TEI and Git in ParlaMint: Collaborative Development of Language Resources

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Matyáš Kopp
Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic
kopp@ufal.mff.cuni.cz

Katja Meden
Dept. of Knowledge Technologies, Jožef Stefan Institute,
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
katja.meden@ijs.si

Abstract

This paper discusses the encoding, validation and development of language resources in the completed ParlaMint I and on-going ParlaMint II CLARIN projects, which centre on the collaborative development of a large set of interoperable corpora of parliamentary proceedings. It focuses on the encoding of ParlaMint corpora, the GitHub development platform, and the evaluation of their use by project partners. We introduce the use of TEI for the encoding guidelines and validation schemas. We motivate and explain the use of Git and GitHub to develop and maintain the encoding schemas, validation and conversion scripts and samples of the corpora. The paper also presents the results of a survey on the use of TEI and Git in the ParlaMint projects among the project participants. Overall, participants were mostly positive about their experience with TEI and Git, although some difficulties were reported. These will serve as a basis for further TEI and Git optimisation in ParlaMint.

1 Introduction

ParlaMint is a CLARIN ERIC supported project[1] which, among other tasks, aims to produce a set of comparable and richly annotated corpora, involving a joint effort of a large number of partners. The concluded ParlaMint I project (2020–2021) already developed corpora containing transcriptions of the sessions of 17 European national parliaments in the time-span 2015–2021 (Erjavec et al., 2022). These corpora are about half a billion words in size, contain rich metadata on 11 thousand speakers, and are linguistically annotated. The on-going ParlaMint II project (2022–2023) plans to extend existing corpora with newer data and add corpora for 14 new, also regional, European parliaments. It will also enhance the corpora by providing machine translations to English, and, for a selected subset of corpora, add speech data, as well as work on the wider use of the corpora.

With a largely bottom-up project involving many partners, parliamentary systems and different sources but aiming to produce a large set of highly comparable corpora, it is important to have robust encoding guidelines, automated validation and a scalable and flexible data workflow. ParlaMint I was already using TEI and Git to achieve these goals, and the first tasks in ParlaMint II were to reevaluate and extend these aspects of the project.

[1] https://www.clarin.eu/parlamint

Tomaž Erjavec, Matyáš Kopp and Katja Meden 2023. TEI and Git in ParlaMint: Collaborative Development of Language Resources. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 44–56. DOI: https://doi.org/10.3384/ecp198005

The paper discusses the project's use of TEI, including developing the encoding guidelines, the formal XML schema, and the validation and conversion scripts (Section 2); the use of Git and GitHub as an open and controlled development environment (Section 3); the results and analysis of a survey conducted among the partners in ParlaMint II on their familiarity, use, and suggestions as to the use of TEI and Git (Section 4), ending with a summary and directions for further work (Section 5).

2 Encoding the ParlaMint Corpora

The definition of a common encoding for parliamentary corpora has a long history in the context of CLARIN, starting with the CLARIN Travelling Campus "Talk of Europe" where, in 2014 and 2015, three hackathons were organised using European Parliament proceedings curated as linked open data. In addition, interdisciplinary workshops on working with parliamentary records and co-locaated with the LREC conference were organised (2018, 2020, 2022) under the guidance of CLARIN. Finally, the Parla-CLARIN recommendations for encoding parliamentary corpora (Erjavec and Pančur, 2021)[2] were proposed at the "CLARIN ParlaFormat Workshop" in 2019.

Parla-CLARIN is a customisation of the TEI Guidelines (TEI Consortium, 2022)[3]. A TEI customisation is specified in a TEI ODD (One Document Does it All)[4] document, which serves a double function: it contains the prose guidelines, as well as the formal schema of the customisation, using the TEI ODD schema specification language. With the TEI XSLT stylesheets the prose guidelines can be converted to HTML for reading, while the ODD schema specification is converted into one of the standard XML schema languages, such as the ISO standard RelaxNG, and such an XML schema is then used for formal validation of the corpora. The design of the Parla-CLARIN recommendation was inspired by previous similar efforts, in particular the TEI Lex-0 encoding recommendations for dictionaries (Tasovac et al., 2018)[5] and the TEI schema for the multilingual ELTeC corpus containing 100 historical novels for a number of languages (Burnard et al., 2021; Schöch et al., 2021)[6].

Although the ParlaMint corpora conform to the Parla-CLARIN schema, we required, in order to ensure interoperability, a much more constrained encoding than the quite general one of Parla-CLARIN. To this end, we have, in ParlaMint I, developed a bespoke RelaxNG schema without using the ODD mechanism. The advantage of this approach is that the schema expresses exactly the kinds of constraints that we wished to make, while the disadvantage is that there were no guidelines accompanying the schema, which is, formally, not even TEI, exactly because the schema was not derived from a TEI ODD. For these reasons, we developed, in ParlaMint II, a ParlaMint ODD, which contains the prose guidelines and as well as the formal schema. The ODD schema allows only the elements and attributes that we wish to have in ParlaMint, however, still with richer content models than those required, as it is quite difficult (short of completely re-specifying the content models of all the elements, or introducing a special namespace) to forbid the multitude of element nestings otherwise allowed by TEI. We have also worked on the documentation of individual elements and attributes in the ODD schema, i.e. changing the default glosses and examples of use of the elements as they appear in the TEI to ParlaMint specific ones.

