DHQ: Digital Humanities Quarterly
2018
Volume 12 Number 2

Using structured text corpora in Parliamentary Metadata Language for the analysis of legislative proceedings

Richard Gartner <richard_dot_gartner_at_sas_dot_ac_dot_uk>, Warburg Institute, London, UK

Abstract

This article examines the potential of employing structured texts, encoded in the Parliamentary Metadata Language XML schema, for the machine-readable analysis of substantial corpora of legislative proceedings. It demonstrates the potential of using PML corpora for combining the results of sentiment analysis with contextual metadata to establish and visualise patterns of divergent attitudes towards a topic such as immigration as they correlate with such features as party affiliation or geographic location. This is readily achieved using such simple techniques as XSLT transformations or XQUERY searches.

Introduction

Large-scale digitisation projects in many countries over the last twenty years have generated significant machine-readable textual corpora of the proceedings of their parliaments or legislative assemblies. In the United Kingdom, for instance, a number of projects have enabled seven hundred years of records to be made available in digital form, including the full text of the official Hansard record of debates from 1803 [UK Parliament 2012]. These corpora are potentially highly significant resources for large-scale machine-based analysis, although at present studies of this type are rendered more difficult by the heterogeneity of approaches to markup and encoding taken by the projects that created them. 1

In recent years, a number of projects have attempted to address the issues raised by this heterogeneity and formalise schemes for rationalising approaches to the encoding of these texts and their associated metadata. In Portugal, for instance, Dublin Core (DC) is used as the basis for a digital library of parliamentary records [Pinto et al. 2005] and in Chile the National Library has produced an extended DC element set to act as the underlying descriptive metadata for its digitisation programme [Fuentes Martínez 2005]. A more ambitious approach has been undertaken by the Akomo Ntoso project for African legislative, parliamentary and judiciary documents [Cervone et al. 2015] which has created a formal XML schema and controlled vocabularies. 2

More recently, the LIPARM (Linking the Parliamentary Record through Metadata) project [Gartner 2012] attempted to define a more sophisticated methodology for parliamentary metadata by devising an XML schema, PML (Parliamentary Metadata Language) [Gartner 2013] that is intended to provide a means of semantically linking core facets of legislative proceedings into a single coherent framework. The schema was intended primarily as a resource discovery tool, allowing a diverse range of previously incompatible digitised resources to be conjoined using common semantic linkages and so become interoperable at a search level. The schema was tested on a corpus of UK Parliamentary materials and proved itself capable of allowing sophisticated searching and browsing of parliamentary records in a prototype interface to PML records constructed for the project [Gartner 2013, 32]. 3

These approaches are primarily aimed at resource discovery rather than the analysis of the textual content of proceedings of the type that has become established in fields such as corpus linguistics. Such studies have been relatively rare in the past: until the last five years, most historical research on parliamentary proceedings has tended to take the form of qualitative analyses based on manual readings. There appears to be have been considerable suspicion 4

of corpus-driven analyses of political vocabulary in particular; some argue that this stems from a prevalent post-structuralist emphasis in the historical community on the micro reading of texts rather than the analysis of broad trends [Blaxill 2014].

Despite these suspicions, more quantitative approaches have begun to emerge in recent years, the most sophisticated of which have been those conducted in the Netherlands using the Political Mashups schema, a TEI extension which is based on the latter's elements sets for theatrical performance [Gielissen and Marx 2009, 19]. This has been used, for instance, to examine attitudes expressed in the country's parliamentary proceedings over a lengthy time period towards a well-known party of the far-right [Piersma et al. 2014]. 5

Although not designed primarily as an analytical tool, the semantic linkages encoded within a PML file, when used in conjunction with the texts whose contents they synthesise, clearly offer the potential for moving beyond resource discovery alone towards something more analytical. Instead of merely allowing these texts to be found, it should allow them to become the raw material for machine-based analyses, using the mechanisms associated with a discipline such as corpus linguistics. The corpora on which this field of study relies integrate texts with metadata, usually the syntactical information associated with each of their constituent words. In a similar way, PML integrates data and metadata and so should be able to facilitate analyses of the same type. 6

