Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account

Naomi Truan, Laurent Romary

Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account

Naomi Truan & Laurent Romary

Version from 18 June 2020

This is a preprint, i.e. a version that precedes formal peer review.

Abstract:

This data paper introduces an integrative and comprehensive method for the linguistic annotation of parliamentary discourse. Initially conceived as a documentation for a specific and rather small-scale research project, the annotation scheme takes into account national specificities and is geared to proposing an annotation scheme that is both highly standardised and adaptable to other research contexts. The paper reads as a specific application of the Text Encoding Initiative (TEI) framework applied to a subset of parliamentary debates. This strategy has two main applications: first, to develop a model for the encoding of parliamentary corpora by providing a systematic way of annotating both elements within the text (e.g. turns, incidents, interruptions) and the metadata associated with it (e.g. variables pertaining to the speaker or the speech event); second, to provide a cross-linguistic empirical basis for further annotation projects.

Keywords: annotation; contrastive linguistics; parliamentary debates; Text Encoding Initiative; open access

1. Introduction: Parliamentary talk as a linguistic object of annotation

Linguistic annotation can be defined as "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 2013: 2). While reflections on linguistic annotation go hand in hand with the development of corpus studies (Ide and Pustejovsky, 2017), we argue that there is still a need for a context-sensitive, fine-grained

annotation of parliamentary corpora, specifically in the context of linguistic research. Since linguistic annotation shapes linguistic research (i.e. allows for specific research questions, but also potentially limits the interpretation), we maintain that it should become an integrative part of linguistic research and, therefore, should be part of the corpus and become available to the research community.

The following data paper aims at tackling this issue by offering methodological reflections on what doing linguistic annotation within the TEI framework means, especially when the annotation serves a small-scale research contrastive project but is geared towards further applications and intends to distribute an open and reusable language resource. In other words, what can be learned from a cross-linguistic linguistic research on parliamentary discourse for other, potentially more comprehensive annotation frameworks for parliamentary discourse? In this paper, we do not propose any analysis or substantive discussion utilising the data. Rather, we intend, through a focus on specialized discourse, to show how and why a reflexion on annotation practices *belongs* to the analysis, and is not only a preliminary step.

On this ground, we present an integrative and comprehensive approach for the linguistic annotation of parliamentary discourse on the basis of "small specialised corpora" (Koester, 2010). We apply the annotation scheme to three electronic corpora based on the stenographic protocols of the British House of Commons, the German Bundestag, and the French Assemblée nationale (Truan 2016a, b, c) available online in Open Access on Ortolang (Open Resources and TOols for LANGuage) since 2016:

- Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]: hdl.handle.net/11403/fr-parl.
- Parliamentary Debates on Europe at the Deutscher Bundestag (1998-2015) [Corpus]: hdl.handle.net/11403/de-parl.
- Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]: hdl.handle.net/11403/uk-parl.

The novelty of the contrastive approach is that it integrates three different parliamentary traditions. In order to ensure not only the interoperability but also the comparison between different parliamentary cultures, we need a common annotation framework flexible enough to accommodate

national specificities, yet standardised enough to be valid for, we expect, any type of parliamentary discourse.

Based on the Text Encoding Initiative (TEI) Guidelines, the annotation framework is both specific to the methodological and technical difficulties encountered while dealing with this particular type of corpora and generalizable to other types of linguistic projects. As we will show, the TEI annotation scheme is indeed the combination of a highly standardised and flexible structure. Accordingly, this paper sets out to propose a model for the encoding of parliamentary corpora by providing a systematic way of annotating both components of the text structure (e.g. turns, incidents, interruptions) and the metadata associated with it (e.g. variables pertaining to the speaker or the speech event). The annotation framework has been conceived as a cross-linguistic empirical basis for further annotation projects. Thus, it can serve as a demonstration of a quite limited, but well-controlled annotation system which allows the researcher to compare (in our case) three rather different sets of data. This annotation system is then recommended for other, similarly focused comparisons. Going further, we believe that the annotation system offers a point of entry into a common methodology that could be used and adapted to scholars working on parliamentary corpus projects more generally.

The following argumentation proceeds in four steps: first, we explain the rationale behind a cross-linguistic encoding of parliamentary debates. Second, we show why the Text Encoding Initiative is a sustainable, reproducible, highly standardised, yet equally flexible annotation framework apt at capturing parliamentary interaction. Third, we describe the annotation scheme at the level of the metadata contained in the TEI header, more specifically the variables associated with each speaker (each Member of Parliament in our case). We also detail the annotation scheme at the level of the text, delineating why the transcription of speech vocabulary should be preferred to a more drama-oriented markup for parliamentary data. A final part is devoted to documenting and archiving the data from an open access perspective.

2. Adopting a contrastive view on the annotation scheme

The corpus annotation and documentation take place in a specific research project focusing on the uses and functions of third-person forms in three communities of practice: the German, French, and British parliaments (Truan forthcoming). While the focus of this research project and the

reasons for the comparison of these three linguacultures will not be discussed in this paper, it appears necessary to sketch out the context in which corpus building has materialised. We first discuss the focus on parliamentary discourse, then move to the contrastive view underlying the project since its inception. We finally set forth the reasons why the Text Encoding Initiative (TEI) is a robust procedure for encoding parliamentary corpora.

2.1. Why parliamentary debates?

Within political discourse, parliamentary debates have recently aroused the interest of linguists (Burkhardt and Pape, 2000; Burkhardt, 2003; Ihalainen, Ilie, and Palonen, 2016), especially because they are particularly insightful corpora for contrastive studies (Bayley, 2004; Ilie, 2010). Yet "in spite of the growing visibility of parliamentary institutions, the scholarly interest for the study of parliamentary discourse has been rather low until recently" (Ilie, 2006: 188). Specifically, linguists have paid attention to personal deixis (Gelabert-Desnoyer, 2008; De Cock and Serrano, 2017), forms of address (Ilie, 2005b), politeness and rudeness (Ilie, 2004, 2005a), and modal verbs (Vuković, 2014). In this context, parliamentary debates increasingly become an object of linguistic annotation (see Fišer and Lenardič, 2018 for an overview of CLARIN parliamentary corpora).