The ParlaMint schema (either the bespoke or the ODD-derived RelaxNG) is only the first step in the validation of the ParlaMint corpora. We have also developed an XSLT script that performs validation regarding the textual content of some elements and checks that redundant metadata (i.e. metadata which is encoded in several places in the corpora but which makes it easier to inspect or process the corpora further) is not contradictory. Another script checks that all cross-references are resolvable, i.e. that a corpus does not contain broken internal links.

[2] https://clarin-eric.github.io/parla-clarin/

[3] https://tei-c.org/guidelines/P5/

[4] https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/

[5] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[6] https://github.com/COST-ELTeC

Furthermore, as the corpora are converted to other formats, such conversions can also expose various errors in the corpus. One of the down-stream formats is CoNLL-U, and the Universal Dependencies validation tool[7] is used to check the validity of the linguistic analyses. The corpora converted to vertical files are mounted on CLARIN.SI concordancers, where the conversion from TEI, the corpus compilation log, and an analysis of corpora on the concordancers can also reveal problems.

The mark-up of a ParlaMint corpus is rather complex compared to most linguistically annotated corpora, first because of extensive metadata about the speakers and the political organisations (e.g., lower/upper house, political party), with temporal attributes attached to the metadata. We have also tried to retain many aspects of the original transcripts, in particular transcribers' comments, which are encoded in several ways, depending on their content. However, this complex encoding is explained and exemplified in the ParlaMint encoding guides and samples from existing corpora are directly viewable via GitHub, all with the intention of making it easier to come to grips with the encoding system,

3 Using Git and GitHub

Git has become the revision control system of choice for many software development projects, and has also proved its worth in the (collaborative) development of language resources, with the most prominent example being the Universal Dependencies treebanks and annotation guidelines (de Marneffe et al., 2021)[8]. It has also been used for the development of TEI customisations, e.g. the already mentioned TEI Lex-0 and ELTeC, for the latter used not only of the schema, but of the corpora as well.

Apart from support for collaborative development with transparent versioning and attribution, and simple comparisons of files, Git hosting platforms, such as GitHub, support social media aspects of development, in particular posting and discussing issues, commenting on commits or pull requests, and a Wiki space. It is also possible to directly publish the documentation of a project using GitHub Pages. Finally, running scripts at a particular point in the Git workflow is supported by GitHub Actions. All these features lead to a more controlled and better documented development process.

We had already used GitHub for the development and publishing of the Parla-CLARIN schema, where the TEI ODD is maintained on GitHub[9], the guidelines are published as GitHub pages[10], and technical instructions for using or further developing the schema are available on the GitHub Wiki[11]. In ParlaMint I, as well, the project development was to a large extent done on GitHub[12]. The Git repository contained the latest RelaxNG schemas for the corpora and the complete validation or transformations scripts, written mostly in XSLT (and some Perl). Problems with the proposed encoding schema were often discussed through GitHub issues, while problems with individual corpora were communicated mostly by email.

Git(Hub) is, of course, not the only revision control platform available, so the question is, why we have chosen to use exactly this option. The reasons are, to a large extent, pragmatic in nature. We are familiar with GitHub, so it is easier to use the platform we know; in ParlaMint I we already used various GitHub's features, so it also made sense to use the same platform in ParlaMint II, because the ParlaMint I partners were already familiar with the platform and established workflow there; finally, GitHub is owned by a very large company, so we can be reasonably sure it will be maintained in the foreseeable future.

But while Git(Hub) is well suited for developing, storing and publishing software tools, schemas, and even quite large hand-annotated corpora, the complete ParlaMint corpora are, in

[7] https://github.com/UniversalDependencies/tools

[8] https://universaldependencies.org/

[9] https://github.com/clarin-eric/parla-clarin/

[10] https://clarin-eric.github.io/parla-clarin/

[11] https://github.com/clarin-eric/parla-clarin/wiki

[12] https://github.com/clarin-eric/ParlaMint

practice, somewhat large to be stored in Git, and, even more so, GitHub. Apart from the sheer size of the corpora (over 240 GB for ParlaMint I) and large number of files (almost 160,000), this is also due to the fact that, say, a new round of automatic annotation changes almost all files, making such a commit a very slow process: an experiment with complete ParlaMint I corpora showed the initial staging and commit to take approximately 4 hours.

It should also be noted that we also use GitHub's social network features, so it is very helpful to be able to view the content of files as rendered in a web browser directly from GitHub. While we haven't found the exact file size limit in GitHub's documentation, our experience shows that it is possible to only view rendered files that are smaller than 2 MB. Another reason for working with a smaller portion of corpora is that annotating one corpus can take a few days, and the errors usually appear in most files so a small sample is sufficient for debugging and discussing doubts in encoding. Therefore we here opted for a compromise, namely, we developed a script that extracts only small samples from individual corpora, and maintain only these samples, also in derived formats, on GitHub. The script reduces the number of files as well as the amount of XML content in particular files so these samples can be directly viewed on GitHub. This gives an impression of how the corpora are structured and makes development more manageable because it is possible to refer to particular parts of files in issues.

