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

Naomi Truan, Laurent Romary

Journal of the Text Encoding Initiative

Issue 14 | 2021
Selected Papers from the 2019 TEI Conference

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

Naomi Truan and Laurent Romary

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/jtei/4164
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Naomi Truan and Laurent Romary, "Building, Encoding, and Annotating a Corpus of Parliamentary Debates in TEI XML: A Cross-Linguistic Account", *Journal of the Text Encoding Initiative* [Online], Issue 14 | 2021, Online since 22 June 2022, connection on 22 June 2022. URL: http://journals.openedition.org/jtei/4164

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

Naomi Truan and Laurent Romary

ABSTRACT

This paper introduces an integrative and comprehensive method for the linguistic annotation of parliamentary discourse. Initially conceived as documentation for a specific and small-scale research project, the annotation scheme takes into account national specificities and is geared to proposing an annotation scheme that is both highly standardized and adaptable to other research contexts. In this paper we present a specific application of the Text Encoding Initiative (TEI) framework applied to a subset of official transcripts of plenary proceedings in three parliamentary cultures. The TEI annotation scheme proposed here has two main applications: first, it serves as a basis for encoding parliamentary corpora by providing a systematic way of annotating both

elements within the text (e.g., turns, incidents, and interruptions) and the metadata associated with it (e.g., variables pertaining to the speaker or the speech event); second, it provides a cross-linguistic empirical basis for further annotation projects.

INDEX

Keywords: annotation, contrastive linguistics, parliamentary debates

EDITOR'S NOTES

Although this article was submitted to our rolling issue, it was added *ex post facto* with the authors' consent to Selected Papers from the 2019 TEI Conference given its thematic similarities to some of the other articles in this issue.

ACKNOWLEDGEMENTS

We would like to thank Erzsébet Tóth-Czifra, Tomaž Erjavec, and Andrej Pančur for valuable comments on earlier versions of this paper.

1. Introduction: Parliamentary Talk as a Linguistic Object of Annotation

1 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). As 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 the issues and decisions pertaining to linguistic annotation are an integral part of linguistic research, and, therefore, should be part of the annotated corpus once released and should become available to the research community together with the data.

2 This 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 contrastive research project but is geared toward further applications and intends to distribute an open and reusable language resource. In other words, what can be learned from cross-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 of the data. Rather, we intend, through a focus on specialized discourse, to show how and why a reflection on annotation practices *belongs* to the analysis, and is not only a preliminary step.

3 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, 2016b, 2016c). The novelty of this 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 standardized enough to be valid for, we expect, any type of parliamentary discourse. Based on the Text Encoding Initiative (TEI) Guidelines (TEI Consortium 2022), the annotation framework combines a high degree of standardization with a flexible structure: it is both specific to the methodological and technical difficulties encountered while dealing with these particular types of corpora and generalizable to other types of linguistic projects that may aim at extending or refining the annotation scheme presented here.

4 The following discussion proceeds in four steps: first, we explain the rationale behind a cross-linguistic encoding of parliamentary debates. Second, we show why the TEI is a sustainable, reproducible, highly standardized, yet equally flexible annotation framework suited to 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. Finally, in order to invite further applications of the TEI framework, we discuss documenting and archiving the data from an open-access perspective.

2. Adopting a Contrastive View on the Annotation Scheme

5 The corpus annotation and documentation takes 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 2018, Truan 2021). The dissertation project and, subsequently, the manual annotation of the three subcorpora and its documentation in open access were conducted by Naomi Truan only, but the decisions behind the use of specific tags within the TEI framework were made in concertation with Laurent Romary. While the focus of this research project and the reasons for the comparison of these three linguacultures[1] will not be discussed in this paper, it appears necessary to sketch out the context in which corpus building has materialized. 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 TEI is a robust procedure for encoding parliamentary corpora.

2.1 Why Parliamentary Debates?

6 In political discourse, parliamentary debates have recently raised the interest of linguists (Burkhardt and Pape 2000; Burkhardt 2003; Ihalainen, Ilie, and Palonen 2016), especially because they are particularly valuable 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). In this context, parliamentary debates increasingly become an object of linguistic annotation (see Fišer and Lenardič, 2018 for an overview of CLARIN parliamentary corpora on which the corpora addressed here are also listed).

7 While we do not engage in a debate on whether parliamentary discourse is of intrinsic research value, we believe that records of parliamentary interaction yield valuable insights in linguistic analysis. First, in most Western countries, parliamentary debates are publicly available in several complementary formats: video, audio, and text. 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.

8 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 problems 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) (see also 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 unauthorized turns, verifications with the video recordings are sometimes impossible since the camera focuses on the speaker and very rarely on the co-interlocutors.

