Journal of the Text Encoding Initiative

Issue 9 | September 2016 - December 2017
Selected Papers from the 2014 TEI Conference

Materiality of TEI Encoding and Decoding: An Analysis of the Western European Union Archives on Armament Policy

Florentina Armaselu, Verónica Martins and Catherine Emma Jones

Electronic version
URL: http://journals.openedition.org/jtei/1463
DOI: 10.4000/jtei.1463
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Florentina Armaselu, Verónica Martins, and Catherine Emma Jones, « Materiality of TEI Encoding and Decoding: An Analysis of the Western European Union Archives on Armament Policy », *Journal of the Text Encoding Initiative* [Online], Issue 9 | September 2016 - December 2017, Online since 19 September 2016, connection on 12 January 2018. URL : http://journals.openedition.org/jtei/1463 ; DOI : 10.4000/jtei.1463

Materiality of TEI Encoding and Decoding: An Analysis of the Western European Union Archives on Armament Policy

Florentina Armaselu, Verónica Martins, and Catherine Emma Jones

ERRATA

This paper was revised on 2017-08-31 to include a missing reference: "(Heiden 2010)." The previous version is archived at https://journals.openedition.org/jtei/1733.

1. On the Materiality of TEI Encoding and Decoding

1 As Manoff (2006, 311) points out, the study of "material aspects of digital objects" may foster new concepts and theories explaining how these properties "alter our ways of creating and consuming information." Through the lens of the history of the book, materiality may be related to things like "typography, binding, illustration, and paper." In the world of digital artifacts, this can bring to light a "whole new range of physical objects and processes, including platforms, interfaces, standards, and coding" (312). From an editorial perspective, McGann (2001, 78) argues that a particular edition may draw attention to the "dynamic engagement between text and its vehicular

(material) form" and in this respect, digital textuality differs from paper-based text as it can be designed for "complex, interactive transformations" (81). Referring to the transition from print, Hayles (2005, 1) assumes that the "transformation of a printed document into an electronic text is a form of translation—'media translation'" that implies an "act of interpretation," an edition representing an instantiation of a work in a "physical form." In particular, relating the process of text encoding to hermeneutics, "the art or science of interpretation," Burnard (1998, section 3) suggests that the "markup maps a (human) interpretation of the text into a set of codes on which computer processing can be performed." Moreover, he sees this process as part of a decoding–encoding succession, the former implying "a selection from the many features implicit in the reading of a text, and their re-encoding in explicit and unambiguous terms." (section 2)

2 This article proposes a new perspective as a new link in the chain, decoding-encoding-decoding, by considering the aspects involved in the digitization, TEI XML encoding, and corpus analysis of a collection of documents from the Western European Union (WEU) paper archives specifically intended for web publication. Placing our approach at the crossroad of digital scholarly editing and digital philology (Pierazzo 2014; Andrews 2013), as well as digital hermeneutics (Capurro 2010), we will focus on the TEI encoding as an addition to the original text, a "material" layer that further supports both machine and human interpretation. This type of materiality is actually related to the intrinsic performative quality of the code (either as markup or as part of a computer program), designed to be interpreted by the machine and to determine processing sequences and output that can become an object for further analysis. In this sense, we may move closer to Ihde's (2003) concept of "material hermeneutics" by understanding code, and digital technology in general, as an "instrument" we can use in "hermeneutic ways" to produce knowledge.

3 In this context, by combining scholarly editing and hermeneutic perspectives, the paper discusses aspects related to building a digital edition of historical documents, the interpretation of documents (*decoding*) on which the XML-TEI annotation is based (*encoding*), as well as the elements and methods subsequently considered for corpus analysis in order to support interpretation of discursive phenomena (*decoding*). Section 2 describes the collection that forms the focus of the study, with reference to the original paper archives and the transformation into electronic form. Section 3 presents a selection of encoded samples and the rationale behind the encoding, from the perspective of its further usage in the decoding process. In section 4, we detail this process placing

emphasis on the importing of the encoded data into the corpus analysis platform, the different types of analysis performed, and a discussion of results. Section 5 concludes the paper with a review of both the benefits and limitations of the applied methods and the materiality underlying the encoding–decoding mechanism.