The complete set of corpora is then made available only for a major release and deposited in the CLARIN.SI repository.

In ParlaMint II, the first step was to update the Parla-CLARIN GitHub to reflect the ParlaMint best practice, while the ParlaMint GitHub was extended with pages[13] which are used to publish the ParlaMint encoding guidelines. Apart from submitting the complete corpora, all the communication was done through GitHub issues, rather than via email, so that problems are documented, can be discussed and the solution linked to a commit.

At the time of writing, over 300 issues, with over 1,700 posts have been opened, with the majority already resolved. 52 different GitHub users contributed to creating an issue, pull request or responding to either. The most discussion was, unsurprisingly, on issues related to problems when merging the pull requests submitting samples for new corpora, with one having almost 70 comments and replies.

As we cooperate with many partners that are supposed to add their sample data with pull requests, a validation procedure for newly inserted data using GitHub Actions[14] has also been developed. Furthermore, when a pull request with valid data is merged into the correct branch, the TEI files are sampled, and derived formats are added to the repository. This approach has several benefits: there is no need for the partners to carefully sample the data themselves, they do not need to compile the derived formats, and we can be sure that the derived files are always up to date with corresponding TEI files.

For local validation and conversion of the complete corpora either by a partner or centrally, we have developed a validation procedure that uses the Unix `make`[15] tool. The Makefile is self-documenting for easier use, i.e. running `make` without arguments prints a list of the available targets, which e.g. check installed prerequisites, validate the corpus against the ParlaMint and Parla-CLARIN schemas, perform advanced content validation, and convert a corpus to derived formats. Finally, the Perl wrapper program is used to prepare a ParlaMint corpus for distribution: it finalizes the corpus header, runs all the validation steps, converts the corpus to derived encodings and packs the corpus, both as a "plain text" corpus and as a linguistically annotated one.

[13] https://clarin-eric.github.io/ParlaMint/

[14] https://docs.github.com/en/actions

[15] https://www.gnu.org/software/make/

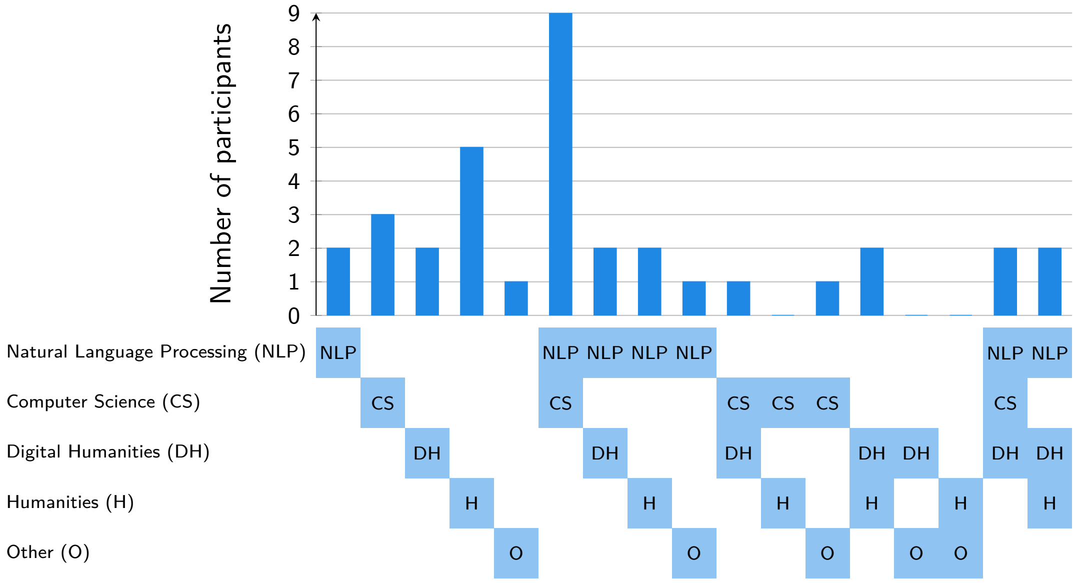

Figure 1: Distribution of research background combinations by individual participants (multiple answer question, 35 participants, 61 answers in total)

4 The Survey on TEI and Git

Since the development of the ParlaMint corpora involved numerous partners with varying degrees of familiarity with TEI and Git, we decided to solicit their feedback to assess the development pipeline and identify opportunities to further improve the workflow. To this end, we designed a questionnaire that included three main sections: an introductory section to identify the key characteristics of our project partners, a section on TEI encoding with questions about the encoding process and experiences with TEI, and finally a section on Git and GitHub that included questions about their familiarity with Git and GitHub. The survey ran from January 3 to January 20, 2023.

The analysis of the survey results is presented separately for TEI and Git. In both cases, we first present the results of the survey, based on which we conducted further analysis to explore various relationships between participants (and their research backgrounds) and their experiences with TEI before and after project work on ParlaMint.

