Creating and analyzing multilingual parliamentary corpora

Erzsébet Tóth-Czifra, Naomi Truan

Research Data Management Workflows
Volume 1: Creating and analyzing multilingual parliamentary corpora

Document initiated by Erzsébet Tóth-Czifra, Open Science Officer, DARIAH Coordination Office Berlin, Germany, and conceived and prepared in collaboration with Naomi Truan, University of Leipzig, Germany
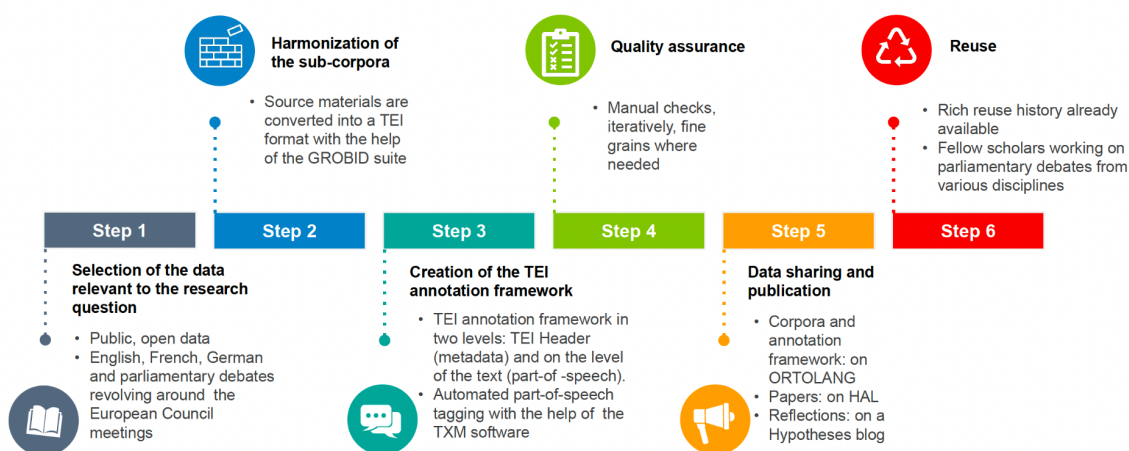
Overview of the data workflow 'Creating and analyzing multilingual parliamentary corpora'
[Download the schema as a PDF on HAL-SHS to make it machine-readable]

In this resource, you can follow a step-by-step description of a research data workflow involving the annotation of multilingual parliamentary corpora (French, German, British) according to the guidelines of the Text Encoding Initiative (TEI). Read further if you are interested in working with the TEI, analyzing parliamentary corpora, or simply would like to see a validated example of how FAIR and open data is implemented in the context of a PhD dissertation in Corpus Linguistics.

How to read this document

This document is the first instance of the research data workflow series provided by the [DARIAH Research Data Management Working Group](#). As research is becoming increasingly digital and **data-driven**, there is a growing pressure on researchers and research teams to find ways to enable technological and cultural **compliance with European-level and national data policies**. Research data management emerged to be a new field of expertise to explore and establish in all range of disciplines. Clearly, the generic research data management guidelines do not always align well with the cultural, conceptual and epistemological complexity of research data in the arts and humanities.

Who are we?

As members of the DARIAH Research Data Management Working Group coordinated by Erzsébet Tóth-Czifra and Marta Błaszczyńska, **we believe that data management practices work best if they are firmly grounded in actual research realities**. Therefore, through the publication of a series of data workflows that summarize **experiences and know-hows from concrete research projects**, our aim is to open the black box of data curation on very specific levels and to provide inspiration to fellow scholars who are working within the same discipline and/or data type.

For whom is this document?

The target audience of this document are all researchers who are potential reusers of part of the data and the workflow summarized below. This includes those who are working with TEI annotations or are interested in working with them, as well as those who work or aim to work with parliamentary corpora.

