Digitization of the Canadian Parliamentary debates

Beelen, K.; Alberdingk Thijm, T.; Cochrane, C.; Halvemaan, K.; Hirst, G.; Kimmins, M.; Lijbrink, S.; Marx, M.; Naderi, N.; Rheault, L.; Polyanovsky, R.; Whyte, T.

DOI
10.1017/S0008423916001165

Publication date
2017

Document Version
Final published version

Published in
Canadian Journal of Political Science

RESEARCH NOTE

Digitization of the Canadian Parliamentary Debates

KASPAR BEELEN *University of Amsterdam*
TIMOTHY ALBERDINGK THIJM *University of Toronto*
CHRISTOPHER COCHRANE *University of Toronto*
KEES HALVEMAAN *University of Amsterdam*
GRAEME HIRST *University of Toronto*
MICHAEL KIMMINS *University of Toronto*
SANDER LIJBRINK *University of Amsterdam*
MAARTEN MARX *University of Amsterdam*
NONA NADERI *University of Toronto*
LUDOVIC RHEAULT *University of Toronto*
ROMAN POLYANOVSKY *University of Toronto*
TANYA WHYTE *University of Toronto*

Introduction: Heritage and the Web of Data

Over the past years, a growing number of digitization projects have archived many forms of social, cultural and political heritage, thereby

Kaspar Beelen, Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam, 1098 XH, email: k.beelen@uva.nl
Christopher Cochrane, Department of Political Science, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, email: christopher.cochrane@utoronto.ca
Graeme Hirst, Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario, M5S 3G4, email: gh@cs.toronto.edu
Maarten Marx, Informatics Institute, University of Amsterdam, Science Park 904, Amsterdam, 1098 XH, email: maartenmarx@uva.nl
Nona Naderi, Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario, M5S 3G4, email: nona@cs.toronto.edu
Ludovic Rheault, Department of Political Science, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, email: ludovic.rheault@utoronto.ca
Tanya Whyte, Department of Political Science, University of Toronto, 100 St. George Street, Toronto, Ontario, M5S 3G3, email: tanya.whyte@mail.utoronto.ca

Canadian Journal of Political Science / Revue canadienne de science politique
50:3 (September / septembre 2017) 849–864 doi:10.1017/S0008423916001165

generating a torrent of electronic resources on which social scientists and historians can draw. Examples of such large-scale projects include the Universal Digital Library, Google Books and the National Digital Newspaper Program.[1] This revolution in data access was further propelled by advances in machine intelligence and computing power, which have helped to increase not only the volume but also the complexity of the digitized material accessible to researchers.

Many of these projects adhere to the principles of the Semantic Web, a global framework for organizing and disseminating data. The conception of this idea dates back to the seminal paper of Berners-Lee and colleagues (2001), which proposes to integrate data into a single global cloud and mould them into a format that allows machines to meaningfully manipulate and process information.[2] Even though the paper's initial vision was more ambitious than its realization to date, the ideas of Berners-Lee and colleagues have nevertheless spawned a comprehensive framework for standardizing and publishing data. The Open Data movement stands out as a key contributor to this project (Auer et al., 2007; Bizer et al., 2009), and comprises a growing number of initiatives led by the governments of democratic countries, for instance Canada and the United Kingdom.[3] For data to be "open," they must be released under a permissive license, remain freely available and allow prospective users to republish or transform them.

Within this context, the Digging into Linked Parliamentary Data (Dilipad) project—an interdisciplinary and international collaboration between researchers at the University of Toronto, the University of Amsterdam and the Institute of Historical Research (University of London)—set out to standardize, enrich and distribute the parliamentary proceedings of the Netherlands (1815–present), the United Kingdom (1803–present) and Canada (1901–present).[4] The project was funded by the Digging into Data Challenge (DiD), an international consortium of granting agencies seeking to promote the dissemination of data in the humanities and social sciences.[5] The goals of the Dilipad project were threefold: the first was to create a uniform, extensible format for the digitized records of parliamentary proceedings in Canada, the UK, and the Netherlands; the second was to facilitate analyses of these proceedings by researchers as well as by non-academic stakeholders such as activists, journalists and enthusiasts; the final goal was to leverage these new data to address substantive research questions about such topics as gender, ideology, immigration and the detection of emotion.