Most responses were submitted by DK, SI (each 4 completed surveys) and IT (3 surveys), while AT, BA, GR, HR, HU, PL, RS, UA shared 2 completed surveys per country. We also received responses (1 completed survey per country) from BE, BG, EE, ES-PV[16], FR, IS, LV, NL, NO, PT, RO, SE, TR, giving us a total of 35 responses from 24 countries of 31 ParlaMint partner countries and regions, i.e. a response rate of 77%. 26 surveys were fully completed (0.75%) and 9 were partially completed (0.25%), with "partially completed" being those surveys where at least one question was answered.

Of the respondents, just over half (54%) were part of the ParlaMint I phase of the project. Most participants held the following three roles in the project: TEI encoding, preparation and submission of corpus samples, and preparation and submission of the entire corpus (all 17%). Most participants have a background in Natural Language Processing (NLP) (57%), followed by Computer Science (CS) (46%), Digital Humanities (DH), and Humanities (H) (31% each). In addition, three other participants chose "other" background, with the explanation that they come from the fields of Linguistics, Machine Learning, and Physics. It should be noted that the question allowed multiple responses and that there are few participants who have only one particular background. The distribution of participants' research backgrounds is shown in

[16] ParlaMint II also includes corpora for regional (autonomous community) parliaments, namely for Catalonia (ES-CT), Galicia (ES-GA), and Basque Country (ES-PV).

Figure 1 – the most common combination among participants with more than one background was a combination of NLP and Computer Science (9 participants), other combinations (such as NLP and Digital Humanities, NLP and Humanities, or Digital Humanities and Humanities) were less common. In four cases, participants chose a combination of three backgrounds; two cases for a combination of NLP, Computer Science and Digital Humanities; the other two cases for a combination of NLP, Digital Humanities and Humanities.

4.1 TEI

The questionnaire included questions about TEI encoding, participants' familiarity with TEI before starting the project, their experience with TEI during the project, and their plans to use TEI in the future. In general, 44% of participants were already at least somewhat familiar with the TEI P5 guidelines used in the encoding procedure, followed by 37% of participants who were not familiar with TEI at all. Of these participants, more than half (64%) did not require external sources to become familiar with TEI before they began working on ParlaMint projects (and only followed the guidelines "The structure and encoding of ParlaMint corpora"). In addition, 20% of the participants were very familiar with TEI.

Regarding the experience with TEI prior to ParlaMint, almost half of the participants were already using TEI encoding for their work (48%), while 22% were only using TEI-inspired encoding and 30% had never used TEI. Concerning the experience with TEI encoding in the ParlaMint project, we asked participants to rate various statements about TEI participation in the project and to assess the extent to which they agreed or disagreed with them (rated on a scale of "strongly disagree" (1) to "strongly agree" (5)). Sentiment towards TEI encoding ranged from neutral to mostly positive. Figure 2 shows the results and averaged values of the responses to each individual statement.

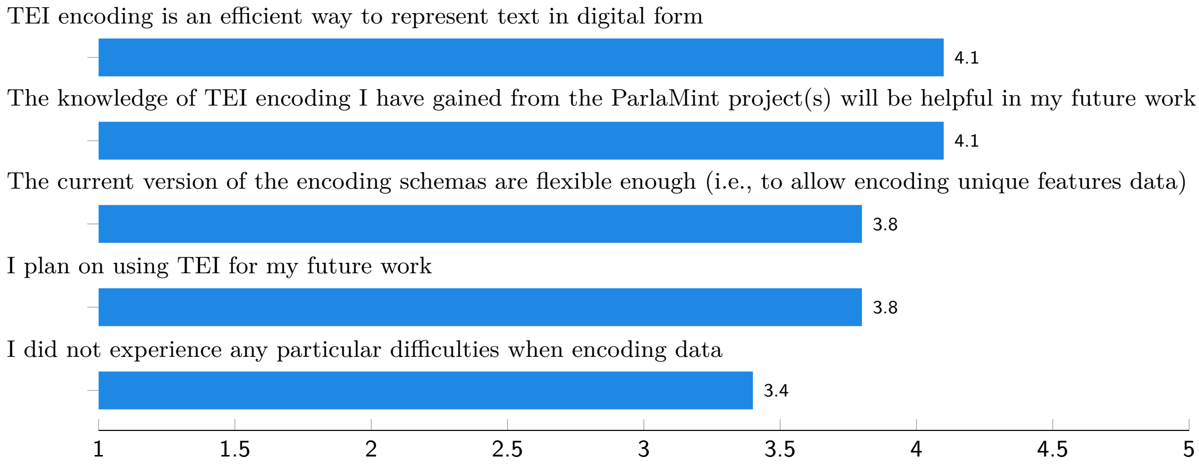

Figure 2: Averaged values of the responses to statements related to the partners' experience with TEI. The values ranged on the scale of "strongly disagree (1)" to "strongly agree" (5), with (3) indicating a "neutral" position.

The highest rated statement was *"TEI encoding is an efficient way to represent text in digital form"* and *"The knowledge about TEI encoding that I gained from the ParlaMint project will be helpful in my future work"* (both 4.1), followed by *"The current version of the encoding schemes is flexible enough"* and *"I plan to use TEI for my future work"* (both 3.8). The most neutral and lowest rated statement is *"I did not experience any particular difficulties when encoding data"*. This was further explored in the next question, in which we asked participants to provide additional feedback on their experiences with TEI encoding, with some expressing the complexity of TEI encoding in terms of the uniform nature of encoding very different parliamentary systems and languages. For further analysis, we examined the proportion of familiarity with TEI based on participants' research backgrounds, as shown in Table 1.

