Translating legislative documents at the European Parliament: e-Parliament, XML, SPA and the Cat4Trad workflow

Pascale Chartier-Brun

May 28th, 2018

urn:nbn:de:hbz:38-82669
www.zerl.uni-koeln.de
The translation services at the European Parliament, under the Directorate-General for Translation, use a chain of applications – e-Parliament – and XML technologies for the automation of the translation of legislative texts and efficient reuse of translated segments. This paper presents a detailed look at the translation workflow implemented at the European Parliament with focus on Safe Protocol Automation (SPA) and the Cat4Trad translation environment.


Les services de traduction du Parlement européen, sous la Direction générale de la traduction, utilisent une chaîne d'applications – e-Parlament – et les technologies XML pour l'automatisation de la traduction des textes législatifs et la réutilisation efficace des segments traduits. Cet article présente un regard détaillé sur le processus de traduction au Parlement européen, en mettant l'accent sur Safe Protocol Automation (SPA) et l'environnement de traduction Cat4Trad.
# Table of contents

1. Introduction .................................................................................................................. 1
2. Translation and reuse of text segments at the EP ...................................................... 1
   2.1. The e-Parliament applications chain and XML ..................................................... 2
   2.2. The translation workflow with Cat4Trad ............................................................... 4
   2.3. Matching translation memories in the FullCat Indices ......................................... 9
   2.4. Concordance searches in Cat4Trad ....................................................................... 11
3. Conclusion and Outlook ............................................................................................... 13
4. References .................................................................................................................... 14

This paper is based on the presentation “IT integrated environment for optimising the translation of legislative documents in the EP” given by Pascale Chartier-Brun at the workshop “Europäische Rechtslinguistik und Digitale Möglichkeiten / EU Legal Linguistics and Digital Perspectives”, held at the University of Cologne July 7th/8th, 2017. She is the Head of Unit, Directorate for Support and Technological Services for Translation, at the Directorate-General for Translation at the European Parliament.
1. Introduction

All 24 official languages of the European Union (EU) enjoy equal status, obligating the institutions of the EU to ensure the highest degree of multilingualism. Thus, the European Parliament (EP) publishes all parliamentary documents in all official languages of the EU. Each language can be translated into 23 other languages, resulting in 552 possible combinations. “In order to meet this challenge, the European Parliament has set up highly efficient interpreting, translation and legal text verification services.”

Further, the translators have access to a wide range of tools and technologies, “speeding up the translation process, reducing the risk of human error and improving consistency through the use of translation memories and reference to documentary and terminological databases.”

The translation services at the EP, provided by the Directorate-General for Translation (DG TRAD), use a chain of applications called e-Parliament as well as XML technologies for the automation of translation of legislative texts and efficient reuse of translated segments. The e-Parliament applications chain and the current translation workflow within Cat4Trad, including text segment matching and concordance search techniques, will be presented in the following pages.

2. Translation and reuse of text segments at the EP

The translation of legal texts in the European Parliament benefits largely from the possibilities of reuse from relevant translation memories. For example, an original text proposed by the European Commission (EC) and to be amended by the EP will already have been translated at the EC. Therefore, it is important to reuse translated texts and text segments in the translation process, not only for efficiency but also for coherency purposes.

---

The legislative procedure in the EP generates documents that are evolutive by nature (draft reports become reports, draft opinions become opinions), and whose content can be reused in other steps.

The identification of relevant texts, i.e., translation memories, for the translation at hand is based on a set of complex predefined rules. Currently the rules to select reference documents are automated by a system called SPA (Safe protocol automation), see section 2.2.

### 2.1. The e-Parliament applications chain and XML

The e-Parliament program integrates various applications needed for the processing of legislative documents in the European Parliament. This integration is possible since all documents are created in XML (Extensible Markup Language). The use of XML brings many advantages, namely

- easy creation and verification of documents,
- fewer errors when producing documents in different applications,
- better reuse of texts for translation and publication,
- better quality of published texts,
- and reduction of costs over the life cycle of documents.

---

4 All illustrations and screenshots are used with permission, courtesy of EP (2014-2017).
The chain of tools brought together for the processing of legislative documents in e-Parliament is shown here in a simplified form:

**Figure 2: Simplified technical architecture of EP translation services applications**

To illustrate the current workflow of the translation process at EP DG TRAD, four tools of the application chain, namely AT4AM, DM-XML, PURE-XML and Cat4Trad, will be briefly presented.

**Figure 3: Four applications in the e-Parliament chain**
• **AT4AM** is an authoring tool that the members of the EP use to create amendments in XML format.