9 Second, being at the interface between spoken and written data, 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). Even if official proceedings/transcripts do not adequately capture the interactional nature of the events such as pauses and hesitations (see Cribb and Rochford [2018] for an example based on the House of Commons), we adopted a TEI structure based on spoken data rather than drama-oriented data in order to allow for further projects that would take into consideration the interactional nature of plenary sessions. 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 Burnett and Bonami [2019] for the Assemblée nationale).

2.2 Why a New Annotated Corpus of Parliamentary Debates?

10 As sketched above, corpus studies based on parliamentary interaction have become numerous in the last decade. Against this background, what can a new annotated corpus of parliamentary data contribute? Why not work with already available parliamentary corpora? While reference corpora such as the Hansard corpus, which consists of British Parliament speeches between 1803 and

2005 (1.6 billion words, 7.5 million speeches), would offer statistically robust results with corpus-assisted techniques, they do not give access to the whole co-text[2] because of property rights. Moreover, no equivalent corpora for the German Bundestag and the French Assemblée nationale were available when the project started (2015–2016).

11 It should be noted, however, that several projects involving parliamentary data have been launched 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 (Blätte 2017a, Blätte 2017b). Apart from the fact that the period covered by the corpus is by no means comparable to the British House of Commons, it 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 other annotated French parliamentary debates was also created (Diwersy, Frontini, and Luxardo 2018) after the first release of the corpus in November 2016 (see sec. 6 for more detail on the platform that hosts the corpora).

12 Parla-CLARIN,[3] a comprehensive project aiming "to develop a TEI customisation for annotating parliamentary debates" by "storing and interchanging linguistically annotated corpora of parliamentary data to be used in scholarly research,"[4] was launched in 2018, two years after the open-access release of the corpora. Similarly, ParlaMint: Towards Comparable Parliamentary Corpora,[5] a project funded by CLARIN, "is a multilingual set of comparable corpora containing parliamentary debates mostly starting at the end of 2015 and extending to mid-2020" (Erjavec et al. 2020, Description). ParlaMint was not, however, available when Naomi Truan started her PhD (2015), as it was launched in July 2020. While these projects offer important and valuable sources of comparison, the annotation scheme described in this paper was conceived before they were launched.

13 Most of the projects presented above differ significantly from the small-scale contrastive project which is the focus of this paper, however, as they involve teams and infrastructures, while the TEI annotation presented here has been implemented by one person only (the first author). The annotation scheme used for the analysis (Truan 2018, 2021) not only invites extension and possibly revision, but also offers a point of entry for further (doctoral) projects working on small

specialized corpora, thus showing what can be annotated for specific research purposes and with limited means. Moreover, small-scale annotation schemes offer other advantages: for example, the possibility of encoding the variable majority/opposition, which had not been implemented otherwise prior to the annotation scheme presented here (Truan 2019, 45), as it needs to be done manually.

14 The variety of sources and formats is a strong point in favor 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.[6]

15 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.

16 We carried out the process of encoding in TEI 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,[7] 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

17 The rationale behind the constitution of "small monolingual corpora"[8] (Koester 2010) is to allow for the interaction between statistical measures and a close-reading analysis that is sensitive to the sociopolitical context in which parliamentary interaction takes place. In order to ensure that

external variables that may shape parliamentary talk are assessed appropriately, the research project that builds the basis for the annotation scheme focused on a limited range of national debates concerning a major European Council meeting (see Truan 2021, chap. 4).

18 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 thematic continuity, Naomi Truan selected one plenary debate per year held between 1998 and 2015 in the British, German, and French national parliaments, respectively, about a major European Council meeting (either before or after the meeting or on the same day). As Auel and Raunio (2014, 17) stress, "[p]roblematic 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 because 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.

19 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 recognize 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 stabilized, routinized and conventionalized than everyday interactions, it can be posited that genres function as an intermediary level of representativeness prior to analysis or as a first step toward the comparison of discourse communities that should be the basis of expectations of a contrastive discourse analysis (see von Münchow 2010).

20 While the annotation scheme described in this paper presents typical features of parliamentary interaction, it also represents a first step toward integrating 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 that is valid and applicable to a large variety of linguistic data and metadata (see also Truan 2019 for methodological reflections on contrastive discourse analysis).

3. Preventing the Built-in Obsolescence of the Corpus

21 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. 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 TEI. In order for this paper to be received outside the TEI community as well, we first briefly 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

22 The Text Encoding Initiative (see Romary 2008) 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 six hundred XML elements. At the core of the TEI Guidelines resides the principle that any TEI-based project should define its own subset (or *customization*) where the elements which are deemed useful for the representational task at hand are selected, documented, and possibly amended.