		Research backgrounds					Total
Response		NLP	CS	DH	H	Other	(Distinct count)
Very familiar	abs. value	3	1	2	3	1	5
	% of col.	15.00	6.25	18.18	27.27	33.33	14.29
	% of row	60.00	20.00	40.00	60.00	20.00	100.00
Somewhat familiar	abs. value	8	7	5	1	1	12
	% of col.	40.00	43.75	45.45	9.09	33.33	34.29
	% of row	66.67	58.33	41.67	8.33	8.33	100.00
Not at all familiar	abs. value	5	4	2	3	1	10
	% of col.	25.00	25.00	18.18	27.27	33.33	28.57
	% of row	50.00	40.00	20.00	30.00	10.00	100.00
No answer	abs. value	4	4	2	4	0	8
	% of col.	20.00	25.00	18.18	36.36	0.00	22.86
	% of row	50.00	50.00	25.00	50.00	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00
				61			

Table 1: Cross-tabulation of the statement "How familiar were you with TEI prior to ParlaMint?" with participants' research background (35 participants, 61 total responses). Each cell contains the absolute value of participants from each research background and their response, the percentage that the value represents relative to all participants from a given domain, and the percentage that the value represents relative to all participants from each response.

Regarding familiarity with TEI, slightly less than half of the NLP participants were somewhat familiar with TEI, followed by participants who were not familiar with TEI at all, and finally, those that were very familiar with TEI. The distribution trend is similar for both Computer Science and Digital Humanities participants but changes for Humanities where only one participant was somewhat familiar with TEI and the others were either not or very familiar with TEI. Lastly, participants with "other" backgrounds (Linguistics, Physics, and Machine Learning) were evenly distributed (1 participant not at all familiar, 1 participant somewhat familiar, and 1 participant very familiar).

We also checked the responses to the question "The acquired knowledge about TEI will be useful for my future work" in relation to the level of familiarity prior to the ParlaMint project work. Participants who were not at all familiar with TEI at the start of the project were relatively evenly distributed on a range from "strongly agree" to "neutral". A neutral position was also slightly stronger according to participants who were somewhat familiar with TEI, while the remaining participants voiced agreement (somewhat agree and strongly agree). Finally, more than half of the participants who were very familiar with TEI strongly agreed with the statement.

When asked if they planned to use TEI in their future work (Table 2), NLP participants expressed either strong agreement or a neutral position, and only one expressed mild disagreement. About 30% of Computer Science participants expressed a neutral position, and the others agreed or strongly agreed with the statement. Digital Humanities participants were somewhat mixed, as agreement ranged from "strongly agree" to "neutral," with a few more choosing "strongly agree." A similar proportion (and percentages) of agreement that TEI should be useful for their future work was also noted by participants from the Humanities. Finally, participants from "Other" backgrounds were evenly split between "strongly agree," "somewhat agree," and "strongly disagree."

4.2 Git

The questionnaire also included questions about the participants' familiarity with version control systems (VCS) prior to the start of the ParlaMint project, questions about communication and

		Research backgrounds					Total
Response		NLP	CS	DH	H	Other	(Distinct count)
	abs. value	7	3	4	4	1	10
Strongly agree	% of col.	35.00	18.75	36.36	36.36	33.33	28.57
	% of row	70.00	30.00	40.00	40.00	10.00	100.00
	abs. value	2	3	2	1	1	5
Somewhat agree	% of col.	10.00	18.75	18.18	9.09	33.33	14.29
	% of row	40.00	60.00	40.00	20.00	20.00	100.00
	abs. value	6	5	2	2	0	10
Neutral	% of col.	30.00	31.25	18.18	18.18	0.00	28.57
	% of row	60.00	50.00	20.00	20.00	0.00	100.00
	abs. value	1	1	1	0	0	1
Somewhat disagree	% of col.	5.00	6.25	9.09	0.00	0.00	2.86
	% of row	100.00	100.00	100.00	0.00	0.00	100.00
	abs. value	0	0	0	0	1	1
Strongly disagree	% of col.	0.00	0.00	0.00	0.00	33.33	2.86
	% of row	0.00	0.00	0.00	0.00	100.00	100.00
	abs. value	4	4	2	4	0	8
No answer	% of col.	20.00	25.00	18.18	36.36	0.00	22.86
	% of row	50.00	50.00	25.00	50.00	0.00	100.00
	abs. value	20	16	11	11	3	35
Total (responses)	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00
		61					

Table 2: Cross-tabulation of the responses for the statement "I plan to use TEI in future work" on the scale of "Strongly agree" to "Strongly disagree" and participants' research backgrounds (35 participants, 61 total responses).

workflow, and a general assessment of their experience with Git in the ParlaMint project.A large proportion of participants had previous experience with VCS (28 participants, 80%), whereas 7 participants (20%) had no experience with VCS.