While we do not engage in a debate on whether parliamentary discourse is of intrinsic research value, we believe that parliamentary interaction constitutes a very insightful corpus for linguistic analysis. First, in most Western countries, parliamentary debates are publicly available in several complementary formats: video, audio, text. Hence, the plenary sessions are already transcribed by a team of professional stenographers familiar with parliamentary procedures as well as with the Members of Parliament, thus enabling the researcher to focus on other levels of transcription and annotation.[1] Second, being at the interface between spoken and written data,

[1] The corpus used for the present study relies on the official transcripts of the plenary debates. The differences between stenographic protocols and the parliamentary debates as well as the problem they raise have been extensively described for the three countries under investigation (see Slembrouck, 1992; Mollin, 2007 for the House of Commons; Gardey, 2005 for the Assemblée nationale; Olschewski, 2000 for the German Bundestag). Notwithstanding these valid reservations, official transcripts are "a valuable basis to start from" (Zima, Brône, and Feyaerts, 2010: 140) (also see Cribb and Rochford (2018: 13), who speak of "a robust reporting procedure"). Moreover, video recordings are not a panacea since they are highly dependent on the choices made by the cameraperson. In the case of unauthorised turns, verifications with

parliamentary discourse gives access to a wide range of discourse features (see Vuković, 2012 for a comparison of pre-prepared and spontaneous parliamentary discourse at the House of Commons). As we will show, this feature of parliamentary interaction is crucial. It therefore explains why we adopted a TEI structure based on spoken data rather than drama-oriented data. Finally, parliamentary debates display a wide range of speakers over a large time span, thus inviting for both diachronic and synchronic sociolinguistic case studies in terms of (expressed) gender, status, or political affiliation (see for instance Burnett and Bonami (2019) for the Assemblée nationale).

2.2. Why a new annotated corpus of parliamentary debates?

As sketched above, corpus studies based on parliamentary interaction have become numerous in the last decade. Against this background, what may a new annotated corpus of parliamentary data bring? Why not work with already available parliamentary corpora? While reference corpora such as the Hansard corpus that consists of British Parliament speeches between 1803 and 2005 (1.6 billion words, 7.5 million talks) would offer statistically robust results with corpus-assisted techniques, they also do not give access to the whole co-text[2] because of property rights. Moreover, no equivalent corpus for the German Bundestag and the French Assemblée nationale currently exists.[3]

---

the video recordings are sometimes impossible since the camera focuses on the speaker and very rarely on the co-interlocutors.

[2] We define co-text as the elements surrounding an occurrence, as opposed to the broader context that may be linked to discourse (see Widdowson, 2004: 59).

[3] In the meantime, the GermaParl R data package, a corpus that includes "all plenary protocols that were published by the German Bundestag between February 1996 and December 2016" (Blätte and Blessing, 2018: 810), has been developed (Blaette 2017). Apart from the fact that the period covered by the corpus is by no means comparable to the British House of Commons, it also raises problems in terms of transcription that will be addressed below. Furthermore, as the authors acknowledge, "[a] thematically specialized corpus […] may offer significantly more detailed metadata and annotation" (Blätte and Blessing, 2018: 810). A provisory version of annotated French parliamentary debates has also been created (Diwersy, Frontini, and Luxardo, 2018) after the first release of the corpus in November 2016 (see Section 4 for more detail on the platform that hosts the corpora.)

The variety of sources and formats is a strong argument in favour of a common annotation framework. All the texts have been retrieved from the official websites of the respective parliaments:

- http://hansard.parliament.uk/ for the House of Commons;
- http://pdok.bundestag.de/ for the German Bundestag;
- http://archives.assemblee-nationale.fr/ for the Assemblée nationale.[4]

Both the British House of Commons and the French Assemblée nationale display the parliamentary proceedings in HTML, which allows for a quick, easy, and accurate retrieval of the content. The German corpus, on the other hand, is based on PDF files. PDF files are noticeably less adequate for further encoding and tagging. In this case, the files have sometimes suffered from inadequate word breaks, thus necessitating minor corrections.

We carried out the encoding process into the TEI Guidelines by combining manual and automatic processing workflows, with the idea of keeping both the content and the metadata of the sources. In particular, we used the GROBID software suite[5], which provides a relatively efficient transformation process from PDFs to a decent TEI format, although not fully compliant with the target encoding scheme. Attention was given to unifying the final format across the three languages and parliamentary settings so that the same phenomena and features would be encoded exactly in the same way for each sub-corpus.

2.3. Small monolingual corpora as the basis for a cross-linguistic perspective

The rationale behind the constitution of "small monolingual corpora"[6] (Koester, 2010) is to allow for the interaction between statistical measures and a close-reading analysis sensitive to the socio-political context in which parliamentary interaction takes place. In order to ensure that external variables that may shape parliamentary talk are accordingly assessed, the research project that

[4] All files are from the *Compte rendu intégral* (and not from the *Compte rendu analytique* when it exists).
[5] GROBID is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents with a particular focus on technical and scientific publications. For a detailed description see Romary and Lopez (2015) for a quick overview.
[6] The German corpus is the largest with 417.095 tokens. The British corpus encompasses 188.913 tokens, the French one 137.620.

builds the basis for the annotation scheme focused on a limited range of national debates concerning a major European Council meeting.