This paper aims to demonstrate how PML can be used in this way; by analysing the content of parliamentary proceedings and using its associated metadata to put the results of these analyses into context, it becomes possible to generate correlations revealing high-level patterns within these texts. In this study, the tools used for this are relatively simple. Sentiment analysis tools are used on a small corpus of materials encoded in a slightly extended version of the PML schema and nothing more sophisticated than basic XSLT (eXtensible Stylesheet Language Transformation) stylesheets are employed to create visualisations and analyses based on the output of these tools. 7

The aim is to demonstrate that, even with these simple techniques, sophisticated and revealing results can readily be obtained from a PML corpus. Extensive synchronic and diachronic studies across substantial datasets are achieveable with these techniques which add a layer of context to the results of standard tools for analysing textual corpora; this additional layer has the potential significantly to enrich these studies, particularly by enabling correlations between the purely linguistic and the contextual to be investigated with relative ease. 8

Parliamentary Metadata Language (PML)

The Parliamentary Metadata Language (PML) schema [LIPARM Project 2012] was originally devised as the key component of the LIPARM project mentioned above. It was conceived as a rigidly structured framework, encoded in XML syntax, for semantically linking core facets of records of legislative proceedings. Seven such facets are incorporated into the schema, most of which are defined very broadly and at a high level of abstraction: they include such concepts as *Unit* (covering everything from political parties, chambers of the legislature, parliamentary constituencies and even gender) and *Proceedings Object* (covering everything that occurs within the proceedings themselves, from procedures such as prayers to speeches and other contributions by individual members). The semantic breadth of most of these facets requires them to be narrowed in use by employing controlled vocabularies of sub-facets — for instance, *Unit*s are refined by these vocabularies into type of unit. A fuller description of the architecture of PML may be found in [Gartner 2013]. 9

PML is designed to act as integrative mechanism for parliamentary proceedings, allowing their textual content to be conjoined with structured metadata: in doing so, it allows diverse content of this kind to be linked into a coherent and interoperable whole. It acts in this way as an "intermediary schema" [Gartner 2011], a mediator between resources that would otherwise be difficult to treat as an integrated resource. The primary mechanism to achieve this is an element available to most facets within a PML document: this `<source>` element, contains an attribute `@sourceRef` which allows a link to an external resource via such mechanisms as URI, Xlink or Xpointer. A speech in a parliamentary debate might, for instance, be linked to its text in a digitised version of *Hansard* as follows: 10

```
<contribution
    type="speech"
    typeURI="http://liparm.ac.uk/contributions/speech"
    contributorID="hh-1980015-person0002">
    <label>Speech</label>
        <sources>
            <source sourceRef="http://hansard.com/commons/1980/jan/15/e-comm-
act-amendment#column_1436" sourceType="URL">
            <label>HC Deb 15 January 1980 vol 976 cc1436-41</label>
            </source>
        </sources>
</contribution>
```

Example 1. Sample PML contribution element (from [Gartner 2013, 29] (truncated))

This example also demonstrates the mechanism by which the broad element `<contribution>` is narrowed semantically to a type of contribution (here a speech) by the use of the `@typeURI` attribute, the use of a `<contributorID>` element to indicate the identity of the member giving the speech and the use of `<label>` elements to provide human-readable handles for each component. 11

As originally conceived in the LIPARM project, PML was intended primarily as a resource-discovery mechanism, allowing the diversely-encoded set of digitized UK parliamentary proceedings to be integrated semantically at a fine level of granularity. For this reason, the PML architecture did not itself incorporate the textual content of the proceedings themselves within its element set: instead, the `@sourceRef` attribute established linkages to this content. This method was employed in the experimental interface to PML data constructed for the LIPARM project [LIPARM Project 2012]: this linked PML files to simple HTML renderings of the original text. 12

The rich set of semantic linkages made possible by the LIPARM architecture should, however, prove amenable to more sophisticated uses than resource discovery alone. It can also be used to support and enrich analyses of the type which are in common usage in the fields of corpus linguistics. In particular, this network of linkages has the potential to put the results of linguistic analyses in context by relating them to the rich set of facets encoded within the PML architecture and, via semantic linkages, to external metadata. The next sections of this article discuss how this can be done using such simple tools as XSLT transformations or XQUERY searches. 13