To date, the project has digitized and annotated the Canadian House of Commons proceedings in English, dating back to 1901. For a number of reasons, including difficulties associated with the optical recognition of Roman characters with accentuation, we are still working to reliably process the accompanying French translation of the debates, as well as all of the debates from the nineteenth century.

Abstract. This paper describes the digitization and enrichment of the Canadian House of Commons English Debates from 1901 to present. We start by laying out the general framework in which this project took place and then present the structure of the database and provide guidelines to prospective users. The paper concludes with the introduction of www.lipad.ca, an online platform designed as a hub for archiving Canadian political data, with the parliamentary proceedings at the centre of its architecture.

Résumé. Cet article décrit la numérisation et l'enrichissement de la publication parlementaire Débats de la Chambre des communes du Canada en langue anglaise, de 1901 à nos jours. Nous commençons par exposer le cadre général dans lequel ce projet s'est inscrit pour présenter ensuite la structure de la base de données et fournir des lignes directrices aux utilisateurs potentiels. L'article se conclut par la présentation de www.lipad.ca, une plateforme en ligne conçue pour être un carrefour d'archivage des données politiques canadiennes, avec les débats parlementaires au centre de son architecture.

Hansard in Canada

The etymology of the word *parliament*, from the French verb *parler*, underscores the importance of speech for parliamentary assemblies. The representational function of parliament plays out in the argument that legislators speak on behalf of citizens (Manin, 1997). The accountability function plays out when citizens use these speeches to judge the parliamentarians entrusted to represent them. Parliamentary speech, therefore, is a pillar of representative government.

Even so, no official transcript of parliamentary debates existed in Canada in the early years after Confederation. Initially, newspapers took up the responsibility by distributing fragmentary reports of parliamentary discussions. These were then published in what was commonly referred to as the "Scrapbook Hansard: a compilation of unofficial, incomplete and often biased transcripts, taken from several press sources" (Marleau and Montpetit, 2000). The House of Commons did not get a complete and authoritative transcript of the parliamentary debates until the Official Report was established in 1880 (Hansard Association of Canada, 2005)—which was then known as the "Hansard." From this year on, a thoroughly trained collective of reporters and amanuenses meticulously noted down every word uttered in the House.[6]

Without doubt, the Official Report stands out as the most comprehensive and complete corpus for studying political speech and decision making in Canada. It spans more than one century, and touches upon a wide range of social, economic and political issues that moved Canadian public opinion at some point in time. As a crucial piece of heritage, its contents are relevant to myriad scholars such as legal historians, political scientists and critical linguists—to name only a few.

The sheer size of the Hansard poses a challenge for researchers, however. It contains over 650 million words, plus translations of the text into the other official language. Few libraries contain the complete 148-volume collection, but even when they are fully available, browsing the proceedings remains a laborious task. It would take an average reader much more than a decade of uninterrupted reading—24 hours a day, seven days a week—to read through the Hansard. Querying its contents is a significant needle-in-haystack problem, notwithstanding the available indices, which, indeed, are detailed and helpful but nevertheless incomplete, since the categories reflect the priorities of the publisher at the date of publication rather than the priorities of researchers today.

Partly as a response to challenges such as these, parliaments around the world began digitizing their proceedings and releasing them online. In Canada, Internet publishing commenced in 1994, and by the early 2000s, the digitization of Hansard had evolved into an integrated data management system called Prism (O'Brien, 2002). Prism integrates all aspects of parliamentary business: bills, committee evidence and debate transcripts.[7] In 2011, the Canadian government launched the Federal Open Data Program,[8] which many enthusiasts, working outside Parliament, interpreted as an endorsement of their efforts to achieve greater transparency of government data (Kitchin, 2014; Milligan, 2014). These open data activists have also participated directly in the dissemination of government data. Michael Mulley, for example, founded Open Parliament (www.openparliament.ca), a platform assembled from freely available digitized parliamentary documents. He further enriched the data and moulded them into a comprehensive and easy-to-use search interface.[9]

Until very recently, the electronic record of Hansard encompassed only the past two decades of parliamentary debates. In 2013, however, the Library of Parliament scanned the entire historical proceedings from 1867 to 1999. The corresponding files were given to *Canadiana,* a non-governmental organization committed to the preservation of Canadian heritage, which in turn deposited the proceedings in an online digital archive of PDF files. Although this archive does not permit efficient and flexible queries and although it does not contain information about the structure of debates or the characteristics of the speakers, *Canadiana*'s archive was an indispensable starting point for our effort to produce an enriched, machine-readable, and highly searchable record of Hansard. It is to this effort that we now turn.