2. Description of the Collection

4 The corpus (WEU-Diplo) chosen for TEI encoding represents a selection from the Archives nationales de Luxembourg[1] of institutional documents concerning armament production and standardization, and armament control within the WEU,[2] from 1954 to 1982. TEI XML has been considered an appropriate format both for building a scholarly online edition and for enabling corpus analysis. The general workflow was conceived with the aim of (partial) re-usability, albeit with some project-specific adaptations and readjustment, in order to support a variety of projects and document types in European integration history (primary/secondary sources: text, image, audio, video, and their transcriptions). As a matter of principle, an alternation of manual and automatic sequential processing has been applied to the corpus in such a way that, independently of the manual interventions, no information should be lost if it is needed to regenerate a certain state of the corpus in subsequent automatic phases (e.g., via XSLT).

5 The first criterion for the selection of the documents was their relevance to a specific research question: what were the French and British positions on the major defense and security matters within WEU and, as subsidiary questions, the identification of the main defense and security matters discussed within the Council of the WEU and the importance placed on the WEU by the two member states, France and the United Kingdom, in their diplomatic strategy between 1954 and 1982.[3] The choice of this case study was influenced by the fact that these two states played a key role in the birth, development, and organization of the WEU, as evidenced in the primary and secondary sources consulted. The idea was that by analyzing these issues we can also shed light on the organization's contributions to European defense under the leadership (or lack thereof) of France and the UK.

6 Other criteria influencing document selection were to obtain a balance among control, standardization, and production, together with a balanced number of documents across the decades, as well as taking into account that some documents are a logical follow-up of others

(e.g., documents mentioning recommendations). Another criterion for selection, of secondary importance, was the actual form of the documents, since the research was also intended to evaluate the accuracy of OCR (optical character recognition) processing for different types of layout (title or content page), paper quality, and legibility of typewritten or handwritten characters and of particular markings as stamps.

7 The selected sample contained 127 documents, 60 in English and 67 in French. For the first phase of the project, 55 documents in French were retained (a total of 290 pages) because of their importance for subsequent corpus analysis[4] and publication. The majority of the documents were notes from the Secretary-General or Secrétaire Général (46, of which 37 were encoded because of their relevance to the research question), followed by minutes of meetings of the WEU Council or the Standing Armaments Committee (SAC) or of the working party on production and standardization of armament of the Interim Commission (15, all of which have been encoded). The sample also included 2 memoranda (both encoded) and 2 studies (1 of which has been encoded).

8 The linguistic aspect was also considered. Although, given the availability of time and resources, the French version was prioritized for encoding (the English processing being planned for a later stage), in general, the chosen documents existed in both French and English versions, and when not mentioned as original language, the French documents were exact translations of the English ones or had the same status, since the documents were produced in both languages. The provider of the translations was the WEU itself (its daily work being undertaken in both languages) and, therefore, a source of official translations.

9 From the 55 French documents selected for encoding, 5 had no English equivalent (3 notes, 1 study, and 1 memorandum); the remainder (all the meeting minutes and remaining categories) were available in both languages. More precisely, there were 16 documents mentioning the original was in French (of which 1 had no English equivalent), 12 with the original indicated as being in English, 10 mentioning the original both in French and English, 11 with no indication of the original language but a comparison between the French and English documents showed a similar structure and content, 4 available only in French bearing no indication of the original language, and 2 with an ambiguous marker (1 indicating French and English as original in the French version

while the corresponding English document mentioned English as the original language, the other bearing no mention of the original in the French version but indicating English as the original in the corresponding English document).

2.1 The Paper Archive

10 As a general rule, documents from the Secretary-General all exist in both French and English. In the nearly 400 folders consulted, there were very few exceptions; only occasionally were documents published in just one language. The French version was printed on blue paper and the English on white paper. Internal documents or notes from the Agency for the Control of Armaments (ACA) were only published in French, on very thin ("tracing") paper. The notes and minutes were formal documents distributed to all the delegations of the Member States.