Creating a PML corpus

To demonstrate the capabilities of PML, a corpus of encoded files was created that covers the period of the last parliament in the UK (2010-2015). This corpus, part of a larger collection covering seventy years of UK parliamentary proceedings, was generated by integrating data from a number of sources: most of the texts themselves were derived from a set of files encoded in the Political Mashups XML schema, which were themselves parsed from the HTML files generated by the Hansard project mentioned above. Other data came from the controlled vocabularies encoded in MADS (Metadata Authority Description Schema) which were produced by the LIPARM project: these included data on Members of Parliament and geographical constituencies. Three consecutive XSLT transformations collected this data from these diverse sources (12,086 files in all) and integrated them into a series of PML files, one covering each day of proceedings. 14

A minor amendment was made to the PML schema to facilitate its use as the basis for an analytical corpus. The `@sourceRef` attribute discussed above was supplemented by the addition of a new element under its parent `<source>` to allow the text of the proceedings to be incorporated directly into the PML file. This `<content>` element is used here to include the text of a member's contribution to a debate: 15

```
<contribution typeURI="http://liparm.ac.uk/id/contributions/speech"
        type="speech"
        contributorID="hh-1980015-person0002"
        regURI="http://uk.proc.d.2010-03-04.2.2.2">
    <label>Kevin Brennan</label>
    <sources>
        <source>
            <content>
                <p>On 2 July 2009 the Government published the consumer White
Paper, which details plans for more effective enforcement against those who
deliberately set out to defraud consumers. I recently announced funding of £4.3
million for the Office of Fair Trading and trading standards to tackle those who
use the internet to con consumers.</p>
            </content>
        </source>
    </sources>
</contribution>
```

Example 2. Sample contribution record with text of contribution included in `<content>` element

Although the `<content>` element in this example contains simple `<p>` (paragraph) elements only, XML data conforming to any namespace would be valid here. 16

The addition of this element allows XML-encoded text of any degree of complexity to be integrated within the heavily structured metadata architecture of PML in a manner analogous to the integrated text and metadata used in many e-text projects. Integrating the text in this way is, of course, not strictly necessary to perform the type of analysis discussed in this article: the linkages established by the `@sourceRef` attribute would allow the disparate corpora of PML and text files readily to be treated as a single entity. The XSLT and XQUERY mechanisms by which the corpora may be interrogated will work just as effectively on such a distributed corpus. 17

Applying sentiment analysis to the PML corpus

To test the feasibility of using PML as an analytical tool, simple sentiment analysis was carried out on a portion of the corpus that covered a single topic, attitudes to immigration. Like all sentiment analysis, this attempted to discern the perspective or attitude of the speaker but was limited to a basic categorization of their expressed views in terms of polarity, whether they were positive, negative or neutral. It did not try to discern more complicated sentiments such as emotional moods or states. 18

This sub-corpus was generated by a simple XQUERY search to extract members' contributions during this period in which the stems `immig*` or `migr*` were present. This sub-corpus was then run through the *OpinionFinder* sentiment analysis software [MPQA Project n.d.] to tag these contributions for positive, negative or neutral sentiments within the same paragraph as the search terms. 19

OpinionFinder is a Java-based application devised by the Universities of Pittsburgh, Cornell and Utah. It can rapidly process multiple files to identify subjective linguistic elements which are then automatically tagged using inline SGML elements [Wilson et al. 2005]. A paragraph element from a PML file tagged in this way might appear in this (slightly simplified) form: 20

```
<p>
    Immigration has
    <MPQAPOL autoclass="positive">enriched</MPQAPOL>
    our culture and
    <MPQAPOL autoclass="positive">enhanced</MPQAPOL>
    our society. Britain can benefit from immigration, but not
    <MPQAPOL autoclass="negative">uncontrolled</MPQAPOL>
    immigration. The levels of net migration seen under the previous
    Government-an annual figure of almost a quarter of a million at
    its peak in 2004-were
    <MPQAPOL autoclass="negative">unprecedented</MPQAPOL>
    in recent times.
</p>
```

Example 3. sample OpinionFinder output (simplified)

For the illustrative purposes of this article, these tags were used to generate a basic measure of the degree of positive or negative sentiment per paragraph by counting the relative numbers of positive, neutral and negative sentiments marked. Clearly more sophisticated techniques are available for analysing output of this type, but these simple measures are sufficient to illustrate the capabilities of PML for contextualising these results. 21

Using the PML architecture to contextualise sentiment analysis data

The primary function of using the network of semantic links embedded within a PML file is to enable data extracted from the textual content encoded within its `<content>` elements, or referenced by the `@sourceRef` attribute, to be contextualised. A number of ways in which this can be done will illustrate something of its potential for enriching the output of linguistic analyses. 22