• **DM-XML** is a document modelling tool. It generates the correct model of the legislative document being processed. So, for example, when a document is opened for translation in Cat4Trad, the standard text in the document will already have been automatically prefilled by DM-XML.

• **PURE-XML** stands for Parliament Unique Repository for XML content. It is a storage repository where all XML content is stored by the applications that create it. The next application in the chain then retrieves the content it needs from that same repository.

• **Cat4Trad** stands for Computer Assisted Translation tool for DG TRAD. It is an application that is currently used for the translation of legislative documents and committee agendas. It was specifically designed for the translators in the language units, with the main aim of
  - automating most of the pre- and post-processing tasks,
  - facilitating the tasks of the EP translation services
  - retrieving material from the other e-Parliament tools,
  - and sending back XML material to the chain for future re-use.

### 2.2. The translation workflow with Cat4Trad

In this section, the translation workflow with Cat4Trad will be described. This tool utilizes the four indices that comprise the repository FullCat, namely the normative, the reference, the retrieval and the shared index.

![Figure 4: FullCat Indices](image-url)
There are different procedures that apply before, during, and after translation. To exemplify these procedures, the processing of a new translation will be illustrated in the following. In this example, a new translation, called FdR (Feuille de Route),\(^5\) has been requested. A series of automatic and manual actions are launched.

<8>

One automatic procedure is performed by SPA, standing for \textit{Safe Protocol Automation}. It is a tool for the automatic pretreatment of documents following the Safe Working Protocols. Depending on the document type of the translation requested, SPA starts the automatic pretreatment of the document. For instance, it retrieves all the relevant and useful TMX (Translation Memory eXchange)\(^6\) files related to the FdR from Euramis.

<9>

Euramis stands for \textit{European advanced multilingual information system} and “acts as a multilingual, multidirectional repository of clearly labelled equivalent phrases ("segments") belonging to official EU documents allowing their re-use in translation in all European Institutions.”\(^8\) New documents that are to be translated will be compared with the contents of existing translation memories and relevant source/target segment pairs including metadata will be retrieved. Entire documents can also be searched and downloaded with their relevant metadata.\(^9\)

\(^5\) An FdR is an “electronic form sent to the Planning unit used to request translations, revisions, editing, terminology, etc.”, see \url{http://iate.europa.eu/FindTermsByLilId.do?lilId=294480&langId=en}, IATE ID: 294480, accessed on 15.03.2018.

\(^6\) TMX stands for Translation Memory eXchange and is an XML-compliant format, see \url{https://www.gala-global.org/tmx-14b}, accessed 19.03.2018.


\(^9\) “Euramis Concordance provides the option to query the available translation memories based by specifying a search string combined with the translator’s login and displaying the result on screen.” See \url{http://ec.europa.eu/dpo-register/details.htm?id=41727}, accessed on 20.03.2018.
Figure 5: FdR import into Cat4Trad, automatic preparation

FullCat indexes the TMX files obtained by SPA for all the previous documents belonging to the same procedure, including the basic reference documents. These are then imported into the reference index. The retrieval TMX file that SPA received from Euramis is imported into the retrieval index. This procedure insures that all the segments needed for pre-translation and during translation are found in the indices.

In a separate procedure, the new FdR\(^{10}\) is imported into Cat4Trad, where it is automatically prepared for translation. For example, for a document containing two document column amendments, Cat4Trad automatically fills the left-hand column based on what is available in the PURE-XML repository. With the help of DM-XML, the translations of the standard texts, such as amendment headings and titles, are prefilled throughout the document.

---

\(^{10}\) The FdR in figure 5 is an amendment (AM).
The text parts that have not been prefilled by DM-XML are then pre-translated at the rate of 100% against the normative index of FullCat.

Finally, Cat4Trad pre-translates the right-hand column against the basic reference document available in the reference index of FullCat. This insures that only segments from the basic reference document are inserted.

As soon as the pretreatment processes in SPA and Cat4Trad are finished, the translator can begin translating it. Maybe the document was assigned to more than one translator, in which case all translators can work on different parts of the same document at the same time in
Cat4Trad. While translating, they all have access to segments in the reference, retrieval and shared indices.

Figure 8: Access to indices during translation

The new segments created are stored immediately in the shared index, which means that all translators have access to each other’s work in real time.

Figure 9: During translation: storage in and access to shared index
After the translation, revision and validation of the whole document are complete, Cat4Trad intervenes to finish the process.