23 In the context of small specialized corpora, TEI annotation can be 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). TEI XML annotation enables researchers to fruitfully visualize the articulation between text and context—that is, between the plenary session and the metadata associated with it.

Interpretative data is situated within the corpus using dedicated TEI elements. As we will detail in section 6, the corpus is available under a CC BY 4.0 license, which enables anyone to correct or extend the metadata if necessary.

24 Based on this general understanding, we have conceived the annotation framework with this contrastive research question in mind: the subset we have devised consists of 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, "The TEI and XML") highlights for TEI. In this sense, and despite the fact that the political context changes over time between France, Germany, and the United Kingdom, TEI gives access to a common technical, practical, and methodological framework between the three subcorpora and the three languages.

3.2 A Sustainable Corpus

25 In designing the TEI-based encoding scheme of our corpus, we intended for it to be easy for other scholars to take it up to carry out various types of research, and also to allow its possible extension (in terms of coverage) or enrichment (e.g., with additional annotated features). Although we would avoid the term *reference corpus*, which is more applicable to large-scale endeavors to build up a representative sample for a language (see, e.g., Kupietz et al. 2010), we strove to create a sustainable corpus that may be combined in time and space with other endeavors to describe language resources in a variety of contexts and for a variety of genres. Within this framework, we saw adopting a sampling strategy focused on our research question not as a restriction in the constitution of the corpus, but rather as a route to a better grasp of the parameters for the linguistic analysis and thus for the encoding.

26 With this perspective in mind, we chose the TEI Guidelines as a basis for the encoding scheme because of 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,[9] existing corpora have been designed mainly on the basis of proprietary formats ranging from flat plain-text representations (Kapočiūtė-Dzikienė, Šarkutė, and Utka 2017; Frantzi 2018) 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 finding consensus to cover similar use cases[10] rather than *ad hoc* solutions. Besides, even for those corpora conforming to the TEI Guidelines, there are some strong discrepancies in the actual TEI encoding styles: whereas some (Research Group of Computational Linguistics, University of Tartu 2018) 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 a third group of researchers (Pančur, Šorn, and Erjavec 2018) who based their work upon the Transcription of Speech module of the TEI Guidelines.

27 The (internal) debate within the TEI community as to which module can optimally deal with parliamentary corpora, Drama or Transcription of Speech, relates to a more essential question: how should parliamentary debates be considered as a scholarly source? Three arguments plead, in our view, for an annotation as transcription of speech rather than drama. First, 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 preexisting script. Indeed, even if MPs may be reading from notes when participating in a parliamentary debate, *seul le prononcé fait foi*, that is, the transcription only records what is actually said.

28 Second, 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 aloud). While the resemblance between parliamentary debates and theater is attested (Ilie 2003), there is always room for improvisation, unplanned reactions, interventions, or comments in 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.

29 Third, although some parliamentary records appear to be strongly edited and may be seen as very close to written prose or drama in style or structure, we think it would go against a general effort toward interoperability to adopt, for a subset of the general corpus of parliamentary records, an encoding strategy that would be different from what is needed for more fine-grained transcriptions. As a matter of fact, the tagset for the transcription of spoken language of the TEI Guidelines does not imply that all details from the source must be encoded and one can implement, with a very small subset of the corresponding elements, exactly what could be achieved when adopting an encoding strategy based upon the tagset intended for drama.

30 For these three main reasons, we have adopted a TEI annotation scheme distinct from drama.

4. Enabling Sociolinguistic Explorations: The TEI Header

31 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, "The TEI Header"); second, the transcriptions of speech within the `<text>` element itself (for instance, the distribution of turns).

4.1 Political Speakers: The TEI Element `<person>`

32 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 (organization, place, and date encoded in `<settingDesc>`: see example 3) and with the speakers (name, sex, political party, political affiliation, and position encoded in `<particDesc>`: see example 1).[11]

33 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 (TEI Consortium 2022, chap. 15[12]), 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:

1. on the one hand, localizing speaker-related information at the utterance level, at the risk of being insufficiently generic, introducing redundancy, and above all introducing contradictory information throughout the document, when annotation is not carried out consistently;
2. 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 accounting for the variation of, for instance, political roles over time and across parliamentary debates.[13]

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

35 Our documentation strategy has been determined by our fundamental decision within our corpus to fragment parliamentary debates into document units corresponding to plenary sessions, with the additional advantage of optimizing 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 TEI XML document corresponds to one plenary debate as a communicative unit, that is, 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.[14]

36 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:

1. it makes the TEI document more readable 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);
2. 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;
3. it permits the development and extension of 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.