Interrogating the corpus of PML data can be done by simple XPATH expressions which follow the chain of linkages between data and metadata: although, for the purpose of the tests documented here, these were used on the modified PML files which included the texts of proceedings, they would work equally well using externally held texts referenced by PML's `@sourceRef` attribute. These XPATH expressions may be used with XQUERY queries or within XSLT transformations to produce the same results. For the purpose of this study, XSLT stylesheets were written to generate the results shown although XQUERY can be just used as effectively. 23

One obvious line of enquiry is to examine the sentiments expressed by individual MPs. This is readily done by using the `@contributorID` attribute shown in Example 2 which links a contribution to the person making it. A simple XQUERY or XSLT transformation can be used to extract the contributions made by a member and perform a statistical analysis of the sentiments recorded within them. Here, for instance, the positivity or negativity of comments on immigration by each member for each of the years 2010-2014 are calculated and colour-coded by an XSLT transformation to HTML: 24

	2010	2011	2012	2013	2014
MacShane, Denis	100.00	25.00	33.33	N/A	N/A
Mactaggart, Fiona	54.55	100.00	47.37	37.50	N/A
Main, Anne	N/A	66.67	N/A	100.00	N/A
Mann, John	61.11	44.44	N/A	N/A	N/A
Maude, Francis	N/A	100.00	N/A	N/A	N/A
May, Theresa	70.83	62.86	53.41	47.58	52.08
McCabe, Steven	33.33	100.00	N/A	0.00	N/A

Figure 1. Extracted sentiment data for MPs on topic of immigration 2010-2014

In this chart the colour black indicates the most negative attitudes, reddish brown those which tend to be more neutral and orange the most positive. Where no comments on immigration were made by a member, `N/A` is shown. 25

Although this example shows only a very simple analysis, it does reveal some interesting results. The then Home Secretary, Theresa May, tended initially to express relatively neutral sentiments towards immigration but these grew significantly more negative as the parliamentary term progressed. As Prime Minister she has subsequently committed the country to leave the European single market, the so-called “hard Brexit”, primarily to exert more control on immigration. This simple analysis reveals that her negative views on immigration formed over a number of years. 26

Although this is a relatively broad analysis which conflates all contributions by a given member in a given year, more detailed ones may be carried out at any level of granularity down to individual sessions or parts of sessions; this is because PML incorporates a temporal element, `<calendarObject>`, to which any contribution can be linked. 27

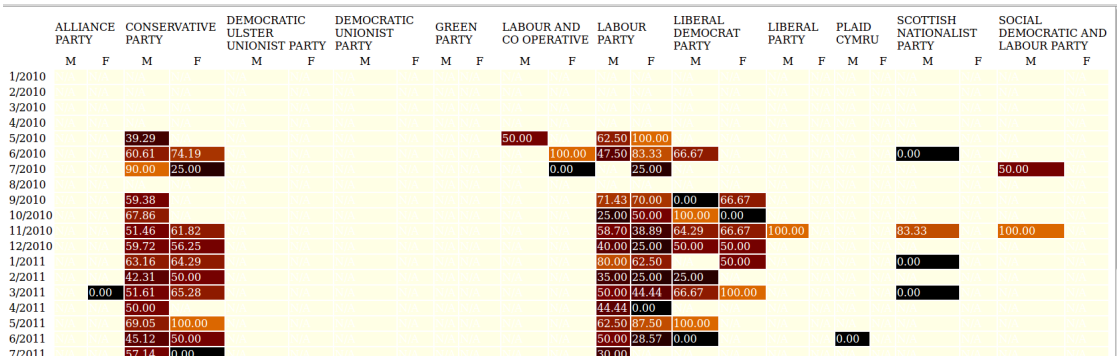
Another level of analysis may be carried out by the same simple statistical technique and colour coding to visualise patterns of sentiment by political party. This time use is made of an attribute, `@categoryID` within PML’s `<person>` element: this includes an attribute which notes a member’s place within a category such as party affiliation. Again a simple HTML table can be generated by an XSLT transformation or an XQUERY query, although here the breakdown of data is by month rather than year: 28