Of the participants who reported previous experience with VCS, when asked which VCS they had used (multiple responses, 27 participants, 56 responses counted), all but one (1 participant did not answer the question) reported experience with Git (GitHub, GitLab: 27, 100%), followed by SVN and Bitbucket (10, 37% each), CVS (6, 22%), and Mercurial (3, 27%). None of the participants reported previous experience with HelixCore or Beanstalk.

Regarding experience with Git prior to starting work on the ParlaMint project, of the participants who indicated previous experience with Git and GitHub in the previous question (27 participants in total), 13 participants (48%) described their experience level as "intermediate", 9 (33%) as "beginner", and 5 participants (19%) as "advanced", which compared to all survey participants (35 in total, 8 or 23% did not provide an answer) equates to 37% "intermediate", 26% "beginner," and 14% "advanced".

Another important aspect of the survey was the communication process, in which the use of GitHub Issues played an important role – GitHub Issues served as the main communication channel between the corpus compilation work-package leads and the project partners, and for communicating issues that partners encountered. They also served to provide information/answers to other project partners facing similar issues. Regarding the communication process with GitHub Issues, participants were mostly positive. Almost 70% used Issues as a means to communicate and discuss problems, and of those, almost all (96%) used Issues to find relevant information pertaining to their particular problem, and more than ¾ participated in discussions about problems through Issue comments.

In addition to the use of GitHub Issues, we also asked about the experience with the project workflow. Most participants (77%) agreed that the Git requirements and workflow were clearly explained, while those who disagreed indicated that some constraints and phases of the workflow were not adequately explained (for example: "the whole process of approving the sample first was unclear to me" and "README was linked to Parla-CLARIN instead of ParlaMint, the upper limit of the sample size was not clearly stated, the paths to the Java libraries were hard-coded, some of the tests assumed Linux (as opposed to OS X), it was unclear that CoNNL-U and vertana were automatically generated"). The same percentage applied to whether they had particular difficulties submitting data samples to GitHub. The difficulties reported ranged from the size of the sample files to the distinction between the process of submitting data samples and the process of submitting full corpora. Over 80% indicated that they received sufficient and good support from the ParlaMint team. When asked if there were other GitHub features that facilitated communication and the work process, several participants mentioned the automatic validation that was started every time they pushed on an open pull request and the helpfulness of the Github Issues (as they provided information for other participants' issues) or the helpfulness of the people who provided responses to the issues. As a final part of the Git usability assessment, we asked participants to rate their experience with Git, GitHub features, and project workflows in the form of a single multi-sentence response table on a scale of "strongly disagree" (1) to "strongly agree" (5). Figure 3 shows the averages of responses to each statement.

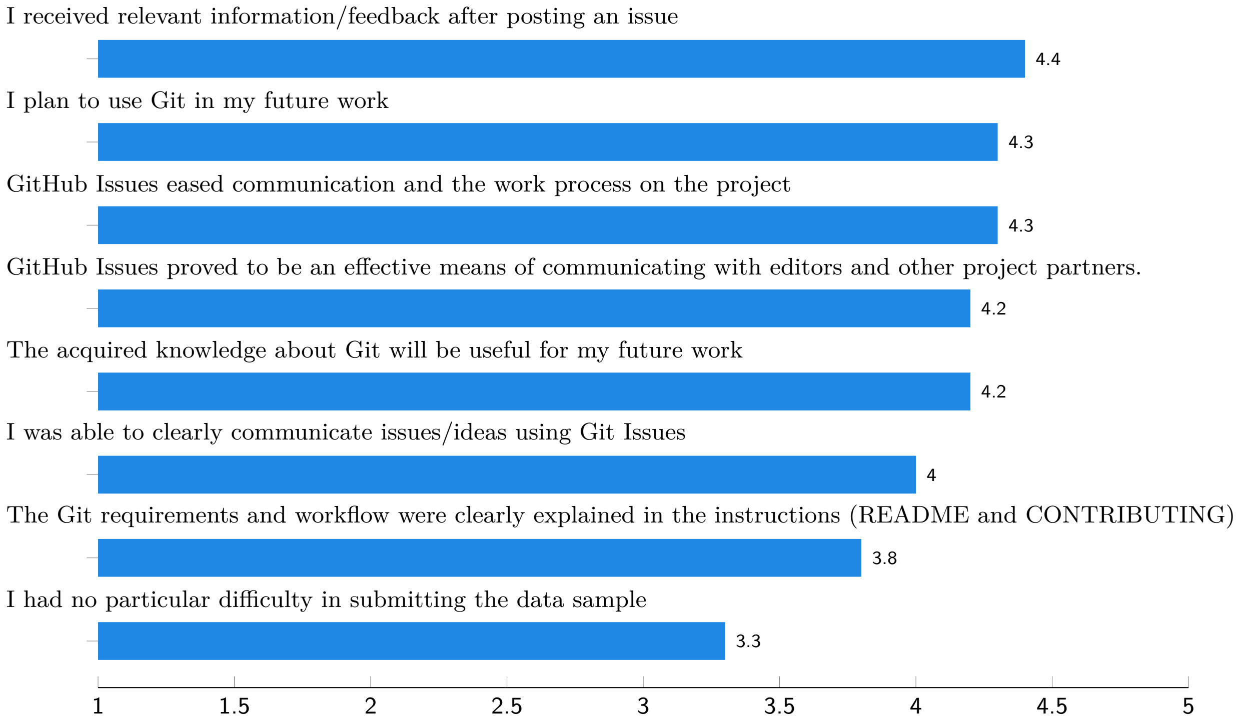

Figure 3: Averaged values of the responses to statements related to the partners' Git-related experience. The values are ranged on the scale of "strongly disagree (1)" to "strongly agree" (5), with value 3 indicating a "neutral" position.