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

1. first, it ensures unique identification of the speakers[15] across their various interventions within a plenary debate;
2. 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.

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

```
<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>
```

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

39 The first group of features attached to the description of an MP within a plenary debate corresponds to stable—or very rarely varying—characteristics pertaining to the identification of the speakers according to long-term properties such as name (<persName>), sex (<sex> [16]), and nationality (<nationality>). The content of <sex> allows for simple comparisons such as length of speeches by gender (see Truan 2021, chap. 4).

40 The second group of features is more specific to each plenary debate and corresponds to the political characteristics of the speakers: their political affiliation (<affiliation>), their relation to current government (<floruit>,[17] with values "majority"and "opposition"), and the district that elected them (<residence>).

41 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 are able to gain insights into 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 (see Truan 2021, chap. 7). 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 manually encoded variables in the form of TEI constructs.

42 The annotation framework was geared toward 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 together parliamentary groups such as PDS/Die Linke that are coded as "Far Left" (see the use of <trait> in example 1) (for some observations on the variable majority/opposition in a Norwegian corpus, see Lapponi and Søyland 2016; Lapponi et al. 2018).

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

Example 2. Exemplifying a change in political party within a plenary debate.

```
<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>
```

4.2 The Speech Event: The TEI Element `<settingDesc>`

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

45 The TEI Guidelines provide a suitable way to do so within the TEI header by using the `<setting>` element within a `<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:

1. the name of the organization (`<orgName>` [18]) where the debate is taking place, namely the corresponding national parliament (for this corpus: House of Commons, Deutscher Bundestag, and Assemblée Nationale);
2. 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);

3. the name of the head of government in place (<persName>, [19] so that the debate can be easily understood in relation to a wider political context. We adopted a complementary numbering marker (e. g., Blair I, Blair II, Blair III, etc.) to signal successive governments with the same leader;
4. the actual legislative session (<name> [20]) within which the debate is taking place.

46 In addition to these generic political parameters, we added two specific descriptors intended to provide information about the European debate per se. In doing this, we pursued our general encoding strategy, and reused existing elements from the TEI Guidelines while slightly adapting their semantics as TEI components. For the description of the main topic(s) of the European Council meeting about which the national parliament is debating, we used the <activity> element. For the description of the place where the European Council meeting took place, we used the <locale> element. Both choices could probably be the most problematic ones if we were to carry out a wider dialogue with the scientific community on the standardization process and the encoding of parliamentary debates.

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

```
<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>
```

5. Encoding the Content

47 In this section, we present the decisions pertaining to the turn level section 5.1 as well as the intra-turn level section 5.2. Importantly, we do not address other levels of annotation such as word-level annotation that could have been marked up in TEI as well. Indeed, for the purpose of

this project, the results are based on automatic part-of-speech tagging. The open-source software TXM[21] (Heiden, Magué, and Pincemin 2010) used in this project indeed proceeds to language-specific part-of-speech tagging when a corpus is imported.

5.1 The Representation of Spoken Political Discourse: The Turn Level

5.1.1 Utterances/Turns

48 The element <u> (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 terminology. Turns are a superficial unit pertaining "to the surface structure of conversation" (Kerbrat-Orecchioni 2004, 8) since they solely indicate a change of speaker. We made this decision to account for the essentially monological nature of parliamentary interaction so that a specific speaker's turn 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>
```

5.1.2 Interruptions

49 Diwersy, Frontini, and Luxardo (2018) observe that the descriptor "speech type (debate, interruption, vote explanation, etc.)," which we do not use 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 unauthorized turns at the British parliament are only indicated as *interruptions* with no further information provided on the nature, source, or content of the disruption, as in the following example:

> 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[22])

50 Although the co-text sometimes gives insights on what kind of *interruption* occurred (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 unauthorized and authorized interventions, which enables us to retrieve automatically all the utterances of a given speaker.

5.2 Segments and Quotes: The Intra-turn Level

51 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>
```

52 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 by Schmidt (2011). This is due to both 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. 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

53 The Text Encoding Initiative has been, right from the onset, the basis for a strong open science vision, where interoperability would be at the service of sharing and reusing digital content encoded according to the TEI Guidelines (for an overview, see Romary 2020). For this reason we provide here an overview of our efforts to make the corpus FAIR ("Findable, Accessible, Identifiable, and Reusable"; Wilkinson et al. 2016).