	ALLIANCE PARTY	CONSERVATIVE PARTY	DEMOCRATIC ULSTER UNIONIST PARTY	DEMOCRATIC UNIONIST PARTY	GREEN PARTY	LABOUR AND CO OPERATIVE	LABOUR PARTY	LIBERAL DEMOCRAT PARTY	LIBERAL PARTY	PLAID CYMRU	SCOTTISH NATIONALIST PARTY	SOCIAL DEMOCRATIC AND LABOUR PARTY
1/2010												
2/2010												
3/2010												
4/2010												
5/2010		39.29				50.00	70.00					
6/2010		67.19				100.00	55.77	66.67			0.00	
7/2010		71.43				0.00	25.00					50.00
8/2010												
9/2010		59.38					70.83	50.00				
10/2010		67.86					41.67	50.00				
11/2010		53.98					53.12	64.58	100.00		83.33	100.00
12/2010		59.09					37.50	50.00				
1/2011		63.46					72.22	50.00			0.00	
2/2011		43.33					32.14	25.00				
3/2011	0.00	58.96					47.37	75.00			0.00	
4/2011		50.00					36.36					
5/2011		70.45					79.17	100.00				
6/2011		46.00					36.36	0.00		0.00		

Figure 2. Extracted sentiment data by party (per month)

This analysis, which uses the same colour coding as Figure 1, shows the possibly surprising result that there is little difference between the parties in their (predominantly negative) sentiments expressed towards immigration during the first year of this Parliament. The left-of-centre Labour party appears as negative as the right-wing Conservative party. This may reflect the beginnings of an initiative by the then Labour leader to attempt to make his party appear tougher on immigration, which some at the time felt was an issue that had lost them votes in the General Election of 2010, and which later led to more pronounced comments and policies on this subject [Miliband 2012]. 29

A further level of analysis could extend this breakdown by party to incorporate gender. This is readily done as the same categoryID mentioned above may be used to delineate by gender. A simple modification of the XSLT stylesheet used to generate Figure 2 produces the following breakdown by party and gender, again using the same colour coding; this reveals little variation by this criterion within each party: 30



	ALLIANCE PARTY		CONSERVATIVE PARTY		DEMOCRATIC ULSTER UNIONIST PARTY		DEMOCRATIC UNIONIST PARTY		GREEN PARTY		LABOUR AND CO OPERATIVE		LABOUR PARTY		LIBERAL DEMOCRAT PARTY		LIBERAL PARTY		PLAID CYMRU		SCOTTISH NATIONALIST PARTY		SOCIAL DEMOCRATIC AND LABOUR PARTY	
	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
1/2010																								
2/2010																								
3/2010																								
4/2010																								
5/2010			39.29								50.00		62.50	100.00										
6/2010			60.61	74.19								100.00	47.50	83.33	66.67						0.00			
7/2010			90.00	25.00								0.00		25.00									50.00	
8/2010																								
9/2010			59.38										71.43	70.00	0.00	66.67								
10/2010			67.86										25.00	50.00	100.00	0.00								
11/2010			51.46	61.82									58.70	38.89	64.29	66.67	100.00				83.33		100.00	
12/2010			59.72	56.25									40.00	25.00	50.00	50.00								
1/2011			63.16	64.29									80.00	62.50		50.00					0.00			
2/2011			42.31	50.00									35.00	25.00	25.00									
3/2011		0.00	51.61	65.28									50.00	44.44	66.67	100.00					0.00			
4/2011			50.00										44.44	0.00										
5/2011			69.05	100.00									62.50	87.50	100.00									
6/2011			45.12	50.00									50.00	28.57	0.00				0.00					
7/2011			57.14	0.00									30.00											

Figure 3. Extracted sentiment data by party and gender (per month)

More sophisticated analyses may be achieved by utilising a further feature of PML, its extensive use of URIs to reference almost every component: these may readily be used to interface with externally-held data. One interesting approach to examining patterns of sentiment expressed in the proceedings is to consider their geo-spatial dimension. Each constituency recorded in a PML file, for instance, may be labelled with a URI which in turn can be used to generate geo-data to enable visualisations demonstrating the geographical patterning of the sentiments recorded in parliamentary speeches. 31

One approach to enabling this is to use files encoded in Keyhole Markup Language (KML) [Google Developers 2015] which have already been published to cover UK parliamentary constituencies. These files, produced by MapIt UK [MapIt UK 2015] cover all 650 UK constituencies and include detailed boundary information for each of these. For the purpose of this demonstration, the seventy-two files representing constituencies for Greater London were used. To interface with the PML data, they had to be edited slightly to incorporate the same URIs for constituencies as those found in the PML files. 32