The document is *not* to be read as a best practice in the sense of 'these are the solutions you need to follow under any circumstances', but rather as a **concrete application of how to implement reflections on Open Science when gathering and collecting your data**, no matter for which purpose and where you come from.

When she began her PhD, Naomi Truan, whose completed project is here presented, had a background in German studies with a focus on the qualitative analysis of texts and no background on neither Digital Humanities (the use of digital tools in humanities), nor on Open Science (the practice of sharing not only one's research findings, but also one's data, for instance).

This Research Data Management Workflow shows that the implementation of **reproducible**, **standardized, and open data practices** is accessible to anyone.

Overview

What?

Creating and analyzing multilingual parliamentary corpora

Who?

Dr. Naomi Truan, University of Leipzig (Germany)
naomi.truan@uni-leipzig.de

Methods

A contrastive, corpus-based linguistic study of French, German and British parliamentary corpora.

Analysis is conducted through a TEI annotation (Text Encoding Initiative) scheme that offers a standardized, still flexible framework that is suitable for the contrastive focus of the study. This includes annotating both elements within the text and the metadata associated with it.

The analysis is then conducted both qualitatively and quantitatively, notably by using the open-source software TXM.

Project

PhD dissertation (Truan 2018) and the resulting book (Truan 2021) as well as several related papers, all available in open access on the HAL-SHS open repository of the author.

No funding for the annotation framework specifically. The work has been carried out by a single person as part of a PhD project.
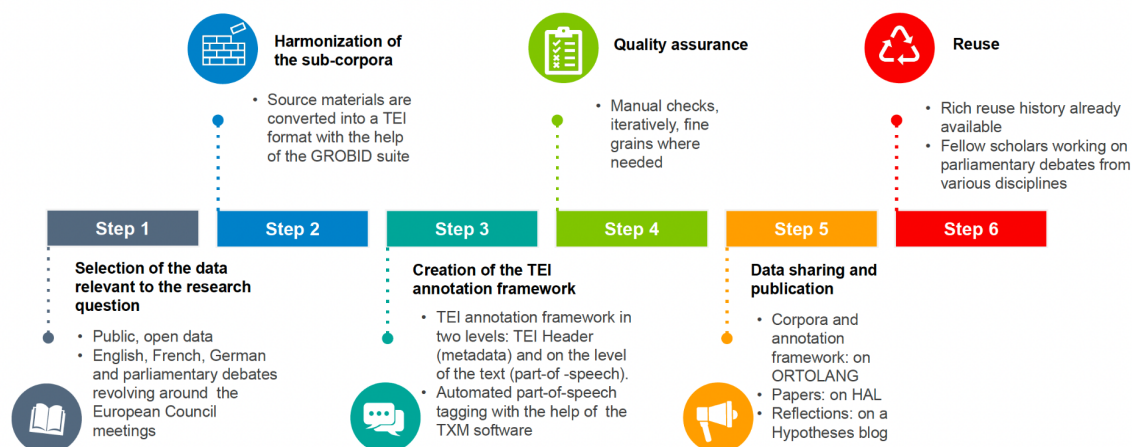
Key values

small-scale, well-documented, open and reproducible workflow, rich reuse history, Open Science in practice, Research Data Management decisions are inherent part of the design and analysis

Keywords

corpus linguistics, contrastive analysis, parliamentary debates, Text Encoding Initiative (TEI), linguistic annotation, discourse analysis, XML mark-up

Workflow

Creating and analyzing multilingual parliamentary corpora

Harmonization of the sub-corpora
- Source materials are converted into a TEI format with the help of the GROBID suite

Quality assurance
- Manual checks, iteratively, fine grains where needed

Reuse
- Rich reuse history already available
- Fellow scholars working on parliamentary debates from various disciplines

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6

Step 1 — Selection of the data relevant to the research question
- Public, open data
- English, French, German and parliamentary debates revolving around the European Council meetings