Data Conversion: From PDF to Linked Data

To make the proceedings truly machine-readable and manipulatable, we set out to semantically annotate the proceedings using eXtensible Markup

Language (XML). For annotation, we followed the Political Mashup (PM) guidelines,[10] whose metadata adhere to the Dublin Core standard and whose schema reuses modules from the Text Encoding Initiative (TEI) wherever possible (Ide and Veronis, 1995; Marx, 2009). In concise terms, the Political Mashup schema builds on the TEI guidelines for encoding theatrical data, which, in its stripped-down form, consists of the following components:

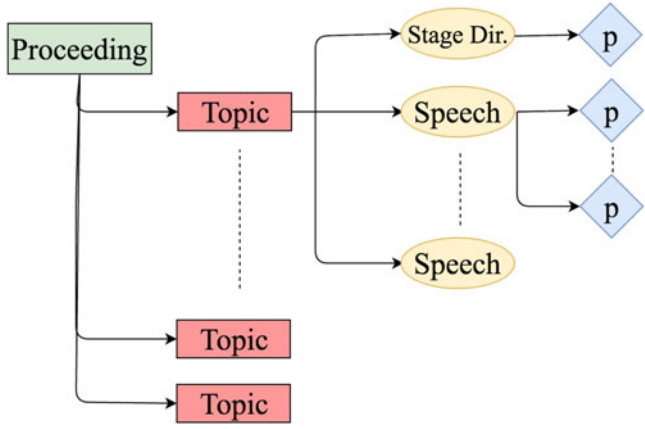
Topic: the points on the agenda.
Speech: uninterrupted string spoken by a member (in one topic element).
Stage-Direction: everything outside of the spoken word, such as procedural text and votes.
Paragraphs: text containers.

These elements fit within a hierarchy as shown in Figure 1 (Marx, 2009)[11]:

For the post-1994 proceedings (which were already available digitally) the transformation to PM annotations was relatively straightforward; it only involved writing code that converted data from one format (SQL) to another (XML). The source data were compiled from Michael Mulley's Open Parliament and transformed by Maarten Marx and Sander Lijbrink.[12]

Processing the historical Hansard to XML amounted, however, to a far more challenging task, and required several steps. In the first step, we converted PDF image files to "flat text" with optical character recognition

FIGURE 1
Simplified structure of the Political Mashup schema. The boxes represent the elements of the schema. The arrows show the nesting.



(OCR) software (ABBYY FineReader). On practical grounds, we chose to build our corpus from raw text, instead of starting with OCR conversions to XML or Rich Text Format. By doing otherwise we would have retained information on font size, font type and character position—valuable information indeed—but this would also have forced us to spend more energy and time on preprocessing: identifying page structure, filtering out noise, and so forth. Because of the project's rather limited timeframe, we chose to let the OCR engine handle these issues. This resulted in relatively clean output, but complicated some other aspects of the transformation, as we will explain below.

After having pushed the scans through the OCR engine, we parsed these unstructured documents. Fortunately, the raw text contained various textual and typographic cues, which permitted us to identify the building blocks of the parliamentary debates (speeches and topics). The example below shows how. (Speeches are truncated to make the fragment more presentable.[13])

DEFENCE EXPENDITURE
APPOINTMENT OF SPECIAL COMMITTEE
Mr. G. C. Nowlan (Annapolis-Kings): Mr. Speaker, I intend to intervene but briefly in his debate. […]
Mr. Jean Francois Pouliot (Temiscouata): Mr. Speaker, I do not intend to speak today only as a member of parliament or as a member of the Liberal party. […]

A distinct, albeit complicated, pattern flags the beginning of a new agenda point or speech turn. Topic titles appear as capitalized strings. Interventions are initialized by a specific pattern: a title (Mr., Ms. or Hon.) and a name followed by additional information (function and/or constituency) between brackets. The colon marks the end of the speaker string. Regular expressions (a pattern recognition technique) can detect these character sequences and annotate the corresponding text, thereby generating a first markup layer that encodes the starting point of each topic and speech. Due to the many OCR errors, these regular expressions had to be flexible and account for small deviations from the ideal pattern; otherwise recall would have dropped drastically.[14] The result for the example above is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
<proceedings>
  <topic_start>DEFENCE EXPENDITURE</topic_start>
  <topic_start>APPOINTMENT OF SPECIAL COMMITTEE
  </topic_start>
```