Despite their high degree of conventionality, parliamentary debates involve a wide range of different activities (or subgenres) such as ministerial statements, speeches, debates, oral/written questions and Question Time (Ilie, 2006: 191). In order to capture a wide array of speakers and to ensure a thematic continuity, one plenary debate per year held in the British, German, and French national parliaments, respectively, about a major European Council meeting (*ex ante, ex post*, or on the same day) was selected between 1998 and 2015. As Auel and Raunio (2014: 17) stress, "problematic for the comparative analysis is that identifying EU debates is rather difficult in some parliaments". While the Bundestag and the Assemblée nationale list what they consider to be EU debates on their websites for the current and previous legislative periods, the House of Commons does not provide such information on its website. Search engines do not enable further distinction between the EU being only mentioned in a debate on, say, agriculture on a national level, and the EU being the specific topic covered during the plenary session. For these reasons, and given the fact that European affairs are not the focus of this work, but only a common variable to ensure the comparability of the data, it has been assumed that the European Council meetings offer a baseline against which to collect the national plenary debates.

To increase the reliability of the comparison, the genre of parliamentary debate has therefore been considered a constant variable, together with the focus on European Council meetings. The main purpose was to avoid contrastive analyses based on the languages but disregarding the specificities of a particular culture or institution. Following Krzeszowski (1989: 61), we recognise that "[t]ext-bound CS [contrastive studies] are corpus-restricted" since no systematic generalizations outside the original data are made. Bearing in mind that institutional settings are accordingly more stabilised, routinised and conventionalised than everyday interactions, it can be posited that genres function as an intermediary level of representativeness prior to analysis or as a first step towards the comparison of discourse communities that should be the horizon of expectations of a contrastive discourse analysis (see von Münchow, 2010).

While the annotation scheme described in this paper presents typical features of parliamentary interaction, it also represents a first step towards the integration of contrastive perspectives while developing an annotation framework. The advantage of the comparison pertains to its heuristic

value: by reflecting on similarities and differences during the annotation process, we come closer to an architecture valid and applicable to a large variety of linguistic data and metadata.

3. Preventing the built-in obsolescence of the corpus

In this section, we outline the principles guiding the documentation of the corpus and show how the choices we made are intended to serve general purposes going beyond the three countries under investigation. We argue that annotating corpora cross-linguistically calls for a very flexible annotation framework that allows for multiple, expansible, and evolving annotations that may change over the course of time—a principle that is deeply rooted in the Text Encoding Initiative (TEI). We first present the TEI Guidelines and show why they are deemed to be appropriate for parliamentary debates. We then link this general framework to what we call a sustainable corpus.

3.1. The TEI annotation scheme

The Text Encoding Initiative (TEI, see Romary 2009) has become, since its inception in 1987, the reference technical standard for the representation of textual content in the humanities. Based upon the W3C XML recommendation, it covers a wide range of genres and provides users with a vocabulary of nearly 600 XML elements. At the core of the TEI Guidelines resides the face that any TEI-based project should define its own subset (or *customisation*) where the elements which are deemed useful for the representational task at hand are selected, documented and possibly amended.

The TEI annotation is used to store the "detailed information about the speakers or writers" (Koester, 2010: 72). Linked with "the goals of the interactions or texts and the setting in which they were produced as part of the corpus database means that linguistic practices can easily be linked to specific contextual variables" (Koester, 2010: 72). The XML-TEI annotation enables researchers to fruitfully visualise the articulation between text and context, i.e. between the plenary session and the metadata associated with it. Interpretative data is situated within the corpus as dedicated TEI elements, which enables anyone to see it, and correct or extend it if necessary.

Based on this general understanding, the annotation framework has been conceived with this contrastive research question in mind: the subset we have conceived consists in elements that are

deemed equally valid for British, French, and German parliamentary debates. We argue that the cross-linguistic view enables us to take into account national specificities while "emphasiz[ing] what is common to every kind of document", as Burnard (2014) highlights for the TEI. In this sense, and despite the fact that the political context changes over time between France, Germany, and the United Kingdom, the TEI gives access to a common technical, practical, and methodological framework between the three subcorpora and the three languages.

3.2. A sustainable corpus

When designing the TEI-based encoding scheme of our corpus we have been led by the idea that it could be easily taken up by other scholars to carry out various types of research, but also to allow its possible extension (in terms of coverage) or enrichment (e.g. additional annotated features). Although we would avoid the term 'reference corpus', which is more applicable to large scale endeavours to build up a representative sample for a language (see e.g. Kupietz et al., 2010), we strived towards a sustainable corpus that may be combined in time and space with other endeavours to describe language resources in a variety of contexts and for a variety of genres. In this framework, adopting a sampling strategy focused on our research question was not seen as a restriction in the constitution of the corpus. Rather, we saw this strategy as a way to have a better grasp on the parameters for the linguistic analysis and thus encoding.

With this perspective in mind, the use of the TEI Guidelines as a reference background for the encoding scheme was motivated by the lack of consistency across the various corpora of parliamentary debates available online in their native source representations. As reflected in the corpus overview page compiled by the CLARIN infrastructure[7], existing corpora have been mainly designed on the basis of proprietary formats ranging from flat plain-text (Kapočiūtė-Dzikienė, Šarkutė, and Utka, 2017; Clarin:el, 2011) representations to ad-hoc XML vocabularies (Pražák and Šmídl, 2012; Hansen, 2018; Vitali and Zeni, 2007), with even some attempts to define a specific metadata schema for parliamentary debates (Gartner, 2013)—a practice that can be seen as opposite to the underlying assumptions of the TEI community that strives towards standards rather than *ad hoc* solutions. Besides, even for those corpora abiding to the TEI Guidelines, there

[7] https://www.clarin.eu/resource-families/parliamentary-corpora, accessed on 05.03.2019.

are some strong discrepancies in the actual TEI encoding styles: whereas some (Research Group of Computational Linguistics, University of Tartu n.d.) have used a simple paragraph segmentation for the encoding of turns and associated features, others (Blätte and Blessing, 2018) have considered parliamentary debates as a possible instance of drama, with finally a third group of researchers (Pančur, Šorn, and Erjavec, 2018) who based their work upon the Transcription of Speech module of the TEI Guidelines.