Step 3 — Creation of the TEI annotation framework
- TEI annotation framework in two levels: TEI Header (metadata) and on the level of the text (part-of -speech).
- Automated part-of-speech tagging with the help of the TXM software

Step 5 — Data sharing and publication
- Corpora and annotation framework: on ORTOLANG
- Papers: on HAL
- Reflections: on a Hypotheses blog

Tóth-Czifra & Truan 2021 | CC BY 4.0

Overview of the data workflow 'Creating and analyzing multilingual parliamentary corpora'
[Download the schema as a PDF on HAL-SHS to make it machine-readable]

In the following pages, we detail the six steps by including excerpts of an interview of Naomi Truan conducted by Erzsébet Tóth-Czifra and Marta Błaszczyńska on 07.10.2020 as well as references and useful resources in boxes.

Step 1: Selection of the data relevant to the research question

To enable the cross-linguistic analysis of third-person reference based on empirical data, British, French, and German parliamentary debates are gathered for a contrastive corpus study. To ensure the comparability of the three national contexts, the selection was restricted to national debates revolving around the European Council meetings.

The analysis was limited to the official transcripts of parliamentary records; audio and video resources were consulted on a qualitative basis only.

> *"I did not engage with audio and video data extensively, which I could have, but chose not to because I did not need it for my aims. I was indeed working mostly in semantics and pragmatics and did not include multimodal elements into the analysis. If it had to be included, it would have been a whole different story because then you would have to add the time stamp to the XML-TEI annotation, etc."*

All the texts have been retrieved from the official websites of the respective parliaments where the transcripts are shared as public data.

Resources:

http://hansard.parliament.uk/ for the House of Commons (HTML)

http://pdok.bundestag.de/ for the German Bundestag (PDF)

http://archives.assemblee-nationale.fr/ for the Assemblée nationale (HTML)

Step 2: Harmonization of the sub-corpora

"[At the time of my first investigations, in 2014-2015, both the British House of Commons and the French Assemblée nationale displayed the parliamentary proceedings in HTML, which allows for a quick, easy, and accurate retrieval of the content. The German corpus, on the other hand, was 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. I 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."

Source materials have been converted into a TEI-XML format with the help of the GROBID suite:

- TEI stands for Text Encoding Initiative, a community of digital humanists who collectively define a type of XML format.
- XML stands for Extensible Markup Language, a markup language that defines encoding standards for documents in formats that are both human-readable and machine-readable.

"In particular, we used the GROBID software suite, 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."

Resources:

GROBID software suite: https://github.com/kermitt2/grobid.

Romary, Laurent & Patrice Lopez. 2015. GROBID - Information Extraction from Scientific Publications. ERCIM, Data Sharing and Re-Use. https://hal.inria.fr/hal-01673305/document.

Step 3: Creation of the TEI annotation framework

Text Encoding Initiative annotation framework enables two levels: the TEI Header (metadata) and the text level (utterances, quotes). The part-of-speech tagging allowing for CQP queries was automated and directly integrated into the open-source TXM software.

CQP stands for Corpus Query Processor, a project developed within the IMS Open Corpus Workbench (CWB). As indicated on the website, "The IMS Open CorpusWorkbench (CWB) is a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words) with linguistic annotations. Its central component is the flexible and efficient query processor CQP".[1]

The integration of CQP for specific linguistic queries and TEI as a general annotation framework for digital texts is specific to linguistics, but the questions it raises are central to all Digital Humanities fields.

"I 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."

Why use TEI standards?

- The TEI builds on and fosters established, **well-documented community standards** that are widely acknowledged in the author's field of study and above: corpus-based linguistic modelling and analysis.
- The TEI is **flexible enough** to allow for customized and easy-to-enrich annotations that are **tailored to a specific research question** (by contrast to ad hoc solutions such as a specific metadata schema for parliamentary debates).
- The thorough documentation of the decisions made to construct the annotation framework and its publication comes with the degree of scholarly **transparency** that enables **reproducibility** and **reusability** of the framework for other research settings and/or purposes.