Cat4Trad saves the XML version of the document to the PURE-XML repository. A TMX file with the newly translated content is created and sent for upload to Euramis. Finally, Cat4Trad creates a DocEP\(^{11}\) compliant document file and sends it to be saved on the primary drive of the translation unit. After the document has been booked out of the translation unit, the workflow is complete.

### 2.3. Matching translation memories in the FullCat Indices

This section illustrates how matches are selected in Cat4Trad from the FullCat indices during the translation process. In this example, Amendment 4 is to be translated from English into French. To do this, the translator will hover the mouse over the respective text. Clicking anywhere in the highlighted text will open the segment in the inline editor. Opening the segment prompts Cat4Trad to search for matches in the FullCat indices. A match is proposed if at least 65% of the source segment that is saved in the indices is similar to the source segment that needs to be translated. If at least one match is found, in either the reference, retrieval or shared index, it will be displayed together with its translation.

Figure 11: Primary match in Cat4Trad

This match is called the primary match. If more than one match is found, the one with the highest match rate is displayed. The differences between the source text of the match found and the source text to be translated are marked with track changes in the source text field of the inline editor. Other available matches from the FullCat indices can also be viewed. Clicking the ‘Matches’ button will result in the display of a list of the best available matches. At the top of the list, the primary match is shown.

< 14 >

The matches are selected as follows. First, a maximum of five best matches are shown in this list. Each match displayed comes either from the reference, retrieval or shared indices, and has a minimum match rate of 65%. The match with the highest match rate is displayed first, down to the one with the lowest match rate. If there are matches with an identical match rate in the list, the one from the basic reference document, if any, is shown before the others. If there are no matches from the basic reference document, the matches are sorted by date of creation, from the most recent to the oldest.
Regarding the metadata for each match, one can see that the matches with the highest match rates are displayed first. Therefore, in this example, the three matches with an 84% match rate are displayed before the two matches with the 79% match rate. Secondly, for matches with the same match rate, Cat4Trad gives priority to the basic reference document, so this match comes first. The next two matches, which also have an 84% match rate, both come from the retrieval index. So Cat4Trad displays the most recent one first. The last two matches have an identical match rate of 79% and were both created on the same date at the same time. In this case, Cat4Trad gives priority to the one from the reference index over the one from the retrieval index.

2.4. Concordance searches in Cat4Trad

Concordance searches can be performed on a term, a sentence, a reference number, a title or any other string of characters in Cat4Trad. This can be useful if no matches were found for a certain segment, but the translator would still like to see whether parts of the segment have been translated before. Different dictionary tools can be used for the concordance search, namely FullCat, FullDoc, Quest and DocFinder.
FullCat is the working server directly supporting Cat4Trad (see figure 4 and 5). FullDoc includes all the documents of the EP and legislative documents from the Commission and the Council. An interinstitutional search for terminology or documents can be performed with the centralized database Quest. Documents with a reference number can be found in the centralized interinstitutional database DocFinder.

To start the concordance search, the translator highlights the selected segment and selects the
desired dictionary tool. Up to five results will be displayed in the matches box in the inline editor. The words for which the translator searched are marked in bold in the source text. Useless results can be removed and useful translations copied to the target field.

 DocFinder is used for finding documents according to their reference number. For example, the translator may want to find the title of a certain regulation.

![DocFinder Interface](image)

**Figure 15: Interinstitutional document search with DocFinder.**

Highlighting the regulation number and clicking on DocFinder initiates an interinstitutional concordance search for the corresponding document. Thus, the proper title of the regulation can be easily found in the target language and copied to the editor within Cat4Trad.

### 3. Conclusion and Outlook

This paper has given an insight into the current translation workflow at the European Parliament. The efficient and productive data exchange between the different applications of the EP translation department is possible due to the use of XML. Further data exchange with interinstitutional IT-structures, e.g., access to the translation memories in Euramis, is also supported by the use of XML-compliant formats.

Currently the SPA system is being modernised with a flexible workflow engine, which will allow to easily change the workflow of the documents and to include additional steps such as quality control of the original text, a different pretreatment or any other new feature.

As shown above, the identification of relevant translation memories for translations is based on a set of predefined rules. These rules are complex and difficult to maintain, and can
probably benefit from the use of Machine Translation (MT) techniques. Already successfully implemented as a proof-of-concept, Machine Translation based on neural networks, i.e., Neural Machine Translation (NMT), will soon be in operation to further improve the selection of translation segments proposed to the translators at the EP (see forthcoming article in ZERL).

4. References


“DocEP for Freelance Translators”.


“EP Translators”.


“Multilingualism in the European Parliament”.