The (internal) debate within the TEI community as to which module can optimally deal with parliamentary corpora between 'Drama' and 'Transcription of Speech' relates to a more essential question: how should parliamentary debates be considered as a scholarly source? When designing the annotation scheme we were quickly set on identifying parliamentary debates as the tangible record of an observable interaction rather than a performance that could be derived from a pre-existing script. Indeed, even if MPs may have notes that they read when intervening in a parliamentary debate, "seul le prononcé fait foi", i.e. the transcription only records what has actually been said.

By the same token, even if one could claim—following the theatrical metaphor—that MPs play a role, specifically depending on their relation to the government (majority, opposition) or their specific positioning on certain political issues, we also observe speakers as concrete entities to which we can associate, as we shall see, concrete personal and sociolinguistic markers in the context of a given political speech. Finally, parliamentary debates display a wide range of phenomena pertaining to spoken (multimodal) interactions such as overlaps, interruptions, background noises or applause, which may all be deemed to bear (an interactional, if not political) meaning and thus cannot equate with blocking as indications pertaining to the staging of actors in order to facilitate the performance. Furthermore, MPs often depart from the script. (At the British House of Commons, they are not allowed to read a text out loud.) While the resemblance between parliamentary debates and theatre is attested (Ilie 2003), there is always a room for improvisation, unplanned reactions, interventions, or comments at the parliament. It is true that some of these characteristics may not be transcribed by the official stenographers (see below for a discussion); yet they remain available. This, to our view, pleads in favour of a TEI annotation scheme distinct from drama.

4. Enabling sociolinguistic explorations: The TEI Header

The criteria for documenting the corpus are directly derived from the model sketched out in the first two sections. In this section, we account for two levels of analysis underlying the annotation scheme: first, the TEI header (<teiHeader> element), which stores information related to "the metadata associated with the digital document itself, analogous to the title page of a printed book" (Burnard, 2014), second, the transcriptions of speech within the <text> element itself (for instance, the distribution of turns).

4.1. Political speakers: The TEI element <person>

In this part, we describe the metadata attached to the TEI element <person> corresponding to each speaker. In this corpus, the TEI header contains, among others, the metadata (or variables) associated with the environment of the parliamentary debate (organisation, place, date encoded in <settingDesc>, see figure 3) and with the speakers (name, sex, political party, political affiliation, position encoded in <particDesc>, see figure 1)[8].

An important decision was to encode speakers' related information in the header of each document and to associate such descriptions with a group of features relevant for the linguistic analysis of parliamentary discourse. In compliance with the TEI Guidelines, and more specifically its *Language Corpora* module, such information is situated in the profile description section (<profileDesc>) of the TEI header within the element (<particDesc>) dedicated to the cataloging of participants in a spoken discourse. Our choice was essentially motivated by the need to find an adequate compromise between two possible strategies:

i. on the one hand, localising speaker-related information at the utterance level, with the risk of lacking genericity, introducing redundancy and above all introducing contradictory information throughout the document, when annotation is not carried out consistently;

ii. on the other hand, grouping all speakers' related information within a global prosopographic document (i.e. an independent digital thesaurus of persons) where each MP would have been identified once and for all, thus preventing a finer grained analysis

[8] For a comprehensive mapping of all the TEI tags used in this work and how they have been applied to parliamentary debates specifically, see [link removed because points to one of the authors].

accounting for the variation of, for instance, political role over time and across parliamentary debates[9].

As a consequence, our documentation strategy has been determined by our ground decision within our corpus to fragment parliamentary debates into document units corresponding to plenary sessions, with the additional advantage of optimising the maintenance of the corresponding information within our corpus at large (e.g. allowing other researchers to easily complement the corpus with additional sessions, as independent TEI documents), as well as facilitating cross-session analysis. Hence, each XML-TEI document corresponds to one plenary debate as a communicative unit, i.e. a given spatiotemporal unit bound to a specific situation in which a group of given participants discusses a given topic (Kerbrat-Orecchioni, 1990: 216), thus making the text the proper linguistic object under investigation[10].

We have chosen to identify the speakers in each debate in the corresponding header and *not* in each utterance (or prior to each utterance) for three main reasons:

i. it allows for a better readability of the TEI document at first glance since the metadata associated with each speaker is not mixed—and thus potentially hard to retrieve—all together in the text (see the ode to simplicity in the next section);

ii. it ensures the consistency of the parameters applied to each speaker since the list of the speakers attending a specific plenary debate is given at the beginning;

iii. it permits to develop and extend the metadata associated with each speaker if necessary by changing the TEI header only once, and not every time a speaker produces a new turn.

In this context, the documentation of speakers in the header plays a double role for the management of our transcription document:

i. first, it ensures a unique identification of the speakers[11] across their various interventions within a plenary debate;

[9] There is a trade-off here as to how much speaker-related information should be localized with the parliamentary debate as opposed to be grouped in a prosopographic document. We expect our encoding to reflect the need to make each plenary debate an autonomous object not requiring a constant back and forth access to an external authority document.

[10] But note that one session can last more than one day, i.e. can be split.

[11] Speakers are mostly MPs, but not only. For instance, in France, members of the government invited to make a speech are not members of the parliament.

ii. second, it provides various descriptive features which are both stable for the corresponding debate and relevant for the purpose of the study of parliamentary discourse at large.

```
<teiHeader>
  <profileDesc>
    ...
    <particDesc>
      <listPerson type="parliamentarians">
        ...
        <person xml:id="ROBERTSON-ANGUS">
          <persName>Angus Robertson</persName>
          <sex>male</sex>
          <occupation>MP</occupation>
                <affiliation>Scottish National Party</affiliation>
          <trait type="party">
            <desc>Independent</desc>
          </trait>
          <floruit>opposition</floruit>
          <nationality>UK</nationality>
          <residence>Moray</residence>
        </person>
        ...
      </listPerson>
    </particDesc>
  </profileDesc>
</teiHeader>
```

Figure 1: Example of a speaker's description entry in the TEI header of a session document

Usually, the <id> of a speaker corresponds to the last name. In case of speakers sharing their last name with another speaker of the corpus, as is the case here, the first name has been added. Another option could have been to add the date of birth for each speaker.