54 As already alluded to, the corpus has been designed with the idea that it could be easily reused and complemented by others. It thus was in accord with our principles to adopt a completely open distribution setting by releasing it on the Ortolang platform (https://www.ortolang.fr/). Ortolang combines a number of important technical features:

1. specialization in linguistic data with the ability to attach several linguistic descriptors (language, genre, source type, etc.) to the corpus itself;
2. provision of unique identifiers to the resources;
3. long-term archiving for all uploaded resources;
4. version management, which allows the publication of corrections and improvements to the corpus while keeping the same underlying digital identity;
5. precise identification of the various contributors to a resource;
6. linking of resources with open licences—in our case a Creative Commons CC BY licence requiring proper attribution to the authors (CC BY 4.0) in `<publicationStmt>`;
7. finally, the ability to add an XSLT stylesheet to the corpus to provide a default search and presentation environment (in HTML).

55 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 endeavors. As a consequence, we decided to distribute all the source documents rather than limiting access through, for example, a query interface, as is the case for the EuroParl corpus. 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 was available.[23]

56 The three corpora are available online at the following addresses:

- hdl.handle.net/11403/uk-parl for the British corpus;
- hdl.handle.net/11403/fr-parl for the French corpus;
- hdl.handle.net/11403/de-parl for the German corpus.

57 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. Online access to the corpus (or the three subcorpora) was opened in November 2016.

58 Since the open-access release of the corpora in November 2016, the corpora (including their documentation) have been downloaded (partly or as a whole) 113 times for the British corpus, 49 times for the German corpus, and 418 times for the French corpus, respectively as of 17 May 2022. The corpora are listed on the CLARIN website.[24] Moreover, they have been discussed in several other annotation projects (see, for instance, Blätte 2018 for German or Diwersy, Frontini, and Luxardo 2018 for French) and used as a comparison corpus for other research projects (see, for instance, Stefanowitsch 2019, Zinn and Müller 2021, Piquer Martinez 2022). These examples show that an early open-access release (November 2016) of the annotated corpora together with their documentation, long before the dissertation project was submitted (October 2018), is beneficial to the community.

7. Conclusion

59 This paper has suggested an integrative and comprehensive approach to 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 TEI framework. We have 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 have also pleaded for an easily reproducible cross-

linguistic annotation framework. Specifically, we have shown that including metadata such as *political affiliation* or the distinction between majority and opposition is crucial to allowing for the comparison between several parliamentary systems.

60 We understand this paper as a first step toward the annotation of parliamentary corpora on a larger scale. We recognize that the small size of the corpora (from approximately 137,000 tokens for the French corpus to 417,000 for the German corpus) allowed for fine-grained annotation that may be more difficult to implement on a larger scale. Accordingly, the application of this annotation scheme 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), "so that a user at a particular point in the report can link through to the audio recording effortlessly and accurately." A more precise linkage between the videos and the transcripts could also enable insightful annotation in terms of kinesics—a dimension which, arguably, would adequately complete a close-reading discourse-analytic endeavor.

61 These further extensions and exploitations of the annotated corpora are at the core of our understanding of annotation as a process rather than a finished product (see also Bucholtz 2000) for a similarly reasoned argument in terms of "the politics of transcription"). In doing science in the digital age it is essential to make decisions explicit, transparent, and replicable. The annotation scheme developed in this project is only a first step.

BIBLIOGRAPHY

Auel, Katrin, and Tapio Raunio. 2014. "Debating the State of the Union? Comparing Parliamentary Debates on EU Issues in Finland, France, Germany and the United Kingdom." *Journal of Legislative Studies* 20 (1): 13–28.

Bayley, Paul, ed. 2004. *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam: John Benjamins.

Blätte, Andreas. 2017a. *GermaParl: Corpus of Plenary Protocols of the German Bundestag*. R data package, v. 1.0.4. http://polmine.sowi.uni-due.de/packages/src/contrib/GermaParl_1.0.4.tar.gz.

———. 2017b. *GermaParl: Linguistically Annotated and Indexed Corpus of Plenary Protocols of the German Bundestag*. CWB corpus v. 1.0.4. 10.5281/zenodo.3735141, https://doi.org/10.5281/zenodo.3735141.

Blätte, Andreas, and Andre Blessing. 2018. "The GermaParl Corpus of Parliamentary Protocols." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 810–16. N.p.: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2018/summaries/1024.html.

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

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

Burkhardt, Armin, and Kornelia Pape, eds. 2000. *Sprache des deutschen Parlamentarismus: Studien zu 150 Jahren parlamentarischer Kommunikation*. Wiesbaden: Springer.

Burnard, Lou. 2014. *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Encyclopédie numérique. Marseille: OpenEdition Press. doi:10.4000/books.oep.426.

Burnett, Heather, and Olivier Bonami. 2019. "Linguistic Prescription, Ideological Structure, and the Actuation of Linguistic Changes: Grammatical Gender in French Parliamentary Debates." *Language in Society* 48 (1): 65–93. doi:10.1017/S0047404518001161.