Resources:

Romary, Laurent. 2009. Questions & Answers for TEI Newcomers. Mentis Verlag. https://hal.archives-ouvertes.fr/hal-00348372 (18 June, 2020).

Romary, Laurent. 2020. TEI guidelines: born to be open. *ACDH-CH: Austrian Centre for Digital Humanities and Cultural Heritage Lectures*, vol. 6. Lecture. Vienna. https://doi.org/hal-02864525. https://hal.inria.fr/hal-02864525 (8 February, 2021).

---

[1] See http://cwb.sourceforge.net/, accessed on 31.05.2021.

Step 3.1: Setting the scene

The TEI header places emphasis on identifying speakers in a way that is most appropriate and flexible for the sake of analysis (e.g. enabling the stable identification of speakers even if their political affiliation changes over the years). See p. 15 here.

Step 3.2: Encoding the content

The body of the XML-TEI corpora went through an automated, language-specific part-of-speech tagging with the help of Tree Tagger part-of-speech annotation tool directly integrated into the external open-source software TXM. You can follow the documentation of the import process and visualize the corpus through an XSL style sheet and an XML stylesheet here.

Output: .csv tables that were exported from the software TXM according to some specific queries / CQP requests (e.g. concordance lists and collocations with personal pronouns or nouns and nominal expressions).

One of the challenges was that the XML-TEI annotation was not as easily imported into the TXM software as we thought, which required email exchanges with the TXM team.

> "*We actually needed at least half a year, if not a whole year, to get TXM to recognize the TEI annotation. To make it work through the software, so to say. And for me, this was the most important thing, because I had had annotated my data set for quite a long time and knew that I wanted to work with the software TXM. But it didn't match. This is also why I worked a lot qualitatively—in the meantime, before we figured out how the corpus could be imported. I actually began working with the data by extracting examples directly from the TEI-XML files. But it's not the real workflow. I was waiting for someone to build the bridge.*"

Another challenge was that we wanted to optimize heterogeneous data for contrastive reuse. The three sub-corpora were available in different formats and offered different possibilities for coding. Naomi had to find a common denominator, a unified annotation framework that works equally well for all three of them but is also suitable for contrastive analysis. This comes with limitations in terms of richness of the annotation for those who wish to reuse only one sub-corpus, for instance, working only on the French one.

Resources:
Open-source software TXM: http://textometrie.ens-lyon.fr/?lang=en.
Heiden, Serge. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto & Yasunari Harada (eds.), 24th Pacific Asia Conference on Language, Information and Computation, 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764/document (16 March, 2017).

Truan, Naomi. "From TEI-XML to TXM, HTML and back" in Bag of tags, 26/12/2016, https://tags.hypotheses.org/81.

Truan, Naomi & Laurent Romary. forthcoming. Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative* (14). Preprint: ⟨halshs-03097333v2⟩.

Step 4: Quality assurance and import into TXM

Due to the relatively small size of the corpus and the qualitative nature of the analysis, quality assurance consisted of manual checks of the encoding performed in an iterative fashion. Also, since the analysis had been carried out as Naomi's PhD, there was no possibility of employing an assistant as co-annotator. As a next step, the TXM team provided support during the import of manual codes to TXM.

"And since the corpus was rather small and I felt like engaging with the data was part of my job anyways, I actually did everything by myself. I listened to a lot of podcasts [during the manual coding and the successive imports into TXM]. It was a great time, but it was not automatic at all."

Step 5: Data sharing and publication

Parallel to the development of the annotation framework, and before the publication of any paper making use of the annotated corpora, Naomi opted for the publication of the annotated corpora in Open Access. Her main motivations were scholarly transparency and reproducibility. In order to allow other scholars to trace and eventually reproduce the analysis, or parts of it, as it unfolded, Naomi shared a rich documentation of the outputs and throughput.