Crucially, providing the speaker's description at the (local) level of each parliamentary debate or TEI document does not prevent from setting up an external, more comprehensive prosopographic document where all biographic indications (and somehow independent from specific political contexts) may be maintained. Referring from the corpus documentation to such a prosopographic document by means of the @corresp attribute on the <person> element is technically simple.

The first group of features attached to the description of an MP within a plenary debate corresponds to stable—or bearing very rare variation—characteristics pertaining to the identification of the speakers according to long-term properties such as name (<persName>), sex (<sex>[12]) and nationality (<nationality>). The second group of features is more specific to each plenary debate and corresponds to the characteristics borne by the speakers from a political point of view: these are their political affiliation (<affiliation>), their relation to current government (<floruit>, with values *majority* and *opposition*) and the electoral circumscription where they have been elected (<residence>).

This approach allowed us to look into the corpus through variables that have not, as far as we know, been consistently integrated into the corpus-based and corpus-driven analysis of parliamentary discourse so far. We could gain insights of the relationship between opposition and majority in terms of person reference that otherwise would have remained hidden. For instance, referring to *certains* ('some'), for a member of the UMP (Conservatives in France), is likely to denote the Communists at the Assemblée nationale. Building categories of discourse participants is closely intertwined with the speaker's construal of who is included and who is excluded. Such a finding could only be attained through the exploration of the correlation between linguistic forms and TEI variables manually encoded.

[12] Although the TEI documentation reports on a @value attribute to normalise the corresponding content of the <sex> element, it does not provide a real standardised set of values as reference (https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sex.html, accessed on 29.04.2019). We thus discarded this attribute in our encoding.

As we will show, the annotation framework was geared towards the coding of external variables (or metadata) which had only rarely been taken into account until the first release of the corpus in November 2016, such as the variable majority/opposition or grouping parliamentary groups together such as PDS/Die Linke coded as "Far Left" (see the use of <trait> in figure 1) (for some observations on the variable majority/opposition in a Norwegian corpus, see Lapponi and Søyland, 2016; Lapponi et al., 2018).

Although we have not encountered this situation in our corpus, it should be signalled that even in the last group of features, a change within a given debate can happen, when for instance an MP changes sides. Such a scenario is attested by the creation of The Independent Group (TIG) in February 2019[13]. In such cases, the flexibility of the TEI toolkit would allow for a meaningful representation, notably through the use of temporal attributes as exemplified in figure 2.

```
<person xml:id="SOUBRY">
  <persName>Anna Mary Soubry</persName>
  ...
  <affiliation notAfter="2019-02-20">Tories</affiliation>
  <affiliation notBefore="2019-02-20">The Independent Group</affiliation>
  …
</person>
```

Figure 2: Exemplifying a change in political party within a plenary debate

4.2. The speech event: The TEI element <settingDesc>

As shown previously, each parliamentary debate constitutes a specific speech event taking place at one time and one place. The speech event constitutes a macro frame in which speakers, who alternatively become hearers as well, produce several turns. The contextual description of the speech event must thus contain the basic features that enable a user of the corpus to situate each

[13] https://news.sky.com/story/live-speculation-more-mps-will-quit-to-join-independent-group-11642586, accessed on 16.03.2019.

utterance within a precise geo-temporal environment, but also to understand the broader political context.

The TEI Guidelines provide a suitable construct to do so within the TEI header by means of the <setting> element within <settingDesc> element, whose usage we have adapted to match our purposes. As illustrated in the example below, we have described the following features attached to a parliamentary debate:

i. the name of the organisation (<orgName>[14]) where the debate is taking place, namely the corresponding national parliament (for this corpus: House of Commons, Deutscher Bundestag and Assemblée Nationale);

ii. the actual date of the debate (<date type="parliamentaryDebateDate">) both as recorded in the original transcript and normalized according to the ISO standard 8601 (yyyy-mm-dd);

iii. the name of the head of government in place (<persName>[15]), so that the debate can easily be put in relation with a wider political context. We adopted a complementary numbering marker (e.g. Blair I) to signal successive government with the same leader;

iv. the actual legislative session (<name>[16]) within which the debate is taking place.

In addition to these generic political parameters, we added two specific descriptors[17] intended to provide information about the European debate *per se*, namely a description of the main topic(s) of the European Council meeting about which the national parliament is debating, together with the place where the European Council meeting took place. We used the <activity> (for EC meeting topics) and <locale> (for the meeting place) elements to this purpose, with the understanding that these could probably be the least consensual choice if we were to carry out a

[14] It may have been more adequate to actually indicate the organisation (<org type="parliament">) rather than just the name, but the corresponding element is currently not allowed in <setting>.
[15] A further development of general guidelines for such encodings should lead to agree on a @type for such elements (e.g. type="governmentHead").
[16] Here also we could think of type the element within a larger standardisation context, and possibly link this with a reference document of legislature events.
[17] In keeping with our general encoding strategy, we reused existing elements from the TEI Guidelines, while slightly adapting their semantic as TEI components.

wider dialogue with the scientific community on the standardisation process and the encoding of parliamentary debates.

```
<settingDesc>
  <setting>
    <orgName type="parliament">Assemblée Nationale</orgName>
     <date type="parliamentarySessionDate" when="2008-12-10">10 December 2008</date>
     <activity>Treaty of Lisbon, General questions</activity>
     <locale>Brussels</locale>
     <persName>Sarkozy</persName>
     <name>XIIIe législature</name>
  </setting>
</settingDesc>
```

Figure 3: Example of a session's description entry in the TEI header of a session document

5. Encoding the content: An ode to simplicity

5.1. The representation of spoken political discourse: The turn level

Utterances/Turns

The <u> element (with gloss *utterance*) in the TEI Guidelines potentially covers any kind of linguistic segmentation in a transcription of a spoken sequence as long as this segment may be attributed to a single speaker. For the purpose of encoding parliamentary debates, we decided to adopt a terser interpretation of this element and considered that it should represent a *turn* in the standard linguistic acceptance. Turns are a superficial unit pertaining "to the surface structure of conversation" (Kerbrat-Orecchioni, 2004: 8) since they solely indicate a change of speaker. The reason behind is essentially to account for the essentially monological nature of parliamentary interaction so that a specific speaker's intervention can be easily identified and distinguished from the preceding and following turns of other MPs.

```
<u who="#ROBERTSON-ANGUS"> On the question of European enlargement and immigration […] </u>
```