11 The research aim was to gather a set of representative documents that expressed the French and British positions within the WEU's different bodies, on different topics—the exploration of the design, production, and control of armament being only one of them. For this purpose, we used the WEU's collection database, held in the Archives nationales de Luxembourg, and consulted several sections and collections including: (1) Interim Period; (2) Brussels Treaty Organisation (BTO); (3) 1954–87 within the Secretariat-General/Council's archives; (4) Armament Bodies—Agency for the Control of Armaments (ACA) and Standing Armaments Committee (SAC). Other collections such as those relating to the military bodies or WEU operations were considered too recent to be consulted or still held a classified status. The documents were selected primarily based on their themes, strictly following the thirty-year rule. Each collection, comprised of several folders, with a "fiche" indicating the name of the section and collection, title of the folder, security classification such as WEU or NATO (NATO classified documents were not available), the period of time covered in the folder, the reference, and either keywords or a small summary of the main questions, although sometimes this was quite general. Once the folders or boxes were located in the WEU collection, they were consulted *in situ* and a selection of several documents was made according to the general theme *armament*. Before the digitization, a closer reading and final selection was performed.

2.2 Paper to Electronic Text

12 The initial documents were typewritten materials from the WEU archives. The transformation into electronic text necessitated the use of document scanning, OCR, manual post-processing error correction, conversion of the resulting styled Microsoft Word files to TEI XML P5 via OxGarage,[5] and further XSLT transformation and enrichment using the oXygen XML Editor and GATE (General Architecture for Text Engineering) for NER (Named Entity Recognition). To prepare the digital documents for publication on the Web, further processing was carried out, to facilitate visualization and navigation in the browser at the document and collection level.

13 Next we identified, for each document, principal metadata and semantic elements necessary for encoding, as well as the form of encoding required for the computational linguistic analysis in order to answer the main research question. The metadata included elements such as (1) the author of the document: for the majority of documents the author is a collective entity (institution), except for rare cases—internal documents—where the author is an individual; (2) the date on which the document was distributed or produced, such as the date of the meeting; (3) the location (generally London, where the Council's Secretariat-General was located, or Paris, home of the ACA and SAC headquarters); (4) the title; (5) the version (whether or not it is a final version); (6) the language of the document with the mention of the original, when present; (7) the classification: most of the documents were classified as *confidential*, *secret*, or even *top secret*, according to the degree of sensitivity of the information. Likewise, documents of this type also had a copy number (element 8), since they often contained military and strategic information and were only distributed to a limited (sometimes very small) number of people, generally officials at national or institutional level. The organization had its own system of codes (references) for each document, which varied depending on the institution or the type of document.[6] The document reference (element 9) only partially identifies a document, since the same code was sometimes used for several documents, particularly for minutes of meetings which were divided into thematic sections and incorporated into different folders. The folder code (element 10) is therefore an important means of identifying the theme and even the institution within the WEU.[7]

14 In order to address the research question, the main aim of the TEI XML encoding was to identify the speakers (element 11) in the various documents and the views (element 12) that can be attributed to them, whether directly or indirectly.[8] The representatives (ministers, parliamentarians,

ambassadors, and experts) from France and the United Kingdom were systematically identified, and, depending on their relevance to the research, the contributions of the German representatives were also encoded. Some examples of British speakers' names are Christopher Steel, Samuel Hood, and Selwyn Lloyd; French speakers' names included Alexandre Parodi, Jean Chauvel, and Geoffroy Chodron de Courcel.

15 A generic nomenclature was developed to maintain consistency among the various speakers and to deal with cases when the speaker was not named: `repres_fr` (French representative), `repres_uk` (United Kingdom representative), `repres_frg` (German representative), `repres_frg_fr_it` (contribution on behalf of the German, French, and Italian representatives), `repres_deleg_fr` (French delegation), `repres_deleg_uk` (United Kingdom delegation),[9] `repres_cons_weu` (representative of the WEU Council), `repres_assb_weu` (representative of the WEU Assembly), `repres_aca` (representative of the Agency for the Control of Armaments), `respres_sac` (representative of the Standing Armaments Committee) (see example 1).

16 A semi-automatic NER processing involved the identification of other elements (13) in the texts, intended to be considered as whole units in the analysis, such as dates and the names of places, people, office positions, organizations, and bodies (*Western European Union*, the *North Atlantic Treaty Organisation*, the *Standing Armaments Committee*, and the *Agency for the Control of Armaments*), the event to which the document refers, as well as any other events mentioned.[10] Finally, a series of "products" associated with armament were also encoded, for example *Mirage*, *short-haul transport aircraft*, *light tank*, *Pluton*, etc.