The highest rated statement was *"I received relevant information/feedback after posting an issue"* (rated 4.4), followed by *"I plan to use Git in my future work"* and *"GitHub issues eased communication and the work process in the project"* (both 4.3). The responses to the statement *"I did not experience any particular difficulties in submitting the data sample"* were neutral (3.3). This was further illustrated by the next question, where we asked for additional comments on the workflow and submission process. There were some comments indicating that new changes to the material in GitHub (e.g., changes to the encoding process and corrections to the validation schemes) were not communicated clearly enough to participants, a video tutorial/workshop was suggested to facilitate the submission process, and that the GitHub Issues and responses from the team behind them were generally helpful.

We explored several relationships between participants' responses regarding their (previous) experience with Git and whether they plan to use it in their future work. First, we examined the relationships between participants' research backgrounds in terms of experience level or familiarity with Git, as shown in Table 3.

		Research backgrounds					Total
Response		NLP	CS	DH	H	Other	(Distinct count)
Advanced	abs. value	2	4	2	0	0	5
	% of col.	10.00	25.00	18.18	0.00	0.00	14.29
	% of row	40.00	80.00	40.00	0.00	0.00	100.00
Intermediate	abs. value	10	6	5	0	3	13
	% of col.	50.00	37.50	45.45	0.00	100.00	37.14
	% of row	76.92	46.15	38.46	0.00	23.08	100.00
Beginner	abs. value	7	5	3	4	0	9
	% of col.	35.00	31.25	27.27	36.36	0.00	25.71
	% of row	77.78	55.56	33.33	44.44	0.00	100.00
No answer	abs. value	1	1	1	7	0	8
	% of col.	5.00	6.25	9.09	63.64	0.00	22.86
	% of row	12.50	12.50	12.50	87.50	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00
				61			

Table 3: Cross-tabulation of the question "What was your experience with Git (prior to ParlaMint)" and participants' research backgrounds.

In NLP, half of the participants described their experience level as "intermediate", followed closely by beginners and finally two participants at the "advanced" level. In Computer Science, the distribution of experience levels was much more even (5 participants at the "beginner" level, 6 at the "intermediate" level, and 4 at the "advanced" level). Of the participants with a background in the Digital Humanities, slightly less than half described their experience level as "Intermediate", while in the Humanities, most participants were at the "beginner" level with a large number of "No answer" responses.

Following the above analysis, we also examined the relationship between participants' opinions about whether the knowledge they gained about Git might prove useful in their future work and their level of experience prior to participating in ParlaMint.

Half of the participants from the beginner group fully agreed that the Git knowledge they acquired would be useful for their future work, while the other half consisted of two participants who tended to agree with the statement and one who slightly disagreed. The sentiment changed in the intermediate group of participants, where more than half strongly agreed with the statement, while almost 30% slightly agreed with the statement and one somewhat disagreed. On the other hand, more than half of the advanced participants held a neutral position.

Finally, we examined the relationship between participants' responses about their intentions to use Git and their future work (agreement with the statement "I plan to use Git in my future work") and their research background. The responses are presented in Table 4.

Regarding Git, half of NLP participants expressed strong agreement, while the others only somewhat agreed with the statement or held a neutral position; the same sentiment prevailed among participants from Computer Science and Humanities backgrounds. More than half of all participants with a background in Digital Humanities strongly agreed with the statement, only one somewhat agreed, while participants in the "other" category all expressed strong agreement.

		Research backgrounds					Total
Response		NLP	CS	DH	H	Other	(Distinct count)
Strongly agree	abs. value	10	6	7	3	3	16
	% of col.	50.00	37.50	63.64	27.27	100.00	45.71
	% of row	62.50	37.50	43.75	18.75	18.75	100.00
Somewhat agree	abs. value	3	3	1	1	0	5
	% of col.	15.00	18.75	9.09	9.09	0.00	14.29
	% of row	60.00	60.00	20.00	20.00	0.00	100.00
Neutral	abs. value	2	3	0	1	0	4
	% of col.	10.00	18.75	0.00	9.09	0.00	11.43
	% of row	50.00	75.00	0.00	25.00	0.00	100.00
Strongly disagree	abs. value	1	0	0	1	0	1
	% of col.	5.00	0.00	0.00	9.09	0.00	2.86
	% of row	100.00	0.00	0.00	100.00	0.00	100.00
No answer	abs. value	4	4	3	5	0	9
	% of col.	20.00	25.00	27.27	45.45	0.00	25.71
	% of row	44.44	44.44	33.33	55.56	0.00	100.00
Total (responses)	abs. value	20	16	11	11	3	35
	% of col.	100.00	100.00	100.00	100.00	100.00	100.00
	% of row	57.14	45.71	31.43	31.43	8.57	100.00
		61					

Table 4: Cross-tabulation of the statement “I plan to use Git in my future work” and participants' research backgrounds.