Each KML file is run through an XSLT transformation or XQUERY query which retrieves contributions by the member for each constituency, calculates the relative weights of the positive, negative and neutral values in the sentiments expressed within these, assigns a score from 1 to 100 based on these sentiments and generates a colour code based on this value. This value then replaces the one already present in the KML element that designates the fill colour for the polygon defining the boundaries of each constituency. This process may be used across the whole corpus or be restricted by date, down to a fine level of granularity if required. 33

The resulting set of KML files may be used to generate colour-coded maps using any compliant software. A sample output using Google Earth and covering Greater London in the year 2010 is shown in Figure 4: the legend at the bottom of the image indicates the range on the negative/positive spectrum within which the MP's contributions fell during that year. 34

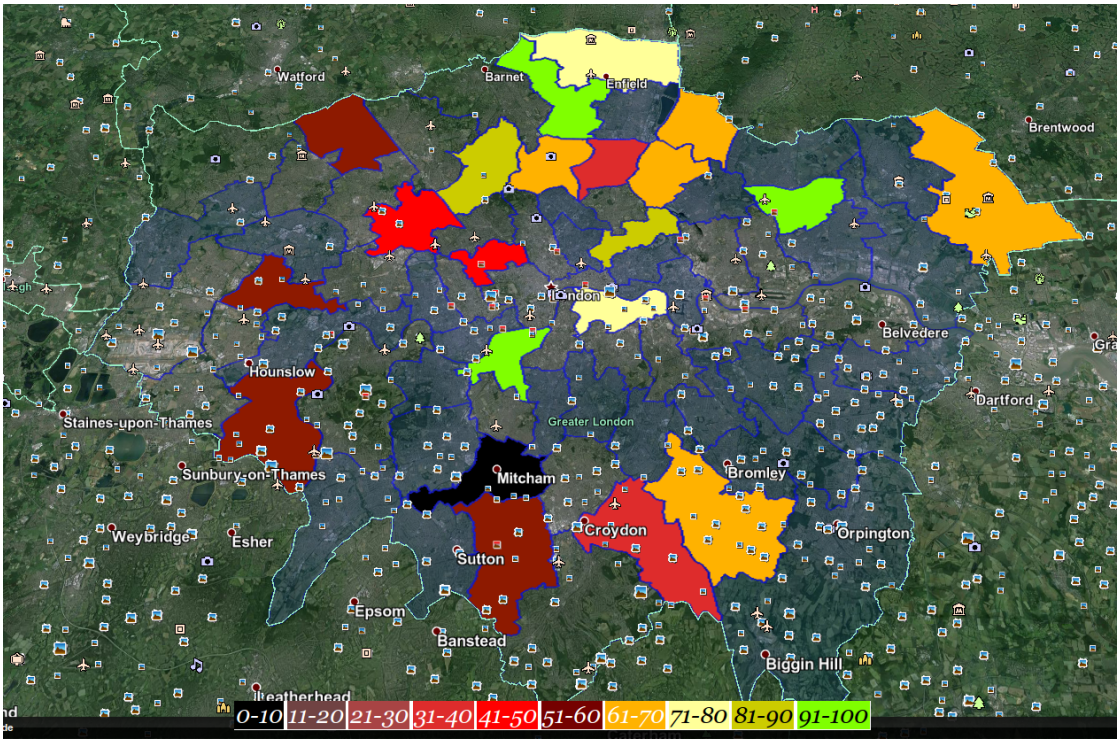


Figure 4. geo-spatial representation of sentiments on immigration expressed by London MPs in 2010

As may be expected in the most diverse and multi-cultural city in the United Kingdom, most sentiments expressed towards immigration by London MPs tend towards the positive, the only exception being those from the member for Mitcham and Morden. Most London's MPs did not comment on immigration during this period and most of those who did come from the outer ring of constituencies, which are generally more affluent and slightly more conservative than those of inner London, but were nevertheless generally positive in their comments on this issue. This is, of course, a visualisation of data from a single year, the year of the General Election: it would be relatively simple to generate animated visualisations representing any changing sentiments within this geographic area over time. 35

Comparison with other approaches to parliamentary metadata

PML's rich set of semantic linkages for parliamentary metadata offers a potentially more powerful base for machine-readable analysis than the other approaches detailed above. At this point, it is useful to compare how its approach differs from these. 36