3. Encoding

17 Although other types of elements were annotated in the corpus (metadata, e.g., title, author, availability date, origin place, and confidentiality status; and structure, e.g., headers, footers, sections, paragraphs, and line breaks), the paper will focus on the content-related encoding—that is, speakers and their discourse—along with the above-mentioned categories of named entities, considered from the perspective of the subsequent decoding phase (analysis and interpretation). The TEI P5 specifications[11] were applied, with no need for adding new classes, elements, or attributes.

3.1 Participants

18 The identification of the agents responsible for the production of texts represents an important step in the analysis of institutional discourse, irrespective of the level of this analysis (Thornborrow 2002; Phillips, Lawrence, and Hardy 2004; Nikander 2008; Van Dijk 1993). Certain types of documents in the WEU-Diplo corpus (usually minutes) explicitly provided indications about the speakers and their position ("*Etaient présents: République Fédérale d'Allemagne: Prof. Dr. L. ERHARD; Belgique: M. A. de STAERCKE …*").[12] For other cases, external knowledge or prior research was needed in order to be able to assign a role to the speaker ("*M. Selwyn LLOYD déclare que …*"[13] as a UK representative) or to identify the contributors to the discourse (the France representative, Geoffroy Chodron de Courcel, and the ACA representative).[14]

Example 1. Generic list of participants. WEU-Diplo: CR/58/8.

```
  <particDesc>
   <p>Liste des représentants des pays/organisations.<list xml:id="repres_list">
     <item xml:id="repres_fr">Représentant(s) de la France</item>
     <item xml:id="repres_uk">Représentant(s) du Royaume-Uni</item>
     <item xml:id="repres_frg">Représentant(s) de la République Fédérale
d'Allemagne</item>
     <item xml:id="repres_frg_fr_it">Représentant(s) de la République Fédérale
d'Allemagne, de la France et de l'Italie</item>
     <item xml:id="repres_fr_uk">Représentant(s) de la France et du Royaume-Uni</
item>
     <item xml:id="repres_deleg_fr">Représentant(s) de la délégation française</
item>
     <item xml:id="repres_deleg_uk">Représentant(s) de la délégation britannique</
item>
     <item xml:id="repres_deleg_fr_be">Représentant(s) des délégations française
et belge</item>
     <item xml:id="repres_deleg_fr_uk">Représentant(s) des délégations française
et britannique</item>
     <item xml:id="repres_cons_weu">Représentant(s) du Conseil de l'U.E.O. (Union
de l'Europe occidentale)</item>
     <item xml:id="repres_sac">Représentant(s) du C.P.A. (Comité permanent des
armements)</item>
     <item xml:id="repres_wpbt">Représentant(s) du G.T.P.B. (Groupe de travail sur
le Pacte de Bruxelles)</item>
     <item xml:id="repres_aca">Représentant(s) de l'A.C.A. (Agence pour le
contrôle des armements)</item>
     <item xml:id="repres_assb_weu">Représentant(s) de l'Assemblée de l'U.E.O.
(Union de l'Europe occidentale)</item></list></p>
  </particDesc>
```

19 The encoding of the participants (considered of interest for the research question) required two types of annotations:

- a mandatory, generic label identifying the participant as an institutional or country representative;

- an identifier (unique for the corpus) provided only when it was possible to refer to a particular person.

20 Since the conversion to TEI XML and the semantic enrichment of the corpus supposed both automatic transformation (via XSLT) and manual annotations, in a preliminary form for all of the documents, we generated a list of all generic labels for the country/institution representatives (<profileDesc> section of the <teiHeader>) (example 1).

21 Then the list was manually customized according to the particularity of each document and the specific "actors" involved in the production of the text. Example 2 illustrates a case where only three representatives were retained and further details were provided on their identity. For other situations, the generic label was enough (when the identity was not required or not available).

3.2 Discourse

22 In order to be able to analyze the discourse of different participants within the WEU's policy on armament issues, we have applied a "kaleidoscopic" approach to the corpus. More precisely, discrete fragments were identified and manually annotated inside each document, with reference to the speaker and his or her role as an institution or country representative manifested in the text (example 2).