The corpora and the annotation framework are deposited in the ORTOLANG repository because:
- ORTOLANG is specifically designed for linguistic data with the possibility to attach several linguistic descriptors (language, genre, source type, etc.).
- As a complementary service to Huma-Num (very large research infrastructure), ORTOLANG exhibits standards of a trustable repository: long-term archiving, permanent identifiers (PIDs), version control, Creative Commons open licensing framework, precise identification of the various contributors to a resource.

The deposits contain:
1. The XML files annotated according to the guidelines of the Text Encoding Initiative (TEI).
2. The German corpus also includes the original PDF files of the transcriptions. For the other two corpora, a link is provided to the transcripts openly available in HTML format.

3. Documents providing 1) an overview of the debates being part of the corpora; 2) a generic and specific description of the annotation framework.

License: Creative Commons CC-BY license requiring proper attribution to the author (CC-BY 4.0). For more information on licenses, check this page: https://creativecommons.org/licenses/.

Screenshot of the file structure of the German corpus

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

The publication of the corpus is interlinked with research papers and complemented with additional documentation in the form of blog posts (see "Resources").

Publications in Open Access: Juggling with Different Formats

What kind of data?	What kind of platform?
Raw data ➤ Choose a license (as open as possible)!	**Open Access Repository** In Linguistics, see ORTOLANG (Open Resources and TOols for LANGuage), https://www.ortolang.fr/
Annotated Data ➤ Choose a license! ➤ Explain how and why (= to which purpose, with which research question in mind) the data has been annotated ➤ Attribution: cite the names of *all* people involved	**Open Access Repository** ➤ Not on a private platform (blog) ➤ Your data needs to be archived with key words and easily available to the academic community Contributors Naomi Truan — Annotator Laurent Romary (INRIA) — Consultant

- Do it as soon as possible to ensure **recognition** and **authorship**
- Don't be afraid that anyone will 'steal' anything: the likelihood that someone else will have the same research question(s) on the same corpus is extremely low...
- It is more likely that you will develop **cooperation projects**!

Publications in Open Access: Juggling with Different Formats

What kind of data?	What kind of platform?
Intermediary thoughts **Methodological Reflections** **Corrections** *afterwards* **Lectures / Seminars**	**Academic Blog** *Hypothèses* has a 'quote function' Cite this article as: Naomi Truan, "Some Useful Free Online Resources When You're Writing a PhD," in *Ici et là*, 07/10/2018, https://icietla.hypotheses.org/383. **Open Access Repository** Working documents
Journal Articles ➤ Post *all* published papers online, even if it's not the full text from the editor (preprint) ➤ On HAL-SHS, it is possible to indicate when the full version of a document should be released (i.e. embargo) ➤ You always have all the rights on a document as an author, the ONLY thing you may not be able to put online is the final *edited version*, often with correct pagination	**Open Access Repository** Always ask the editor if they allow it: ➤ Shows the editors that there is a need for open access, if you're an early-career researcher, mention the "publish or perish" pressure to get tenure ➤ Puts pressure on the editors ➤ Some may give you ambiguous answers, use it at your advantage

Resources:

Truan, Naomi. 2019. How to make the most of your publications in the humanities?" in the hands on session "Do-it-yourself Self-archiving". [Presentation slides]. https://www.fosteropenscience.eu/node/2605

Step 6: Reuse

A rich reuse history is already available and presumably, the reuse cases Naomi is aware of are only fractions of the real reuse instances. Typical reusers are fellow scholars working on parliamentary debates, with disciplines ranging from linguistics to political sciences and sociology.

access to parliamentary debates corpus
2 messages

To: naomi.truan@uni-leipzig.de
Dear Dr. Naomi Truan,

I am a PhD student at   University.
I am deeply interested in the corpus of parliamentary debates 1998-2015 I found on your publications.