```
    <speech_start>Mr. G. C. Nowlan (Annapolis-Kings):
    </speech_start>
    <p>Mr. Speaker, I intend to intervene but briefly in his debate.</p>
    <speech_start>Mr. Jean Francois Pouliot (Temiscouata):
    </speech_start>
    <p>Mr. Speaker, I do not intend to speak today only as a member of
    parliament or as a member of the Liberal party. […]</p>
</proceedings>
```

After determining the scope of the detected elements—figuring out where they start and end—the parser can order topics and speeches and hence force a hierarchy on what was previously a vertically structured document. To this end, the parser only needed a few hard-coded procedures to perform tasks such as "If a topic title is detected, then create a topic element, and nest all following speeches within this topic element, until the next title appears." The result of this step is shown below. (The "pm:" prefix indicates that the element belongs to the Political Mashup namespace.)

```
<?xml version="1.0" encoding="UTF-8"?>
<pm:proceedings pm:id="ca.proc.d.1953-01-20">
  <pm:topic pm:id="ca.proc.d.1953-01-20.1"
  pm:title="DEFENCE EXPENDITURE">
    <pm:scene pm:id="ca.proc.d.1953-01-20.1"
    pm:title="APPOINTMENT OF SPECIAL COMMITTEE"
    pm:type="topic">
      <pm:speech pm:id="ca.proc.d.1953-01-20.1.1"
        pm:speaker="Mr. G. C. Nowlan (Annapolis-Kings):">
        <pm:p pm:id="ca.proc.d.1953-01-20.1.1.1">
              Mr. Speaker, I intend to intervene but briefly in his
              debate.
        </pm:p>
      </pm:speech>
      <pm:speech pm:id="ca.proc.d.1953-01-20.1.2"
      pm:speaker="Mr. Jean Francois Pouliot (Temiscouata):">
        <pm:p pm:id="ca.proc.d.1953-01-20.1.2.1">
          I do not intend to speak today only as a member of
          parliament or as a member of the Liberal party.
        </pm:p>
      </pm:speech>
    </pm:scene>
  </pm:topic>
</pm:proceedings>
```

The markup thus recovers the original flow of the debate. To solidify the hierarchy, we inserted document object identifiers (DOI) (coded as "id" in the Political Mashup namespace) that permanently anchor each element within the overall architecture of the corpus. These DOIs serve as Uniform Resource Identifiers (URIs) and indicate the fixed location of each component. Their format is compliant with the recommendations on publishing government data made by World Wide Web Consortium (W3C) (Alonso et al., 2009; Marx, 2009).

Adjusting the transformation script to distinct historical periods required only minor adaptations. Even though the Hansard corpus spans more than one century, the shape of the transcripts remained more or less stable, except for some minor typographic changes. The main reason to break off the conversion at 1901 was a drop in OCR quality, not a technical issue with the transformation software.

We amended the format to account for the idiosyncrasies of the Canadian Hansard. The PM schema comes with an open version to which anyone can contribute new elements. This was necessary, since the schema initially didn't fit the Canadian proceedings well. The topic structure proved more complex than those of the Dutch proceedings upon which the schema was modelled. But by simply adding a "subtopic" element (as part of the Dilipad namespace) we could accurately replicate the structure of the debate. The maximum nesting depth for topics was set to three levels. For example, the following topic sequence:

IRRIGATION
SOUTH SASKATCHEWAN RIVER
MOTION FOR ADJOURNMENT

was encoded as:

```
<?xml version="1.0" encoding="UTF-8"?>
<pm:topic pm:id="ca.proc.d.1953-01-20.1"
pm:title="IRRIGATION">
   <pm:scene pm:id="ca.proc.d.1953-01-20.1.1"
      pm:type="topic"
      pm:title="SOUTH SASKATCHEWAN RIVER">
      <dp:subtopic pm:id="ca.proc.d.1953-01-20.1.1.1"
      pm:title="MOTION FOR ADJOURNMENT">
      </dp:subtopic>
   </pm:scene>
</pm:topic>
```

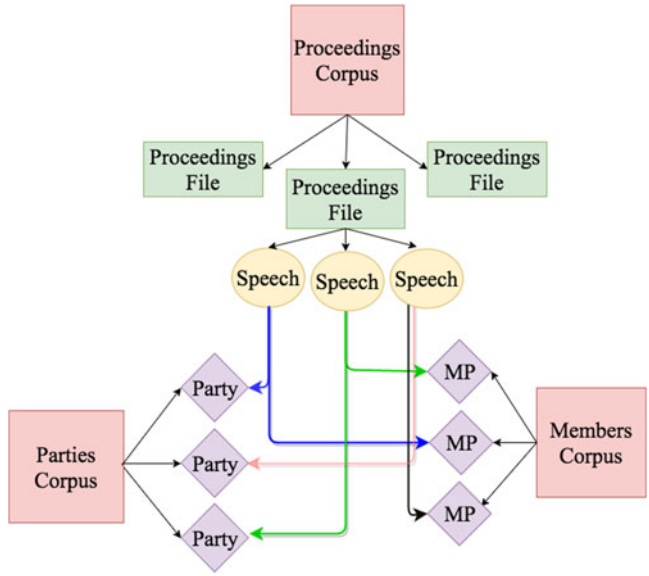
There remains, however, one important caveat: despite our attempts to preserve the original Hansard structure, this turned out to be impossible in

practice, especially for many of the earlier proceedings. As mentioned above, the OCR software converted all data to flat text, thereby discarding typographical cues such as font size and font type. Unfortunately, these elements contained some hints about topic nesting—with smaller fonts signalling the lower-level entities—that could not be extracted from the plain text.

Even though each of the preceding steps incrementally enriched the data, the corpus itself remained highly fragmented. Except for date and position, the elements of our corpus still missed internal and external links. In the final stage of the conversion we therefore associated speeches with biographical information. We determined *who*—which historical person—actually took the floor at a certain point in time, something which is of special interest to social scientists and historians, who often wish to relate language use to social and political categories. The addition of biographical information opens up a wide range of new research opportunities.

Before linking speeches to their authors, we created an authority list. Every MP who held a seat in the House of Commons between 1901 and 2015 received a stable and unique ID. These IDs connect all the speeches made by the same member, and serve at the same time as a pointer to an external biographical database.[15] Figure 2 illustrates the linked structure of the corpus.

FIGURE 2
Structure of the corpus. The squares show the separate corpora. The arrows show the links between the datasets.



Associating speeches with MPs involved disambiguation and OCR error correction. First, we clustered all speeches on the document level—which covers one sitting—by MP. In the Hansard text, a member's name is fully recorded at his or her first intervention (in addition to name and surname, this often includes constituency and portfolio) but is shortened afterwards, during the consecutive speech turns. We selected, for each cluster, the string that contained the most information and proceeded with matching this version against the entries in our biographical authority list. After identifying the components of the name string, we fed the parsed name to a multistage hierarchical matching procedure. As matching conditions become less stringent with each subsequent step, the confidence we have in the selected candidate (if there was any) varies significantly. At the top level, all relevant conditions (first name, last name, function and/or constituency) are matched, while in the later stages only one or two rather ambiguous elements happened to correspond to an entry in the member database. To keep track of these differences, we stored a measure of "speaker-link-confidence" as an attribute of each speech. Moreover, to avoid the linking procedure failing due to small OCR errors in the name string (which occurs rather frequently), the matching algorithm contains an inbuilt OCR flexibility system to handle cases in which the member database contains a clear match that is only one or two edits away from the name string as it appears in the proceedings.[16] Using this procedure we associated 98 per cent of all the speeches with an MP.[17]