Example 2. Customized list of participants. WEU-Diplo: CR/58/8.

```
  <particDesc>
   <p>Liste des représentants des pays/organisations.<list xml:id="repres_list">
     <item xml:id="repres_fr">Représentant(s) de la France <name type="person"
xml:id="faure">Maurice Faure</name></item>
     <item xml:id="repres_uk">Représentant(s) du Royaume-Uni <name type="person"
xml:id="lloyd">Selwyn Lloyd</name></item>
     <item xml:id="repres_frg">Représentant(s) de la République Fédérale
d'Allemagne <name type="person" xml:id="von_brentano">Heinrich von Brentano</
name></item>
    </list></p>
  </particDesc>
```

23 Given its flexibility of use (either inside a paragraph or encompassing several paragraphs), the <said> tag was chosen for delimiting the different pieces of discourse corresponding to a particular agent (example 3). The choice also facilitated assembling these pieces of information for analysis in the decoding phase (section 4).

Example 3. Excerpt from a discourse (oral). WEU-Diplo: CR/58/8.

```
  <p><name type="person">M. Faure</name> souligne avec force que <said
direct="false" ana="#oral_disc" who="#faure" corresp="#repres_fr">le succès
de l'entreprise<lb/>dépend de la volonté politique des gouvernements d'assurer
une<lb/>coopération effective. Les propositions de <name type="person">M. von
Brentano</name><w>cons<lb rendition="#hyphen_before" break="no"/>tituant</w> un
pas important dans cette direction et il s'y rallie.</said></p>
```

24 The @corresp and @who attributes were used in order to link the marked-up fragment with its producer defined in the <particDesc> unit. Additional attributes (@ana and @direct) were needed to differentiate situations referring either to transcribed *oral* (direct or indirect speech) or to what we considered *written* discourse, such as the text of notes (example 4) or studies usually resulting from internal meetings or discussions among institutional bodies (sometimes including "narrative" prose[15] or arguments not necessarily coming from an oral account), and then circulated for further discussion/approval within the WEU.

Example 4. Excerpt from a discourse (written). WEU-Diplo: Note[16] .

```
  <said ana="#written_disc" corresp="#repres_aca">
   <p>Un an après, dans des conditions analogues, le <name
type="person">Ministre<lb/>LUNS</name> était amené à prendre comme président une
position dans<lb/>le même sens.</p>[…] </said>
```

25 A particular occurrence of "discourse within a discourse" is presented below (example 5): a direct citation of the oral intervention of a WEU Council representative, M. Heath, from a previous meeting of the WEU Assembly, within the written account of the ACA.[17]

Example 5. Excerpt from a discourse within a discourse. WEU-Diplo: Note[18].

```
  <said ana="#written_disc" corresp="#repres_aca">
   […] <p>La réponse de <name type="person">M. HEATH</name> était : <said
direct="true" ana="#oral_disc" who="#heath" corresp="#repres_cons_weu">"Nous ne
sommes<lb/>pas juridiquement tenus d'autoriser l'inspection de ces dépôts,<lb/
>car ils ont été constitués dans le cadre de l'<name type="org">OTAN</name>, et
sont<lb/>donc, strictement parlant, uniquement soumis à l’inspection de<lb/>cette
organisation."</said>[…].</p>[…]</said>
```

3.3 Named Entities (NE)

26 The project also included the identification and annotation in the text of named entities intended for later use (such as indexing and linking to an authorities list) or as a prerequisite of the corpus analysis phase. This identification and annotation allows multiword expressions to be counted as single units of a given type (e.g., organization) in the analysis, rather than as separate words (for instance, *Union de l’Europe Occidentale* instead of *Union, de, l’, Europe, Occidentale*).

27 The NER task involved a semi-automatic approach using GATE (French NE, Gazetteer, and Gazetteer List Collector plugins)[19] for the detection of seven classes of entities: persons, places, organizations, events, dates, products, and functions (official positions). Manual corrections were applied, when necessary, in a post-processing phase. Since the GATE XML output format was different from TEI, an XSLT dedicated stylesheet was created for the transformation of the GATE tags (such as `<Person>`, `<Location>`, `<Organization>`, and `<Date>`) into corresponding TEI tags (`<name>` with the attribute `@type`, and `<date>`, respectively). A few examples of `<name type="person">`, `<name type="org">` are presented in the previous examples. Further transformation was necessary during the importing of the annotated corpus into the software for textual analysis (see section 4.1).