Interruptions

In Diwersy, Frontini, and Luxardo (2018), the authors observe that the descriptor "speech type (debate, interruption, vote explanation, etc.)", which is not given in our corpus annotation, proves to be "particularly important when it comes to differentiate effects of register variation ranging from highly formulaic to less formal speech (as in the case of e.g. interruptions)". The main reason for not annotating this level of analysis is, once again, to be found in the contrastive perspective we adopt. Whereas interruptions are thoroughly transcribed in the official recordings of the Bundestag and the Assemblée nationale, enabling new research questions on the special kind of dialogue emerging during these interactions, unexpected or unauthorised turns at the British parliament are only indicated as 'interruption' with no further information provided on the nature, source or content of the disruption, as in (1):

(1) Mr. David Cameron (Tories) [majority]: There is a case for saying that the institutions that Europe put in place after the second world war and I would include NATO as well as the European Union have played a role in making sure that we settle our problems around conference tables rather than on the fields of Flanders. To that extent, yes, I think that it is right. ***Interruption*** Someone says, "Why not go?". (UK 2012.10.22)

Although the co-text sometimes gives insights on what kind of 'interruption' was at stake (and although the video recordings are available online), it is clear that transcription practices (to name only one factor) have a considerable impact on a contrastive research overall. For statistical purposes, it appeared more suitable to encode changes of speakers without discriminating between unauthorised and authorised interventions, which enables us to retrieve automatically all the utterances of a given speaker.

5.2. Segments and quotes: The intra-turn level

Finally, we had to resort to the very generic <note> element to mark up additional commentaries present in the transcripts of the debates and usually added by the parliamentary clerks:

<note>Official Report, 15 January 2014, Vol. 573, c. 11MC.</note>

For the purpose of our corpus, we have not fully used the richness of the Transcriptions of Speech module of the TEI Guidelines, as described in Schmidt (2011). This is both due to the specific

scope of the linguistic study that we were pursuing and the actual informational simplicity of the available sources. Still, the choice we made of using this module offers the possibility of a variety of potential enrichments, either by ourselves, or indeed by anyone who would want to further complement the corpus. Among the important enrichment that could be carried out, we can mention linking the transcriptions with the actual source in audio or video format. The possibility to align with precision, but means of a timeline, the various turns, sub-segments or any kind of incident, offers the potential to have a better insight in the nature of the interactions carried out in parliamentary contexts, from a prosodic or gestural point of view for instance.

6. Documenting and archiving the data

As already alluded to, the corpus has been designed with the idea that it could be easily reused and complemented by others. It thus appeared coherent to adopt a completely open distribution setting for it, by releasing it on the Ortolang platform. Ortolang combines several important technical features:

i. specialisation on linguistic data with the possibility to attach several linguistic descriptors (language, genre, source type etc.) to the corpus itself;

ii. provision of unique identifiers to the resources;

iii. long-term archiving for all uploaded resources;

iv. version management, which allows to publish corrections and improvements to the corpus while keeping the same underlying digital identity;

v. precise identification of the various contributors to a resource;

vi. linking of resources with open licences—in our case a Creative Commons CC-BY licence requiring proper attribution to the authors (CC-BY);

vii. finally, the possibility to add an XSLT stylesheet to the corpus to account for a default search and presentation environment (in HTML).

Beyond the technical setting, we conclude with dissemination issues that, to our view, are an essential part of the annotation project. First, we considered that beyond seeing the corpus as reusable (linguistic) content, presenting the annotation framework as an ongoing process could

also play a role as a methodological point of comparison for other comparable endeavours. As a consequence, we decided to distribute all the source documents rather than limiting access through, e.g., a query interface, as is the case for the EuroParl corpus for instance. Second, although there are often fears of being plundered when data is disseminated at too early a stage in a research process, the author who compiled the corpus as part of her dissertation project took the decision to have the data online even before the actual doctoral publication would be available.[18]

The three corpora are available online at the following addresses:

- hdl.handle.net/11403/fr-parl for the French corpus (Truan 2016a);
- hdl.handle.net/11403/de-parl for the German corpus (Truan 2016b);
- hdl.handle.net/11403/uk-parl for the British corpus (Truan 2016c).

These links are dynamic persistent identifiers that always reference the latest published version of the subcorpora; thus no specific date of access of the given sources is provided. The online access of the corpus (or the three sub-corpora) has been released in November 2016.

7. Conclusion

This data paper suggested an integrative and comprehensive approach for the linguistic annotation of parliamentary discourse that takes into account national specificities and is specifically geared to proposing an annotation scheme that is both highly standardized and adaptable. The method is based on the Text Encoding Initiative (TEI) framework. We argued that the linguistic features of parliamentary interaction call for an annotation scheme distinct from the ways theatrical plays have been accounted for within the TEI community. We also pleaded for a cross-linguistic annotation framework easily reproducible. Specifically, we have shown that the metadata information such as 'political affiliation' or the opposition between majority and opposition are crucially needed in order to allow for the comparison between several parliamentary systems.

We understand this paper as a first step towards the annotation of parliamentary corpora on a larger scale. We recognize that the small size of the corpora (from approximately 137.000 tokens

[18] One of the points of contention could be that the dissemination through Ortolang is not a fully open source project such as GitHub. We see GitHub, which is a private platform and thus does not fulfil all our criteria of a sustainable environment, as a possible front end for the further development of such a corpus as ours, while keeping an environment such as Ortolang as the final end publication setting.

for the French corpus to 417.000 for the German corpus) allowed for such a fine-grained annotation that may be more difficult to implement more massively. Accordingly, the application of these guidelines to a bigger corpus needs to be systematized. On the other hand, it would also be possible to further complement the detailed annotation scheme, for instance by providing timestamps and the hyperlinks to the videos, as suggested by (Cribb and Rochford, 2018: 13), in particular "so that a user at a particular point in the report can link through to the audio recording effortlessly and accurately". A narrower linkage between the videos and the transcripts could also lead to an insightful annotation in terms of kinesics—a dimension which, arguably, would adequately complete a close-reading discourse-analytic endeavour.

