph and can be used by professionals in the legal domain to examine and monitor the lawmaking process, from influence, to drafting, to editing, and the passage into law. Creating this type of structured resource also lays the groundwork for other research in the social sciences. One concrete application in political economy uses education bills to determine how elected officials respond to the preferences of their constituency. When the school performance of students in their electoral district weakens, do they respond by authoring a particular kind of education bill? How is the support for a bill in the legislature affected by additions or removals of certain clauses? To what extent do bill authors make policy tradeoffs between the preferences of his constituency and the preferences of the opposition? Such questions can only be answered using information from corpora such as CoUSBi.

Due to its consistent encoding in TEI, CoUSBi can also be utilized as-is for further syntactic and semantic parsing or can be indexed and used for direct query processing. It is also — to the best of our knowledge — the first attempt to automatically compare US bills to model legislation. The identical language in passages suggests a connection which

(a) Content of bill S140 from North Carolina



(b) Subgraph 1 of Figure 4a



(c) Subgraph 2 of Figure 4a

merits further analysis. Examining content on this level requires an extremely labor-intensive effort for human readers and the automatic method presented in this paper illustrates just one technique which could prove valuable to this end. As the corpus is expanded to include further legislatures and more model legislation, this type of research could be expanded, providing the public with its own measure of such organizations. Other research into paraphrasing, as well as other text matching methods could help to identify corresponding sections between model legislation and bills proposed in the states. Thus far, this information has only been available through painstaking reading through several bills. While this is only a first step, this kind of monitoring becomes possible with the advent of standardized and openly accessible legislative corpora.

Besides our goal to include more bills across states and topics, we also aim at implementing a framework that automatically compares different versions of the same bill and offers some insights on the types of changes between different versions. This will include, for example, an at-a-glance overview of sections withdrawn from or added to bills or highlight those that were only slightly modified. We also see a potential in applying topic-modeling techniques to the corpus in order to annotate individual bills with their key topics. We would additionally like to make use of the full potential of NERs by linking entities of different bills to each other and to the LOD [5] cloud, in this way making a vast amount of knowledge accessible. We furthermore consider the conversion of the TEI-formatted resource into the MetaLex standard, also to facilitate a comparison of legislation across languages and traditions.

[5]Linking Open Data, available under `http://lod-cloud.net/` and `http://linkeddata.org/`

7. Bibliographic References

Ajani, G., Boella, G., Lesmo, L., Martin, M., Masszei, A., Radicioni, D. P., and Rossi, P. (2010). Multilevel legal ontologies. In *Semantic Processing of Legal Texts*, pages 136–156. Springer: Berlin Heidelberg.

Bench-Capron, T. and Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143.

Consortium, T. (2017). Tei p5: Guidelines for electronic text encoding and interchange.

Dozier, C., Kandadadi, R., Light, M., Vachher, A., Veeramachanenin, S., and Wudali, R. (2010). Named entity recognition and resolution in legal text. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 27–43. Springer: Berlin Heidelberg.

El-Assady, M., Sevastjanova, R., Gipp, B., Keim, D., and Collins, C. (2017). Nerex : Named-entity relationship exploration in multi-party conversations. In Jeffrey Heer, editor, *EuroVis 2017 Eurographics / IEEE VGTC Conference on Visualization 2017*, number 36,3 in Computer Graphics Forum, pages 213–225.

Farzindar, A. and Lapalme, G. (2004). Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop*, pages 27–34.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics ACL2005*, pages 363–370. Association for Computational Linguistics.

Francesconi, E., Montemagni, S., Peters, W., and Tiscornia, D. (2010). Integrating a bottom-up and top-down methodology for building semantic resources for the

multilingual legal domain. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 95–121. Springer: Berlin Heidelberg.

Galgani, F., Compton, P., and Hoffmann, A. (2012). Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (Hybrid2012), EACL 2012*, pages 115–123.

Greenblatt, A. (2003). Governing.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Grover, C., Hachey, B., and Korycinski, C. (2003). Summarising legal texts: Sentential tense and argumentative roles. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 33–40.

Hertel-Fernandez, A. (2014). Who passes business's "model bills"? policy capacity and corporate influence in us state politics. *Perspectives on Politics*, 12(3):582–602.

Hollatz, J. (1999). Analogy making in legal reasoning with neural networks and fuzzy logic. *Artificial Intelligence and Law*, 7(2):289–301.

Keim, D., Andrienko, G., Fekete, J.-D., Górg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, et al., editors, *Information Visualization*, pages 154–175. Springer, Berlin.

(2010). Metalex (open xml interchange format for legal and legislative resources).

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law. Association for Computing Machinery*, pages 225–230.

Pala, K., Rychl'y, P., and Smerk, P. (2010). Automatic identification of legal terms in czech law texts. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 83–94. Springer: Berlin Heidelberg.

Quaresma, P. and Goncalves, T. (2010). Using linguistic information and machine learning techniques to identify entities from juridicial documents. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 44–59. Springer: Berlin Heidelberg.

Rizzo, S. (2012). Some of christie's biggest bills match model legislation from d.c. group called alec.

Surdeanu, M., Nallapti, R., and Manning, C. (2010). Legal claim identification: Information extraction with hierarchically labelled data. In *Proceedings of the 3rd Workshop on Semantic Processing of Legal Text, co-located with LREC 2010*, pages 22–29.

Wyner, A., Mochales, R., Moens, M.-F., and Milward, D. (2010). Approaches to text mining arguments from legal cases. In E. Francesconi et al., editor, *Semantic Processing of Legal Texts*, pages 60–79. Springer Berlin Heidelberg, 01.