```
<?xml version="1.0" encoding="UTF-8"?>
<pm:speech pm:id="ca.proc.d.1953-01-20.1.2"
   pm:speaker="Jean-François Pouliot"
   pm:member-ref="ca.m.9510"
   pm:party="Liberal"
   pm:party-ref="ca.p.liberal"
   pm:role="mp"
   dp:speaker-link-confidence="0.75">
   <pm:p pm:id="ca.proc.d.1953-01-20.1.2.1">
   I do not intend to speak today only as a member of
   parliament or as a member of the Liberal party.
   </pm:p>
</pm:speech>
```

As a result, the information appended to each intervention increased rather drastically. Most of the inserted attributes carry time-dependent information such as party-affiliation ("pm:party" and "pm:party-ref"), role description ("pm:role" which has values "MP," "chair" or "government"), or portfolio ("pm:function"). The "member-ref" attribute remains stable and identifies the author by pointing to the location of his or her

biographical summary in the authority list. We outsourced more extensive descriptions of a member's career and background to this controlled vocabulary. These individual profiles, which we downloaded from ParlInfo,[18] record political memberships (caucus, constituency and portfolio) and other personalia (gender, place of birth, education and profession). To further embed our data in the Web, we linked all members to their Wikipedia and DBpedia profiles. A short biographical description was appended to facilitate exploration.

After having completed the semantic annotation, we plan to insert the corpus into the broader Semantic Web by converting the XML data to the Resource Description Framework (RDF). RDFs unambiguously identify objects, such as people, locations, or abstract concepts, and describe the nature of their relation in the form of subject-predicate-object triples. Besides modelling the parliamentary debates as RDF triples, we will integrate it into other relevant vocabularies (FOAF,[19] DBpedia Ontology,[20] BIO[21]) (Tarasova and Marx, 2013).

Access and Distribution

Up to this point, we have focused on the distribution of the digitized corpus in a markup language. However, not all social scientists and historians have received the training required to process large quantities of XML data for applied research. To facilitate the exploration of the Canadian Hansard, we created an easy-to-use website: www.lipad.ca.[22] The interface has been designed to be modern, accessible and user-friendly for both researchers and the general public. On loading the home page, users are presented with a familiar Google-style search box (Figure 3). The search engine is built on the Solr search backend, and customization of a more complex search query with filtering is available on the Advanced Search section of the site.

A date timeline provides a quick and intuitive way to find specific debates and navigate across time. The complete dataset plus supplementary data is downloadable from the site in XML and PostgreSQL formats. Users can also custom export any date range or search result to a CSV file for offline study. Finally, the site's research blog allows us to communicate research results and highlight interesting uses of the data. Future development planned for the site includes interactive data visualization apps, expansion of the database to include French language debates, Senate and provincial-level debates, and aggregation of supplementary Canadian political data such as polling results and news.

For comparativists, the Political Mashup site offers tools to study the historical trajectory of the Canadian House of Commons in relation to parliaments of the United Kingdom or the Netherlands. The search interface can found at http://search.politicalmashup.nl.

FIGURE 3
Lipad.ca Home Page

Concluding Remarks

This paper has laid out a detailed account of the digitization of the Canadian Hansard since 1901, a large-scale project undertaken as part of the Digging into Data Challenge international programme. The resulting corpus promises to shed new light on a part of Canadian political history that has hitherto been inaccessible to many researchers. Reading and processing millions of speeches to uncover empirical patterns is beyond the capacity of human readers, at least not without unreasonable investments of time and labour. Computers, on the other hand, are ideal to perform such tasks, given data in a structured format that they can be instructed to process and retrieve. We believe that this corpus can be helpful not only for academic research but also for practitioners, the many people who work in public institutions who could use the database to gain insights into the history of perhaps the most important political institution in Canada. By adopting the same markup language and conventions used in the design of equivalent databases for Britain and the Netherlands, this project also provides opportunities for comparative research.

We believe that the social impact of this open and linked data project can be significant. The release of substantial quantities of digitized political texts opens the door to scientific research in a number of disciplines, for