4. Decoding

28 The so called “decoding” phase, for corpus analysis and interpretation, consisted of importing and processing the TEI XML annotated documents within a specialized platform, TXM (Heiden 2010),[20] that allows the analysis of a large body of texts by means of lexicometrical and statistical methods. The previous encoding served as a basis for discerning or grouping together different types of semantic or structural elements needed for analysis.

4.1 Importing

29 Since TXM supports XSLT transformation at the moment of import (XML/w+CSV option), an XSLT stylesheet was created to accommodate particular formats or conversions required by the software. Therefore, it was not necessary to store different versions of the corpus, one for TXM analysis, the other for Web publication.

30 First, a lowercase conversion[21] was provided for consistency reasons relating to the varying ways of capitalizing (e.g., *Comité militaire de standardisation*, *Comité militaire de Standardisation*, *Comité Militaire de Standardisation*). Second, for the named entities to be interpreted as a whole instead of as separate units, a supplementary conversion was needed, all the <name> tags being converted to <w> tags each denoting a “word” of a given type (e.g., person or organization). The special case of hyphenated words where the hyphen appears at a line break (see example 3) had to be considered in an earlier transformation, before import, so that the whole word and not its parts could be counted in the analysis (in the example, *constituant* instead of *cons* and *tituant* as the software would treat it without a <w> tag).

31 Part-of-speech tagging via the TreeTagger module integrated into TXM was also applied to the corpus at import in order to allow lemma and part-of-speech statistics and queries.

4.2 Analysis

32 The annotated corpus (only the content inside <text> tags, without metadata) contained 6,512 items (unique words) with 76,558 occurrences in the text.[22]

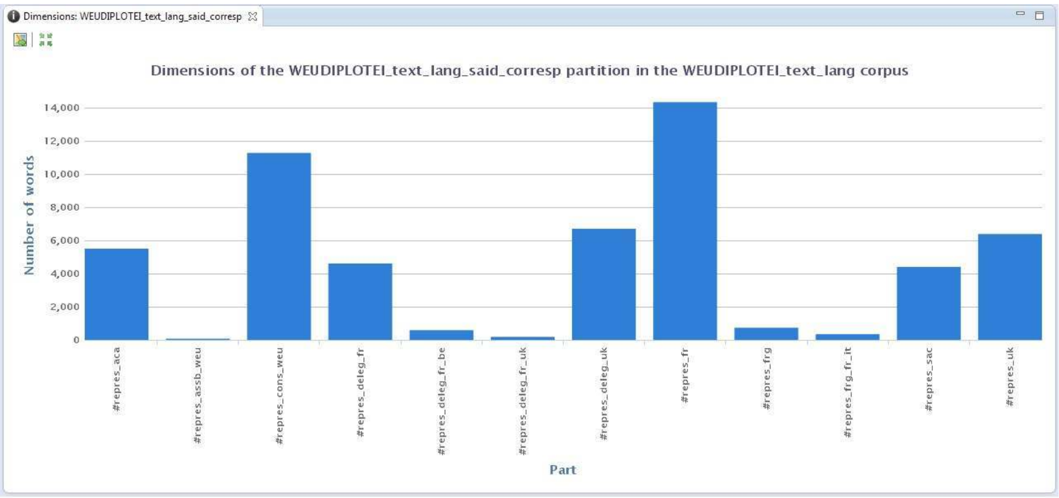
4.2.1 Partitioning

33 Given the identification and annotation of different semantic and structural elements in the encoding phase, TXM allows the creation of partitions (Textométrie 2014, section Construire une partition) by selecting a *Structure* unit and a corresponding *Property* (i.e., an XML element and one of its attributes) from the list of structural units and properties recognized by the software for the imported corpus.

34 For instance, as fragments of discourse spread throughout the documents were assigned to particular countries or institution representatives (section 3.2), a partition was created based on the `<said>` element and its `@corresp` attribute. Figure 1 shows the dimensions of the *representatives' discourse* partition, in number of words (occurrences). One can observe that for the selection of documents, there is a "dominance" of the French (14,338 occurrences) and WEU Council (11,276 occurrences) representatives' discourse, the categories with the lowest size being those corresponding to the French-English delegation (185) and WEU Assembly (72).