Cribb, V. Michael, and Shivani Rochford. 2018. "The Transcription and Representation of Spoken Political Discourse in the UK House of Commons." *International Journal of English Linguistics* 8 (2): 1–14. doi:10.5539/ijel.v8n2p1.

Diwersy, Sascha, Francesca Frontini, and Giancarlo Luxardo. 2018. "The Parliamentary Debates as a Resource for the Textometric Study of the French Political Discourse." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 73–77. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/22_W2.html.

Erjavec, Tomaž, Vladislava Grigorova, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Andrej Pančur, Michał Rudolf, and Kiril Simov. 2020. *Multilingual Comparable Corpora of Parliamentary Debates ParlaMint* 1.0. October 15, 2020. CLARIN ERIC. https://hdl.handle.net/11356/1345.

Fišer, Darja, and Jakob Lenardič. 2018. "CLARIN Corpora for Parliamentary Discourse Research." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 2–7. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/14_W2.html.

Frantzi, Katerina T. 2018. "Tour de CLARIN: Clarin:el Presents the Hellenic Parliament Sittings and Hellenic Parliamentary Corpus H-ParCo." Blog post, edited by Maria Gavriilidou. July 30. https://www.clarin.eu/blog/tour-de-clarin-clarinel-presents-hellenic-parliament-sittings-and-hellenic-parliamentary-corpus.

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

Gartner, Richard. 2013. "Parliamentary Metadata Language: An XML Approach to Integrated Metadata for Legislative Proceedings." *Journal of Library Metadata* 13 (1): 17–35. doi:10.1080/19386389.2013.778728.

Hansen, Dorte Haltrup. 2018. *The Danish Parliament Corpus 2009–2017, v1*. CLARIN-DK-UCPH Centre Repository. Copenhagen: Centre for Language Technology, NorS, University of Copenhagen; The Danish Parliament. http://hdl.handle.net/20.500.12115/8.

Heiden, Serge. 2010. "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–98. Waseda, Japan: Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764/document; https://aclanthology.org/Y10-1044/.

Ide, Nancy, and James Pustejovsky, eds. 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer.

Ihalainen, Pasi, Cornelia Ilie, and Kari Palonen, eds. 2016. *Parliament and Parliamentarism: A Comparative History of a European Concept*. New York: Berghahn.

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

———. 2006. "Parliamentary Discourses." In *Encyclopedia of Language & Linguistics*, edited by Keith Brown, 2nd ed., 188–96. Oxford: Elsevier.

Ilie, Cornelia, ed. 2010. *European Parliaments under Scrutiny: Discourse Strategies and Interaction Practices*. Amsterdam: John Benjamins.

Kapočiūtė-Dzikienė, Jurgita, Ligita Šarkutė, and Andrius Utka. 2017. *Lithuanian Parliament Corpus for Authorship Attribution. CLARIN-LT Digital Library in the Republic of Lithuania*. [Kaunas, Lithuania]: Vytautas Magnus University. http://hdl.handle.net/20.500.11821/17.

Kerbrat-Orecchioni, Catherine. 1990. *Les interactions verbales*. Tome 1. Paris: Armand Colin.

Kerbrat-Orecchioni, Catherine. 2004. "Introducing Polylogue." *Journal of Pragmatics* 36 (1): 1–24. doi:10.1016/S0378-2166(03)00034-1.

Koester, Almut. 2010. "Building Small Specialised Corpora." In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 66–79. London: Routledge.

Krzeszowski, Tomasz P. 1989. "Towards a Typology of Contrastive Studies." In *Contrastive Pragmatics*, edited by Wieslaw Oleksy, 55–72. Pragmatics & Beyond New Series 3. Amsterdam: John Benjamins.

Kupietz, Marc, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. "The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 1848–54. N.p.: European Language Resources Association (ELRA). https://aclanthology.org/L10-1285/.

Lapponi, Emanuele, and Martin G. Søyland. 2016. *Talk of Norway. Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository*. [Oslo]: University of Oslo. http://hdl.handle.net/11509/123.

Lapponi, Emanuele, Martin G. Søyland, Erik Velldal, and Stephan Oepen. 2018. "The Talk of Norway: A Richly Annotated Corpus of the Norwegian Parliament, 1998–2016." *Language Resources and Evaluation* 52 (3): 873–93. doi:10.1007/s10579-018-9411-5.

Leech, Geoffrey. 2013. "Introducing Corpus Annotation." In *Corpus Annotation*, edited by Roger Garside, Geoffrey N. Leech, and Tony McEnery, 1–18. London: Routledge.