Canadian and international scholars alike. First, the corpus represents a new source of data to conduct studies in the field of computational linguistics. Advances in natural language processing require access to large and reliable corpora from different domains. The integration of links to external data sources in our database implies that annotations can be used to train classifiers for various purposes, using methods of machine learning. Second, the digitized corpus has a concrete potential for new advances within the rest of the social sciences and humanities, in disciplines as varied as political science, communication, psychology, linguistics, history, economics and sociology. Political scientists, in particular, are interested in social actors who interact primarily with words, and having access to such an extended record of parliamentary debates removes an obstacle to empirical research on Canadian politics. Applied research using computational text analysis is currently burgeoning in political science, with examples ranging from the detection of ideology and partisan polarization in speech to structural topic modelling (for example, Diermeier et al., 2012; Grimmer and King, 2011; Monroe et al., 2008; Proksch and Slapin, 2010; Roberts et al., 2014; see Grimmer and Stewart, 2013 for a recent summary).

So far, these data and the companion website underpin at least three doctoral dissertations, five article manuscripts, and a number of presentations at academic conferences. By offering new possibilities that are of interest to both quantitatively oriented scholars and to those who prefer qualitative methods—for instance in the fields of political development and political history—it is our hope that the use of these new data will transcend methodological and disciplinary divides to support vibrant new research agendas about the Canadian Parliament.

Looking forward, completing a French version of good quality is a priority, as stated in the introduction. The bilingual Canadian Hansard has already gained an international reputation in computational linguistics due to the quality of its translation and the precise alignment of the two languages (see Brown et al., 1990, 1991, 1993; Fraser and Marcu, 2007). A historical corpus accessible in two languages would represent a useful source material, both for French-speaking researchers interested in Canadian politics and for the development of machine translation tools.

Endnotes

1 See, respectively, http://www.ulib.org/, https://books.google.com/, and https://www.loc.gov/ndnp/.

2 See Barbera, 2013; Bizer et al., 2009; Blanke et al., 2012; Meroño-Peñuela et al., 2012; Rademaker et al., 2015; Sztyler et al., 2014.

3 For Canada see http://open.canada.ca/en/open-data and for the United Kingdom, https://data.gov.uk/.

4 We decided to limit the scope of the database to the post-1901 period, mainly because the OCR quality was extremely poor prior to this period, which affected the overall

reliability of the conversion. Going back in time will require more work, but there are no insurmountable technical obstacles.

5 See http://diggingintodata.org/about. The Digging into Data Challenge Round Three was sponsored by NSERC and SSHRC.

6 Although these transcripts follow the original wording closely, they are not strictly verbatim (Slembrouck, 1992). The House of Commons rulebook describes Hansard as "a transcript *in extenso*. In the case of repetition or for a number of other reasons, such as more specific identification, it is acceptable to make changes so that anyone reading Hansard will get the meaning of what was said" (quoted in Hansard Association of Canada, 2005, from the sixth edition of *the House of Commons Procedure and Practice*. For the latest version see Marleau, and Montpetit (2000). Indeed, conveying meaning takes priority over replicating the exact wording. Nonetheless, Hansard is considered to be the *authoritative* record of words spoken in the House. When, for example, the then–Prime Minister Pierre Elliot Trudeau was asked if he mouthed the words *fuddle duddle* at his Progressive Conservative opponent John Lundrigan, he replied: "If the unimpeachable Hansard has noted that I said *fuddle duddle* then it must be so" (Hansard Association of Canada, 2005).

7 The latest debates can be accessed online at: http://www.parl.gc.ca/HouseChamberBusiness/ChamberSittings.aspx

8 http://open.canada.ca/en.

9 https://openparliament.ca.

10 See for other parliamentary databases: http://search.politicalmashup.nl. The schema is already in use for proceedings of seven other parliaments (Europe, the Netherlands, Flanders, Denmark, Sweden, Norway and the UK).

11 The details of the schema can be found on http://schema.politicalmashup.nl.

12 We used the SQL Data Dump and transformed it to a format, which is compliant with the Open PM schema. The latest version of the Open Parliament data can be found at https://openparliament.ca/data-download.

13 This example may be useful to illustrate how some typographical errors may have remained in the corpus, despite our efforts to correct the majority of them. The first speech from Hon. George Nowlan contains the words "his debate" instead of "this debate". This was caused by a misaligned scan of that specific page of the Hansard, which has blanked the first characters of the leftmost column.

14 We empirically fine-tuned the regular expressions against a random sample of topic titles and speaker entities.