5 Conclusions

The paper attempted to show how TEI can be used to specify the encoding of complex language corpora (or other types of language resources), providing both the guidelines of those wishing to encode the corpora, as well as XML schemas that are used to formally validate their encoding. We also presented Git which is well suited for controlled and distributed development and publishing of not only the guidelines and schemas but also the language resources themselves. As mentioned, the size of the produced ParlaMint corpora makes it somewhat problematic to store them in their entirety on GitHub but this is not the case for smaller language resources, particularly manually annotated ones.

Fully mastering Git is also not a simple process, especially for the typical researcher of the target Social Sciences (SSH) community. However, with an appropriate set-up, such as we have attempted to provide for ParlaMint, we believe that with only basic knowledge, partners can successfully submit their data sample with an automatic check if they validate, while the validation of the complete corpus still relies on local processing.

The paper also presented a survey on the use of TEI and Git by the ParlaMint project partners. Overall, participants were mostly positive about their experience with TEI and Git, although some difficulties were reported, mainly related to the distinction between the submission process for the sample corpus and the full corpus, as well as some workflow limitations. The difficulties identified will serve as an opportunity to update and optimise the current project workflow. With respect to TEI, initial responses were mixed. There was agreement that TEI is an efficient way to represent text in digital form and that the lessons learned will help participants in their future work. Contrary, there was less agreement, or even neutrality, on the question of whether the current schema is flexible enough to support the encoding of unique features of the data, given that it is still a very uniform and sometimes complex way of encoding sometimes drastically different parliamentary systems. On the other hand, reactions to Git were very positive, from relevant information and feedback received via GitHub Issues, to effectiveness in the communication process, to plans to use Git in the future. There was less agreement on whether the requirements and workflow were adequately explained, which points to the difficulties in data submission mentioned above.

The survey helped us to gain some insights into the relationships between participants' research backgrounds, their level of experience and opinion of their TEI and Git experience in the ParlaMint project, and participants' opinions on whether they plan on integrating them into their future work. We recognise that given the small number of participants (though still reasonably representative of the ParlaMint project group, as there are not many project participants), firm or decisive conclusions can not be drawn. Nonetheless, the results of the cross-tabulations can give us some insight into the role that research background and project work play in the likelihood that project partners would include Git and TEI in their "digital toolbox."

The survey showed that participants with Digital Humanities, NLP, and especially Humanities backgrounds tended to be more familiar with TEI. Among participants who had less or no prior familiarity with TEI, there was high agreement that the knowledge they gained would be helpful in their future work.

In comparison, as expected, the responses on the use of Git showed that participants from the NLP and Computer Science fields already relied on and used it regularly. In contrast, participants from the Digital Humanities and especially the Humanities were either unfamiliar or not very familiar with Git at the beginning of the project work. Regarding prior familiarity with Git and agreement on whether the knowledge gained could help them in their future work, most participants agreed, with the exception of those who were already very familiar with Git, most of whom were neutral about it, likely due to the fact that the project did not provide them with any new experiences they had not had before. Finally, most participants agreed that they would like to use Git for their work in the future, and this sentiment was particularly evident in the areas of Digital Humanities and NLP.

In our future work, we plan to continue working on the ParlaMint ODD, which will be updated as we move on to new types of annotation, in particular semantic annotation, and new types of resources, in particular machine-translated corpora and speech data.

The development of the ParlaMint corpora is also currently still rather centralised. In the longer perspective, we would like to encourage anyone that would wish to produce a ParlaMint-compatible corpus to be able to do so independently, for which we have to make the set-up (even) more flexible and also provide a tutorial on how to independently produce a ParlaMint compatible corpus. It is important to note that the corpus generated and validated with the automatic validation will be ParlaMint compatible only at the schema level, as, at least so far, we have also done manual verification in order to ensure the quality of the corpus data.

We believe that both TEI and especially Git, or, in principle, some other Version Control System – and the possibilities of combining the two – are not as well known in the SSH community as they should be, and that learning about them and adopting them into the work process could go a long way in making the development of encoding guidelines and language resources a much smoother and more controlled process, also leading to better reproducibility, a point that is very relevant to the goals of the CLARIN infrastructure.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions, and all the partners in ParlaMint I and II for their suggestions for improving the ParlaMint schema and tools and for taking part in the survey. The work presented in this paper was supported by the CLARIN ERIC projects ParlaFormat (2019) and "ParlaMint: Towards Comparable Parliamentary Corpora" (2020–2021 and 2022–2023), by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ, and individual project partner grants.

References

Lou Burnard, Christof Schöch, and Carolin Odebrecht. 2021. In Search of Comity: TEI for Distant Reading. *Journal of the Text Encoding Initiative*, (14).

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Tomaž Erjavec and Andrej Pančur. 2021. The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.4133.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

Christof Schöch, Roxana Patraș, Diana Santos, and Tomaž Erjavec. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1). http://doi.org/10.3828/mlo.v0i0.364.

Toma Tasovac, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado, and Andreas Witt. 2018. TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1. Technical report, DARIAH Working Group on Lexical Resources.

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.