Compared to the Portuguese metadata system discussed by [Pinto et al. 2005], PML offers much finer granularity and more sophisticated linking. Pinto et al's approach, designed to support an audiovisual delivery system of parliamentary proceedings, only offers granularity down to the page level and a simple XHTML encoding of the textual data [Pinto et al. 2005, 2]. Its mapping of facets of the parliamentary record to DC is limited to a small number (including legislature and session) and it includes none of the contextual information, such as the information on members, gender and geographical data, that PML can support. While functioning very efficiently in its purpose to enable the audiovisual system it supports, it provides very limited facilities for analysis of the type described here. 37

The Akomo Ntoso scheme provides a richer metadata set than the Portuguese for the purpose of marking-up parliamentary documents, as it includes not only debates but also acts, bills, reports and a number of other document types. Its element set is, however, highly document-centred and its potential for encoding wider contextual linkages more limited than PML's. Much of its element set is derived from XHTML and so most of its linking mechanisms use this schema's `@href` attribute to point to externally-held ontologies. Much of the contextualised analytic power of Akomo Ntoso will therefore depend on these external sources and the semantic linkages that they enable: its internal features remain limited for analyses of the type discussed here. 38

The Political Mashup (PM) schema is designed more specifically for analysis but contains limited facilities within its architecture for encoding machine-readable contextual information. The schema is designed specifically for examining the structure of debates as events, and so adapts element sets from the TEI designed for describing theatrical performance: it relies heavily, for instance, on a semantically-imprecise element `<stage-direction>` used to describe most features of the proceedings except for speeches. As is the case for Akomo Ntoso, the schema also uses simple identifiers to externally-held metadata on which it relies for contextualisation: a speech, for instance, may open as follows: 39

```
<speech
    pm:speaker="Andrew Miller"
    pm:party="Lab"
    pm:role="mp"
    pm:party-ref="uk.p.Lab"
    pm:member-ref="uk.m.10435"
    pm:id="uk.proc.d.2011-11-23.1.3.6">
```

Example 4. sample Political Mashups element

None of the referents for these attributes are contained within the PM file: the only metadata components within its architecture are simple Dublin Core elements providing document-level descriptive metadata. Nor does the schema allow the detailed voting analyses provided by PML as it encodes votes as procedural markup, the application of which is inconsistently applied over time. 40

PML appears to offer a better alternative to these other XML approaches for providing the contextual information by which the results of machine-generated linguistic analyses may be enhanced. It offers more powerful and effective mechanisms for contexualising the results of analyses within its tightly-constrained but flexible structures. 41

Other potential uses of PML data

The above discussion is intended to illustrate the potential of using the PML architecture and its semantic linkages in order to contextualise the results of analyses based on the textual content of proceedings. The discussion deliberately concentrates on using simple tools, notably XSLT transformations or XQUERY queries, to extract and visualise the data. 42

Even with these simple tools interesting and revealing results can readily be obtained. There is much scope for further analyses which could exploit the potential of PML to provide new insights into the parliamentary record. Debates currently taking place (at the time of writing) in the UK Parliament on “Brexit”, the UK’s departure from the European Union, for instance, may well yield interesting results when subjected to sentiment-based analyses of this type. Similarly, in the current climate within the US, where much debate appears to take place via Twitter, some potentially valuable research could be undertaken on tracing sentiments within Congressional proceedings and relating them to social media. Robust analyses of this type with a quantitative evidence base could be most valuable and revealing in the current political climate. 43

Some other avenues for research are possible within the PML architecture, although at present these are limited because of the quality of textual corpora available. A potentially powerful area would be to correlate a linguistic analysis of a member's contributions to debates with their voting record. PML allows for the detailed recording of voting patterns using an number of attributes to `<voteEvent>`;, its container element for information on votes or divisions. Using these a complete pattern of an individual's voting record can be derived which could then be incorporated into such an analysis. Unfortunately at present the voting data within most corpora is often a simple, unstructured transcription of the proceedings text itself: to generate an accurate set of data would require manual editing for which sufficient resources 44

have not yet been available. A complete and accurate set of voting data encoded in PML would represent a valuable resource for historians and social scientists.