15 See below for more information about the biographical database.

16 We used the Levenshtein distance metric to compute the similarity between two strings.

17 The vast majority of links going from speeches to MPs are correct, but not all of them. For this reason, we will allow users of Lipad.ca to report potential errors.

18 These data were derived from Parlinfo: http://www.lop.parl.gc.ca/ParlInfo/default.aspx?Menu=Home.

19 http://xmlns.com/foaf/0.1.

20 http://dbpedia.org/ontology/.

21 http://vocab.org/bio.

22 The *lipad.ca* website was created with Django, a Python web development framework. The dataset has been imported into PostgreSQL and combined with additional data from openparliament.ca and the Library of Parliament's PARLINFO database, to create a comprehensive and current digital collection of the Canadian Hansard.

References

Alonso, José, Owen Ambur, Miguel A. Amutio, Oscar Azañón, Daniel Bennett, Rachel Flagg, Dave McAllister, Kevin Novak, Sharron Rush and John Sheridan. 2009. "Improving access to government through better use of the web." World Wide Web Consortium.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives. 2007. "DBPedia: A Nucleus for a Web of Open Data." In *The Semantic Web: Lectures Notes in Computer Science 4825*, ed. Karl Aberer, Key-Sun Choi, Natasha Noy Dean Allemang, Kyung-Il Lee, Lyndon Nixon Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber and Philippe Cudré-Mauroux. Berlin: Springer.

Barbera, Michele. 2013. "Linked (open) data at web scale: research, social and engineering challenges in the digital humanities." *Journal of Law and Information Science* 4: 91–101.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." *Scientific American*, May 1, 1–5.

Bizer, Christian, Tom Heath and Tim Berners-Lee. 2009. "Linked data—the story so far." *International Journal on Semantic Web and Information Systems* 5: 205–27.

Blanke, Tobias, Gabriel Bodard, Michael Bryant, Stuart Dunn, Mark Hedges, Michael Jackson and David Scott. 2012. "Linked data for humanities research—The SPQR experiment." Paper presented at the *6th IEEE International Conference*, IEEE.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. "A statistical approach to machine translation." *Computational linguistics* 16: 79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. "Word-sense disambiguation using statistical methods." In *Proceedings of the 29th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics*, pp. 264–270.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics* 19: 263–311.

Diermeier, Daniel, Jean-François Godbout, Bei Yu and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42: 31–55.

Grimmer, Justin and Gary King, G. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108: 2643–50.

Fraser, Alexander and Daniel Marcu. 2007. "Measuring word alignment quality for statistical machine translation." *Computational Linguistics* 33: 293–303.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21: 267–297.

Hansard Association of Canada. 2005. "Tradition and Innovation, Celebrating 125 years of Hansard." Ottawa.

Ide, Nancy, and Jean Veronis, eds. 1995. *Text encoding initiative: Background and contexts*. vol. 29. Berlin: Springer Science & Business Media.

Kitchin, Rob. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Manin, Bernard. 1997. *The principles of representative government*. Cambridge: Cambridge University Press.

Marleau, Robert, and Camille Montpetit. 2000. *House of Commons Procedure and Practice*. http://www.parl.gc.ca/marleaumontpetit/ (December 1, 2015).

Marx, Maarten. 2009. "Advanced information access to parliamentary debates." *Journal of Digital Information* 10: 1–11.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Laurens Rietveld and Rinke Hoekstra. 2012. "Linked humanities data: The next frontier? A case-study in historical census data." In *Proceedings of the 2nd International Workshop on Linked Science*, Boston.

Milligan, Ian. 2014. "Open Data's Potential for Political History." *Canadian Parliamentary Review* 35: 34–43.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16: 372–403.

O'Brien, Audrey. 2002. "Prism: The House of Commons Integrated Technology Project." *Canadian Parliamentary Review* 25.

Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. "Position Taking in European Parliament Speeches." *British Journal of Political Science* 40: 587–611.

Rademaker, Alexandre, Dário Augusto Borges Oliveira, Valeria de Paiva, Suemi Higuchi, Asla Medeiros e Sá, and Moacyr Alvim. 2015. "A linked open data architecture for the historical archives of the Getulio Vargas Foundation." *International Journal on Digital Libraries* 15: 153–67.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58: 1064–82.

Slembrouck, Stef. 1992. "The parliamentary Hansard 'verbatim' report: the written construction of spoken discourse." *Language and literature* 1: 101–19.

Sztyler, Timo, Jakob Huber, Jan Noessner, Jaimie Murdock, Colin Allen and Mathias Niepert. 2014. "LODE: Linking digital humanities content to the web of data." In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE Press.

Tarasova, Tatiana and Maarten Marx. 2013. "ParlBench: A SPARQL Benchmark for Electronic Publishing Applications." In *The Semantic Web: ESWC 2013 Satellite Events*. Berlin: Springer.