I would be grateful if you could assist me with a reference to download the corpus. Is it please please possible?

I tried to look at HAL archives but found only the paper itself: *On the Pragmatics of Interjections in Parliamentary Interruptions*. I could not find the data :-(

Thank you so much in advance,

Best

Mail from a scholar in Israel requesting access to
'my' corpus of British parliamentary debates, 2019

A thematically specialized corpus such as the one prepared by Naomi Truan on the parliamentary discourse on Europe may offer significantly more detailed metadata and annotation (Truan, 2017).

Quote with reference to the ORTOLANG corpus (with
a date error...), in: Blätte & Blessing 2018

A user would like some additionnal informations about the resource uk-parl

Hi,

A user let you a message in relation with a resource that you manage:

[Parliamentary Debates on Europe at the House of Commons (1998-2015)]
Hi Naomi I would like to access the corpus for my research. We have compiled the Malaysia Hansard Corpus and is currently working on the mark-up. thank you. Prof

You can answer directly by answering this email sender address.

Sincerely,
The ORTOLANG Team.

Mail from a scholar in Malaysia requesting access to
'my' corpus of British parliamentary debates, 2019

Doing 'Interrupting' in Parliamentary Debates in British English, Finnish, French, and German: A Cross-Linguistic Perspective

The data for this project stem from the **parliamentary corpora** in **British English, Finnish, French, and German** taken from open access resources made available by several contributors and listed on the CLARIN infrastructure. The data consist of four mostly manually annotated corpora with TEI-XML markup, offering a ready-to-analyse dataset.

Collaboration with Dr. habil. Johanna Isosävi (University of Helsinki),
who uses 'my' corpora for new research questions
Project funded by the Ella and Georg Ehrnrooth Foundation

Feedback on the annotated corpora by Naomi Truan
after their publication in Open Access (2016)

Citations of the corpora in research projects reusing them

Source: Truan, Naomi. "Outils numériques pour les SHS", in *Ici et là*, 13/02/2020, https://icietla.hypotheses.org/1280. Slide 18 translated into English.

> *"And for instance, the fact that my corpus or my corpora have been downloaded 100 times and have been used for several different projects, although the corpus annotation did not go through a peer-review process at the time [it has in the meantime], shows me that this is valuable and that my standards actually are used and useful to the community."*

A reuse challenge: proper citation is sometimes missing from reuse cases even if there is a 'Cite as' snippet prominently displayed on the landing pages of the deposits. In order to cite data properly, have a look at these recommendations:

Resources:
Andreassen, Helene N., Andrea L. Berez-Kroeker, Laura Collister, Philipp Conzett, Christopher Cox,

Koenraad De Smedt, Bradley McDonnell & Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. https://doi.org/10.15497/rda00040. https://zenodo.org/record/3672840#.Xku-92j7Q2w (4 March, 2020).

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*. De Gruyter 56(1). 1–18. https://doi.org/10/gft4g7.

Tóth-Czifra, Erzsébet. 10 practical tips to fight against the culture of non-citation in the humanities in DARIAH Open, 29/02/2020, https://dariahopen.hypotheses.org/747

To conclude

To conclude, and also because in this data story, we aimed to initiate an honest discussion about data workflows and openness, it is worth highlighting that opening a bigger window on one's scholarly processes means a **significant amount of investment**, not only in terms of time and effort but **also ethically and even emotionally**. Working with sharing in mind from the beginning involves being ready to be exposed. As Naomi puts it:

> *"You have to feel ready to be exposed and to feel like vulnerability is part of your research. And for me, it has been amazing because I think sharing my data has opened a lot of new things. Now I am working more qualitatively as well, practicing autoethnography, and basically involving what research does to you. So I think that [opening my workflow and publishing my annotated corpora] has been very helpful. But it is hard. It is really hard because people do not always value it the way you would want them to."*

Still, what is also clear from Naomi's words is that sharing your data is feasible within the timeframe of a PhD project. Each of these windows, each of these open documentation components can create rich ways of engagement with the work and can thus advance one's career and position in the field. We hope that this document has shown you in a practical way the possible impacts of working on your project with sharing in mind.