35 Other types of partitions were also created and analyzed: by *speaker*, based on the `<said>` element and its `@who` attribute, with dimensions varying from 41 (Brindeau) to 6,141 occurrences (Parodi); by *type of discourse*, using `<said>` and `@ana`, and counting 22,780 occurrences of oral discourse versus 32,355 occurrences of written dicourse; and by *subtype of institutional documents*, taking into account the `<text>` element and its `@subtype` attribute, with occurrences numbering between 2,684 for the study category and 38,565 for the minutes.

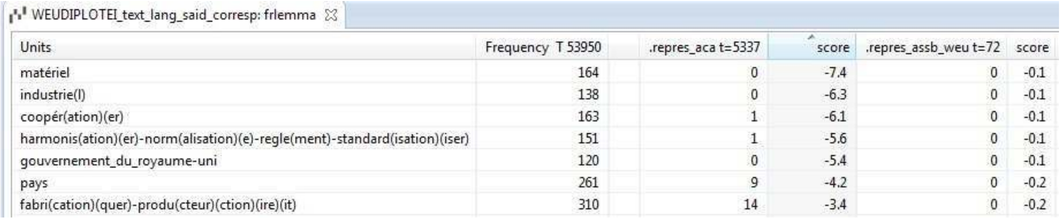
Figure 1. Dimensions of the *said_corresp* partition.



4.2.2 Specificities

36 The use of the *Specificities* feature (Textométrie 2014, section Spécificités) allows a comparison of the vocabularies: what is “specific” (either as “overuse” or “deficit”) in a part of a partition, as compared with the parent corpus and a certain threshold.[23] The feature is based on a probabilistic model (Lafon 1980) used in TXM to compute a log10 specificity score of a word property (e.g., word form, lemma, or part of speech) for a given part. In the analysis of the WEU-Diplo corpus, it was assumed that the specificity score may draw attention to forms “specific” to the discourse of different country/institutional representatives as compared with the whole. Figure 2 shows an extract from the specificities table computed for the *lemma* property and the *said_corresp* partition, sorted by increasing order of the specificity score corresponding to the *respres_aca* part.

Figure 2. Specificities table extract (lemma, *said_corresp*) sorted by increasing scores for *repres_aca*.



Units	Frequency T 53950	.repres_aca t=5337	score	.repres_assb_weu t=72	score
matériel	164	0	-7.4	0	-0.1
industrie(l)	138	0	-6.3	0	-0.1
coopér(ation)(er)	163	1	-6.1	0	-0.1
harmonis(ation)(er)-norm(alisation)(e)-regle(ment)-standard(isation)(iser)	151	1	-5.6	0	-0.1
gouvernement_du_royaume-uni	120	0	-5.4	0	-0.1
pays	261	9	-4.2	0	-0.2
fabri(cation)(quer)-produ(cteur)(ction)(ire)(it)	310	14	-3.4	0	-0.2

37 Each line in the table corresponds to a value of the chosen property (lemma or a group of lemmas) displayed in the *Units* column. The second column indicates the frequencies or number of occurrences of the property values in the corpus (with a total T). The other columns contain the number of occurrences of the property values in a part (cumulated by t) and are followed by a corresponding logarithmic score of specificity that can be positive or negative. The table may be sorted in increasing or decreasing order, according to a given column. In the case of an increasing score (as presented in the figure for *repres_aca*), the first property values displayed (e.g., *matériel, industrie(l)*)[24] indicate a deficit in use as compared to the whole corpus and the last ones displayed indicate an overuse, while the values with scores around 0 (inside a certain interval) are considered “trivial” (i.e., the specificity measure may not be pertinent for them).