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

Münchow, Patricia von. 2010. "Langue, discours, culture: quelle articulation? (1ère partie)." In "Visions du monde et spécificité des discours," ed. Mioara Codleanu and Sandina Iulia Vasile, special issue, *Signes, discours et sociétés: Revue semestrielle en sciences humaines et sociales dédiée à l'analyse des Discours* 4. http://revue-signes.gsu.edu.tr/article/-LXz7yiZKgVO69fy49uT.

Olschewski, Andreas. 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, Andrej, Mojca Šorn, and Tomaž Erjavec. 2018. "SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 8–14. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html.

Piquer Martinez, Jose. 2022. *Corpus of Political Speeches: Policy responses to the Great Recession in the United Kingdom and Spain (2008–2014)* [Dataset]. Department of Politics And International Studies. https://doi.org/10.17863/CAM.79047. https://www.repository.cam.ac.uk/handle/1810/332648.

Pražák, Aleš and Luboš Šmídl. 2012. *Czech Parliament Meetings*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Pilsen, Czech Republic]: University of West Bohemia, Department of Cybernetics. http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Research Group of Computational Linguistics, University of Tartu. 2018. *Reference Corpus of Estonian: Transcripts of Riigikogu (Estonian Parliament)*. Last modified December 21, 2018. https://www.cl.ut.ee/korpused/segakorpus/riigikogu/.

Risager, Karen. 2014. "The language-culture nexus in transnational perspective." In *The Routledge Handbook of Language and Culture*, edited by Farzad Sharifian, 87–99. London: Routledge.

Romary, Laurent. 2008. "Questions & Answers for TEI Newcomers." *Jahrbuch für Computerphilologie* 10: 69–90. https://hal.archives-ouvertes.fr/hal-00348372; http://computerphilologie.digital-humanities.de/jahrbuch/jb10-content.html; http://computerphilologie.de/jg08/romary.pdf.

Romary, Laurent, and Patrice Lopez. 2015. "GROBID - Information Extraction from Scientific Publications." *ERCIM News* 100 (January). https://hal.inria.fr/hal-01673305/document; https://ercim-news.ercim.eu/en100/r-i/grobid-information-extraction-from-scientific-publications.

Romary, Laurent. 2020. "TEI Guidelines: Born to be Open." ACDH-CH (Austrian Centre for Digital Humanities and Cultural Heritage) Lecture 6.1, June 10, 2020, Vienna, Austria. https://hal.inria.fr/hal-02864525.

Schmidt, Thomas. 2011. "A TEI-Based Approach to Standardising Spoken Language Transcription." *Journal of the Text Encoding Initiative* 1. https://journals.openedition.org/jtei/142; doi:10.4000/jtei.142.

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

Stefanowitsch, Anatol. 2019. "Delivering a Brexit Deal to the British People: Theresa May as a Reluctant Populist." *Zeitschrift für Anglistik und Amerikanistik* 67 (3): 231–63. https://doi.org/10.1515/zaa-2019-0022.

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.4.0. Last updated April 29 2022. https://www.tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/.

Truan, Naomi. 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.

———. 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.

———. 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.

———. 2018. "'Who Are You Talking About?.' The Pragmatics of Third-Person Referring Expressions: A Contrastive Corpus-Based Study of British, German, and French Parliamentary Debates." PhD diss., Sorbonne Université and Freie Universität Berlin.

———. 2019. "Möglichkeiten und Herausforderungen einer pragmatisch orientierten kontrastiven Diskursanalyse: Ein Vorschlag am Beispiel deutscher, französischer und britischer Parlamentsdebatten." *Diskurse – digital* 1 (3): 29–50.

———. 2021. *The Politics of Person Reference: Third-person Forms in English, German, and French*. Pragmatics & Beyond New Series 320. Amsterdam: John Benjamins. doi:10.1075/pbns.320.

Vitali, Fabio, and Flavio Zeni. 2007. "Towards a Country-Independent Data Format: The Akoma Ntoso Experience." In *Proceedings of the V Legislative XML Workshop*, edited by Carlo Biagioli, Enrico Francesconi, and Giovanni Sartor, 67–86. Florence: European Press Academic Publishing.

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

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

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, article no. 160018. 10.1038/sdata.2016.18, https://doi.org/10.1038/sdata.2016.18.

Zima, Elisabeth, Geert Brône, and Kurt Feyaerts. 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: John Benjamins.

Zinn, Jens O., and Marcus Müller. 2021. "Understanding discourse and language of risk." *Journal of Risk Research*. 1–14. https://doi.org/10/gnwxbv.