45 The simple XSLT transformations used to visualise PML data could also be used as the basis of interactive web-based services for presenting parliamentary data for non-specialist users. They could, for instance, to used to provide on-the-fly visualisations across the corpus, potentially enhancing web services which are designed to enhance citizens' access to legislative proceedings (such as the service *They Work For You* in the UK [mySociety 2015]). Projects such as these could be facilitated perhaps by producing modularised XSLT stylesheets for diverse sets of queries or analyses. The sharing of such stylesheets throughout the academic and extra-academic communities could significantly augment the usability of PML-encoded data and make its potential for sophisticated access to the parliamentary record more widely realised.

Conclusions

46 Being able to contextualise the results of fine-grained machine-generated analyses of linguistic data offers the potential to enrich greatly their findings. For the specific area of parliamentary proceedings, the approach offered by Parliamentary Markup Language appears to offer great potential for new avenues of research by allowing contextualisation of this type to be readily achieved. When combined with resources external to the PML file itself, for instance the MADS-encoded controlled vocabularies produced by the LIPARM project, the scope for large-scale studies becomes readily achievable.

47 The range of potential analyses that could be performed with PML is wide-ranging: they could include extensive diachronic historical analyses or on-the-fly visualisations of current proceedings. PML corpora can provide an substantial evidence base for sophisticated machine-readable analyses which extend well beyond the resource discovery functions for which the schema was initially devised. The use of electronic text corpora for parliamentary studies remains relatively nascent, but its potential for seeding future research is clear.

Works Cited

Blaxill 2014 Blaxill, L. 2014. "Digging into Linked Parliamentary Data". Available from: http://dilipad.history.ac.uk/2014/09/ (accessed 6.23.14).

Cervone et al. 2015 Cervone, L. et al. "Akoma Ntoso | Akoma Ntoso Site". Available from: http://www.akomantoso.org/ (2015).

Fuentes Martínez 2005 Fuentes Martínez, M. A. "Metadata at the Library of the National Congress of Chile: a multidisciplinary experience" *DC-2005: Proceedings of the International Conference on Dublin Core and Metadata Applications* (2005).

Gartner 2011 Gartner, R. "Intermediary schemas for complex XML applications: an example from research information management", *Journal of Digital Information*, 12.3 (2011).

Gartner 2012 Gartner, R. "The LIPARM project: a new approach to Parliamentary metadata", *Ariadne*, 70. Available from: http://www.ariadne.ac.uk/issue70/gartner (2012).

Gartner 2013 Gartner, R. "Parliamentary Metadata Language: an XML approach to integrated metadata for legislative proceedings", *Journal of Library Metadata*, 13.1 (2013): 17–35.

Gielissen and Marx 2009 Gielissen, T. and Marx, M. J. "Exemelification Of Parliamentary Debates", *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)* (2009): 19–25.

Google Developers 2015 Google Developers. "Keyhole Markup Language". Available from: https://developers.google.com/kml/ (2015).

LIPARM Project 2012 LIPARM Project. "Controlled Vocabularies | Linking Parliamentary Records through Metadata". Available from: http://www.liparm.ac.uk/?page_id=61 (2012).

MPQA Project n.d. "OpinionFinder | MPQA". Available from: http://mpqa.cs.pitt.edu/opinionfinder/ (2015).

MapIt UK 2015 MapIt UK. "MapIt : map postcodes and geographical points to administrative areas". Available from:

http://mapit.mysociety.org/ (2015).

Miliband 2012 “Ed Miliband’s immigration speech in full”. Available from: http://www.politics.co.uk/comment-analysis/2012/06/22/ed-miliband-s-immigration-speech-in-full (2012).

Piersma et al. 2014 Piersma, H. et al. “War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History”, *Digital Humanities Quarterly*, 8.1 (2015).

Pinto et al. 2005 Pinto, J. S. et al. “Portuguese Parliamentary Records: a Multimedia Digital Library Distributed Architecture, based on Web Services” *International Conference on Next Generation Web Services Practices* (2005).

UK Parliament 2012 “HANSARD 1803–2005”. Available from: http://hansard.millbanksystems.com/ (2012).

Wilson et al. 2005 Wilson, T. et al. “OpinionFinder: A system for subjectivity analysis”, *Proceedings of hlt/emnlp on interactive demonstrations* (2005): 34–35.

mySociety 2015 “TheyWorkForYou: Hansard and Official Reports for the UK Parliament, Scottish Parliament, and Northern Ireland Assembly - done right”. Available from: http://www.theyworkforyou.com/ (2015).