References

Annotated corpora and annotation framework

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

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

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

Paper describing the TEI annotation framework

Truan, Naomi & Laurent Romary. forthcoming. Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A Cross-Linguistic Account. Journal of the Text Encoding Initiative (14). Preprint: ⟨halshs-03097333v2⟩.

The paper has been written with reusability considerations in mind: "Making decisions explicit, transparent, and replicable are primary prerequisites for doing science in the digital age."

Key publications

Truan, Naomi. 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 Dissertation. Sorbonne Université / Freie Universität Berlin.

Truan, Naomi. 2019. Talking about, for, and to the People: Populism and Representation in Parliamentary Debates on Europe. *Zeitschrift für Anglistik und Amerikanistik* 67(3). 307–337. https://doi.org/10.1515/zaa-2019-0025. ⟨halshs-02376375⟩

Truan, Naomi. 2020. Narratives of dialogue in parliamentary discourse: Constructing the ethos of the receptive politician. Journal of Language and Politics. https://doi.org/10.1075/jlp.20018.tru. ⟨hal-02988202⟩

Truan, Naomi. 2021. *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.

Softwares

Romary, Laurent & Patrice Lopez. 2015. GROBID - Information Extraction from Scientific Publications. ERCIM, Data Sharing and Re-Use. https://hal.inria.fr/hal-01673305/document.

Heiden, Serge. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto & Yasunari Harada (eds.), 24th Pacific Asia Conference on Language, Information and Computation, 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764/document (16 March, 2017).

Blog posts

Truan, Naomi. "From TEI-XML to TXM, HTML and back" in *Bag of tags*, 26/12/2016, https://tags.hypotheses.org/81

Truan, Naomi. "Je te donne, tu me donnes, nous nous donnons… nos données" in *Ici et là*, 03/07/2017, https://icietla.hypotheses.org/53.

Truan, Naomi. "Going Open Access is Like Transitioning Into a Vegetarian Lifestyle" in *Ici et là*, 21/01/2019, https://icietla.hypotheses.org/872.

Truan, Naomi. "Open Access: An Early Career Researcher Perspective" in *Ici et là*, 22/01/2019, https://icietla.hypotheses.org/994.

Truan, Naomi. "Outils numériques pour les SHS" in *Ici et là*, 13/02/2020, https://icietla.hypotheses.org/1280.

Truan, Naomi. "How sharing my data has changed how I write" in *Ici et là*, 24/03/2021, https://icietla.hypotheses.org/1929.

Glossary

CQP

CQP stands for Corpus Query Processor, a project developed within the IMS Open Corpus Workbench (CWB). As indicated on the website, "The IMS Open CorpusWorkbench (CWB) is a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words) with linguistic annotations. Its central component is the flexible and efficient query processor CQP".[2]

TEI

TEI stands for *Text Encoding Initiative*: comprising a set of community practices (as well as the communities themselves) to mark-up and encode textual scholarship in a collectively defined a type of XML format. For the official definition, see also: https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html (accessed on 31.08.2021).

TXM

TXM is an open source research analysis software that had been developed within the framework of the French Textométrie project and is designed to perform a new generation of textometrical research, in synergy with existing corpus technologies (Unicode, XML, TEI, NLP tools, CQP, R). You can learn more about the software here: http://textometrie.ens-lyon.fr/?lang=en (accessed on 31.08.2021).

---

[2] See http://cwb.sourceforge.net/, accessed on 31.05.2021.

XML

XML stands for Extensible Markup Language, a syntax used for defining mark-up languages for encoding documents in standardized formats that are both readable by humans and machines. For a commonly agreed definition, see also: https://en.wikipedia.org/wiki/XML (accessed on 31.08.2021).