NOTES

1 A concept that refers to culture in language or the cultural dimensions of language (see Risager 2014).

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 Tomaž Erjavec and Andrej Pančur, *Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings*, v. 0.2, accessed February 8, 2021, https://clarin-eric.github.io/parla-clarin.

4 Tomaž Erjavec and Andrej Pančur, *Parla-CLARIN GitHub repository*, accessed February 8, 2021, https://github.com/clarin-eric/parla-clarin/.

5 Accessed February 8, 2021, https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora.

6 All files are from the *Compte rendu intégral* (and not from the *Compte rendu analytique* when it exists).

7 GROBID is a machine-learning library, with a particular focus on technical and scientific publications, for extracting, parsing, and re-structuring raw documents such as PDFs into structured TEI-encoded documents. For a detailed description, see Romary and Lopez (2015).

8 The German corpus is the largest, with 417,095 tokens. The British corpus contains 188,913 tokens, the French one 137,620.

9 "Parliamentary Corpora," accessed February 8, 2021, https://www.clarin.eu/resource-families/parliamentary-corpora.

10 As another illustration of this, we can mention what the TEI Lex-0 has aimed to achieve in the domain of interoperable lexical data and for which it was awarded the Rahtz Prize for TEI Ingenuity in 2020 (see Eliza Papaki, "DARIAH Working Group on Lexical Resources Wins Innovation Prize," November 20, 2020, https://www.dariah.eu/2020/11/20/dariah-working-group-on-lexical-resources-wins-innovation-prize/).

11 For a comprehensive mapping of all the TEI tags used in this work and how they have been applied to parliamentary debates specifically, see Truan 2016a, 2016b, and 2016c.

12 Accessed June 6, 2022, https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/CC.html

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

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

15 Speakers are mostly, but not only, MPs. For instance, in France, members of the government who are not members of the parliament can be invited to make a speech.

16 Although the TEI documentation reports on a `@value` attribute to normalize the corresponding content of the `<sex>` element, it does not provide a real standardized set of values as reference (TEI Consortium 2022, Appendix C: Elements, `<sex>`, https://tei-c.org/Vault/P5/4.4.0/doc/tei-p5-doc/en/html/ref-sex.html). We thus discarded this attribute in our encoding, but we used normalized values within the corpus (male/female/none).

17 Although the description of the `<floruit>` element in the TEI Guidelines may suggest that `<floruit>` should remain limited to the description of a temporal time span, we consider it acceptable to extend this description to the nature of the activity of the person in the given time span, especially when this activity may change over time, as is the case for the variable majority/opposition in the political sphere.

18 It might have been preferable to actually indicate the organization (`<org type="parliament">`) rather than just the name, but the corresponding element is currently not allowed in `<setting>`.

19 A further development of general guidelines for such encodings should lead to agreement on a `@type` for such elements (e.g., `@type="governmentHead"`)

20 Here also we could think of adding a `@type` to the corresponding `<name>` element within a larger standardization context, and possibly link this with a reference document of legislature events.

21 Accessed February 5, 2021, http://textometrie.ens-lyon.fr/?lang=en.

22 The source of the utterance (or metadata) is to be found at the end of the excerpt in round brackets: ISO 3166 country code (DE for Germany, FR for France, UK for the United Kingdom), year.month.day of the debate.

23 One of the points of contention could be that the dissemination through Ortolang is not fully open-source, as it would be through such a platform as GitHub. We see GitHub, which is a private platform and thus does not fulfill 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 publication setting.

24 "Parliamentary Corpora," accessed February 8, 2021, https://www.clarin.eu/resource-families/parliamentary-corpora.

AUTHORS

NAOMI TRUAN

Naomi Truan is a Research and Teaching Fellow in German Linguistics at the University of Leipzig, Germany. Her research interests include political discourse, digital interactions, multilingual practices, educational contexts, and language ideologies. Her first book, *The Politics of Person Reference* (2021), explores various instantiations of the third person in British, German, and French parliamentary debates. As an Alumna of the Open Science Fellows Program, she is committed to open science and science communication. You can find her publications in open access here: https://cv.archives-ouvertes.fr/naomi-truan.

LAURENT ROMARY

Laurent Romary is Senior Researcher at Inria, France, and former initiator and director general of the DARIAH European infrastructure. He carries out research on the modeling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He has been active in standardization activities with ISO, as chair of the ISO/TC 37/SC 4 (2002–2014) and ISO/TC 37 (2016–) committees, and with the Text Encoding Initiative, as member (2001–2011) and chair (2008–2011) of its technical council. He has been involved in scientific information (now called open science) policies and corresponding infrastructure deployments (HAL and Episciences) since 2005 within various research-performing organizations.