These further extensions and exploitations of the annotated corpora are at the core of our understanding of annotation as a process rather than a finish product (also see Bucholtz, 2000) for a similar argument in terms of "the politics of transcription"). Making decisions explicit, transparent, and replicable are primary prerequisites for doing science in the digital age. We hope to have shown that the annotation scheme developed in this project in only a first step.

References

Bayley, P., ed. (2004). *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam/Philadelphia: John Benjamins.

Blaette, A. (2017). *GermaParl. Corpus of Plenary Protocols of the German Bundestag*. R Data Package (v1.0.4). http://polmine.sowi.uni-due.de/packages/src/contrib/GermaParl_1.0.4.tar.gz

Blätte, A., and Blessing, A. (2018). "The GermaParl Corpus of Parliamentary Protocols." In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, 810–16. Miyazaki. http://www.lrec-conf.org/proceedings/lrec2018/pdf/1024.pdf

Bucholtz, M. (2000). "The Politics of Transcription." *Journal of Pragmatics* 32 (10): 1439–65. https://doi.org/10.1016/S0378-2166(99)00094-6

Burkhardt, A. (2003). *Das Parlament und seine Sprache. Studien zu Theorie und Geschichte parlamentarischer Kommunikation*. Tübingen: Max Niemeyer.

Burkhardt, A, and Pape, K. eds. (2000). *Sprache des deutschen Parlamentarismus. Studien zu*

150 Jahren parlamentarischer Kommunikation. Wiesbaden: Springer.

Burnard, L. (2014). *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Encyclopédie Numérique. Marseille: OpenEdition Press. http://books.openedition.org/oep/426

Burnett, H., and Bonami, O. (2019). "Linguistic Prescription, Ideological Structure, and the Actuation of Linguistic Changes: Grammatical Gender in French Parliamentary Debates." *Language in Society*, no. 48: 65–93. https://doi.org/10.1017/S0047404518001161

Clarin:el (2011). "Tour de CLARIN: Clarin:El Presents the Hellenic Parliament Sittings and Hellenic Parliamentary Corpus H-ParCo | CLARIN ERIC." 2015 2011. https://www.clarin.eu/blog/tour-de-clarin-clarinel-presents-hellenic-parliament-sittings-and-hellenic-parliamentary-corpus

Cribb, V.M. and Rochford, S. (2018). "The Transcription and Representation of Spoken Political Discourse in the UK House of Commons." *International Journal of English Linguistics* 8 (2): 1. https://doi.org/10.5539/ijel.v8n2p1

De Cock, B., and Nogué Serrano, N. (2017). "The Pragmatics of Person Reference: A Comparative Study of Catalan and Spanish Parliamentary Discourse." *Languages in Contrast*.

Diwersy, S., Frontini, F. and Luxardo, G. (2018). "The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse." In *Proceedings of the ParlaCLARIN@ LREC2018 Workshop*, 6. Miyazaki.

Fišer, D., and Lenardič, J. (2018). "CLARIN Corpora for Parliamentary Discourse Research." In Miyazaki. http://lrec-conf.org/workshops/lrec2018/W2/pdf/14_W2.pdf

Gardey, D. (2005). "Turning Public Discourse into an Authentic Artefact: Shorthand Transcription in the French National Assembly." In *Making Things Public. Atmospheres of Democracy*, edited by Bruno Latour and Peter Weibel, 836–43. Cambridge, Massachuetts: MIT Press. https://archive-ouverte.unige.ch/unige:76415

Gartner, R. (2013). "Parliamentary Metadata Language: An XML Approach to Integrated Metadata for Legislative Proceedings." *Journal of Library Metadata* 13 (1): 17–35.

https://doi.org/10.1080/19386389.2013.778728

Gelabert-Desnoyer, J. (2008). "Not so Impersonal: Intentionality in the Use of Pronoun *Uno* in Contemporary Spanish Political Discourse." *Pragmatics* 18 (3): 407–24.

Hansen, D. (2018). "The Danish Parliament Corpus 2009 - 2017, v1." https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/8

Ide, N. and Pustejovsky, J. eds. (2017). *Handbook of Linguistic Annotation*. Dordrecht: Springer.

Ihalainen, P., Ilie, C. and Palonen, K. eds. (2016). *Parliament and Parliamentarism. A Comparative History of a European Concept*. New York/Oxford: Berghahn.

Ilie, C. (2003). "Histrionic and Agonistic Features of Parliamentary Discourse." *Studies in Communication Sciences* 3 (1): 25–53.

Ilie, C. (2004). "Insulting as (Un)Parliamentary Practice in the British and Swedish Parliaments. A Rhetorical Approach." In *Cross-Cultural Perspectives on Parliamentary Discourse*, edited by Paul Bayley, 45–86. Discourse Approaches to Politics, Society and Culture 10. Amsterdam/Philadelphia: Johns Benjamins.

Ilie, C. (2005a). "Interruption Patterns in British Parliamentary Debates and Drama Dialogue." In *Selected Papers from the 9th IADA Conference, Salzburg 2003*, edited by Anne Betten and Monika Dannerer, 418–30. Tübingen: Max Niemeyer.

Ilie, C. (2005b) "Politeness in Sweden: Parliamentary Forms of Address." In *Politeness in Europe*, edited by Leo Hickey and Miranda Stewart, 174–88. Clevedon, England: Multilingual Matters.

Ilie, C. (2006). "Parliamentary Discourses." In *Encyclopedia of Language & Linguistics*, edited by Keith Brown, 2nd ed., 188–97. Oxford: Elsevier.

Ilie, C., ed., (2010) *European Parliaments under Scrutiny. Discourse Strategies and Interaction Practices*. Amsterdam/Philadelphia: John Benjamins.

Kapočiūtė-Dzikienė, J., Šarkutė L., and Utka, A. (2017) "Lithuanian Parliament Corpus for Authorship Attribution." *http://dangus.vdu.lt/~jkd/eng/*. https://clarin.vdu.lt/xmlui/handle/20.500.11821/17

Kerbrat-Orecchioni, C. (1990). *Les interactions verbales. Tome 1*. Vol. 1. Paris: Armand Colin.

Kerbrat-Orecchioni, C. 2004). “Introducing Polylogue.” *Journal of Pragmatics* 36: 1–24. https://doi.org/10.1016/S0378-2166(03)00034-1

Koester, A. (2010). “Building Small Specialised Corpora.” In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O’Keeffe and Michael McCarthy, 66–79. London/New York: Routledge.

Krzeszowski, T. (1989). “Towards a Typology of Contrastive Studies.” In *Contrastive Pragmatics*, edited by Wieslaw Oleksy, 55–72. Pragmatics & Beyond New Series. Amsterdam/Philadelphia: John Benjamins.

Kupietz, M., Belica, C., Keibel, H. and Witt, A. (2010). “The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research.” In *LREC*.

Lapponi, E. and Søyland, M. (2016). “Talk of Norway.” *http://www.mn.uio.no/ifi/english/research/projects/ton/*. https://repo.clarino.uib.no/xmlui/handle/11509/123

Lapponi, E., Søyland, M. Velldal, E. and Oepen, S. (2018). “The Talk of Norway: A Richly Annotated Corpus of the Norwegian Parliament, 1998–2016.” *Language Resources and Evaluation* 52 (3): 873–93. https://doi.org/10.1007/s10579-018-9411-5

Leech, G. (2013). “Introducing Corpus Annotation.” In *Corpus Annotation*, edited by Roger Garside, Geoffrey N. Leech, and Tony McEnery, 1–18. London/New York: Routledge.

Mollin, S. (2007). “The Hansard Hazard: Gauging the Accuracy of British Parliamentary Transcripts.” *Corpora* 2 (2): 187–210. https://doi.org/10.3366/cor.2007.2.2.187

Münchow, P. von (2010). “Langue, discours, culture : quelle articulation ? (1ère partie).” *Signes, discours et sociétés*, Visions du monde et spécificité des discours, , no. 4. http://www.revue-signes.info/document.php?id=1439

Olschewski, A. (2000). “Die Verschriftung von Parlamentsdebatten durch die stenographischen Dienste in Geschichte und Gegenwart.” In *Sprache des deutschen Parlamentarismus. Studien zu 150 Jahren parlamentarischer Kommunikation*, edited by Armin Burkhardt and Kornelia Pape, 336–53. Wiesbaden: Springer.

Pančur, A. Šorn, M. and Erjavec, T. (2018). “SlovParl 2.0: The Collection of Slovene

Parliamentary Debates from the Period of Secession."

Pražák, A. and Šmídl, L. (2012). "Czech Parliament Meetings." https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0005-CF9C-4

Research Group of Computational Linguistics, University of Tartu. n.d. "Reference Corpus of Estonian: Transcripts of Riigikogu (Estonian Parliament)." Accessed March 4, 2019. https://www.cl.ut.ee/korpused/segakorpus/riigikogu/

Romary, L. (2009). Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*, Mentis Verlag. https://hal.archives-ouvertes.fr/hal-00348372

Romary, L. and Lopez, P. (2015). "GROBID - Information Extraction from Scientific Publications." *ERCIM News* 100 (January). https://hal.inria.fr/hal-01673305/document

Schmidt, T. (2011). "A TEI-Based Approach to Standardising Spoken Language Transcription." *Journal of the Text Encoding Initiative*, no. 1 (June). http://jtei.revues.org/142

Slembrouck, S. (1992). "The Parliamentary Hansard 'Verbatim' Report. The Written Construction of Spoken Discourse." *Language and Literature* 1 (2): 101–19. https://doi.org/10.1177/096394709200100202

Truan, N. (2016a). Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). hdl.handle.net/11403/fr-parl.

Truan, N. (2016b). Parliamentary Debates on Europe at the Deutscher Bundestag (1998-2015) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). hdl.handle.net/11403/de-parl.

Truan, N. (2016c). Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). hdl.handle.net/11403/uk-parl.

Truan, N. (forthcoming). *The Politics of Person Reference. Third-person forms in English, German, and French* (Pragmatics & Beyond New Series). Amsterdam/Philadelphia: John Benjamins. 10.1075/pbns.320.

Vitali, F., and Zeni, F. (2007). "Towards a Country-Independent Data Format: The Akoma Ntoso

Experience." In *Proceedings of the V Legislative XML Workshop*, 67–86.

Vuković, M. (2012). "Positioning in Pre-Prepared and Spontaneous Parliamentary Discourse: Choice of Person in the Parliament of Montenegro." *Discourse & Society* 23 (2): 184–202. https://doi.org/10.1177/0957926511431507

Vuković, M. (2014) "Strong Epistemic Modality in Parliamentary Discourse." *Open Linguistics* 1 (1). https://www.degruyter.com/view/j/opli.2014.1.issue-1/opli-2014-0003/opli-2014-0003.xml

Widdowson, H. G. 2004. *Text, Context, Pretext. Critical Issues in Discourse Analysis*. Language in Society 35. Malden, MA: Blackwell.

Zima, E., Brône, G., and Feyaerts, K. (2010). "Patterns of Interaction in Austrian Parliamentary Debates. On the Pragmasemantics of Unauthorized Interruptive Comments." In *European Parliaments under Scrutiny. Discourse Strategies and Interaction Practices*, edited by Cornelia Ilie, 135–64. Amsterdam/Philadelphia: John Benjamins.