38 Before creating a specificities table, a set of basic operations (merge, delete, export, import) are allowed via the *Lexical Table* feature (Textométrie 2014, section Table lexicale). The groups of units presented in figures 2 and 3 were created using *Lexical Table* and the *Merge Lines*

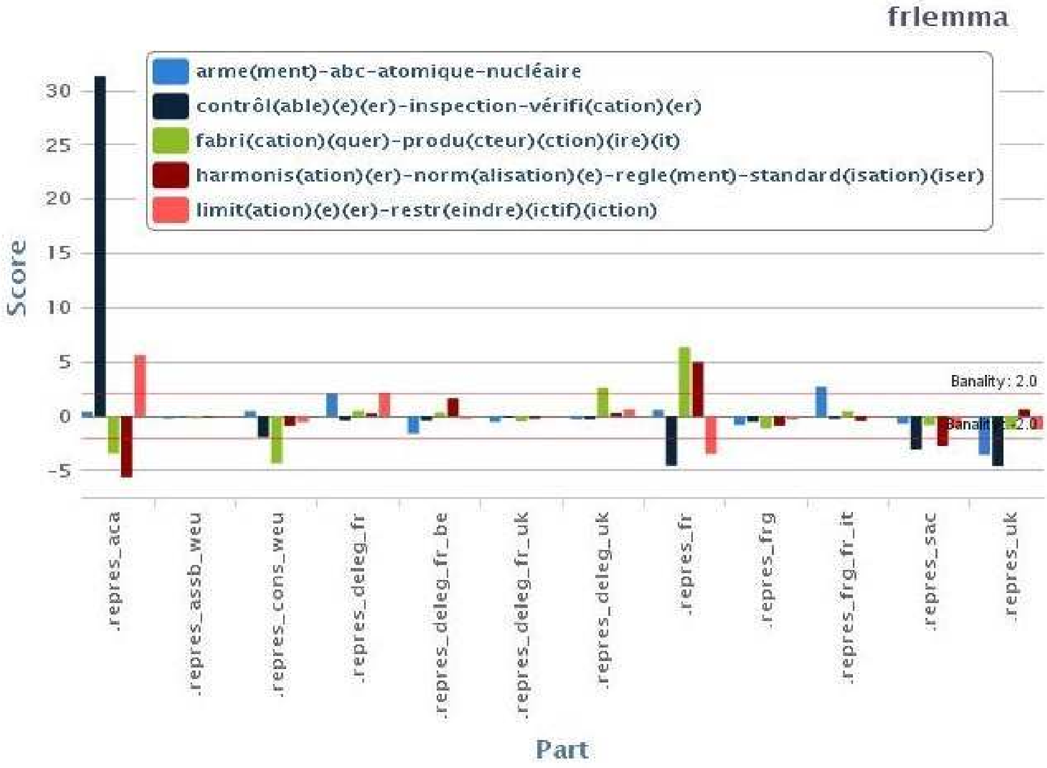
feature, in order to merge units considered to be close from a semantic point of view in the context, like, for instance, *harmonisation/harmoniser-norme/normalisation-règle/règlement-standard/standardisation/standardiser*).[25]

4.2.2.1 Graphical Representation

39 From a lexical table, a specificities table can be generated, as well as the corresponding diagram for a number of selected units, as shown in figure 3.

40 According to the diagram, the ACA representative's vocabulary is characterized by a high positive specificity score for the groups *control-inspection-verification* and *limitation-restriction*, and by negative specificity scores for the groups *fabrication-production*, *harmonization-normalization*.

Figure 3. Specificities diagram (lemma, *said_corresp* partition). Selection: *armament-control- production-standardization-limitation*.



41 The result is not very surprising given the role of the ACA: it was created to control the stocks of armament of its member states on the European continent (and less related to production/standardization). The negative specificity for *arme(ment)-abc-atomique-nucléaire*[26] in

the `repres_uk`'s discourse is not surprising either, considering that the United Kingdom, although interested in the topic, was not primarily concerned with this issue. Likewise, expected results are revealed for the positive specificity scores for *fabrication-production* and *harmonization-normalization* and the negative specificity scores for *control-inspection-verification* in the `repres_fr`'s discourse. The former are most probably linked to the selection of documents and, in particular, to the French memorandum presenting the armament agency which focuses on *fabrication-production* and *harmonization-normalization* (PWG/A/2).[27] Experiments excluding this document[28] from the corpus analysis have confirmed the hypothesis (with scores for *fabrication* and *harmonization* groups being lowered to the positive banality area). However, the negative specificity score for the *control* group in `repres_fr`'s discourse persisted[29] after the exclusion of the document from the analysis and may be associated either with the assertion of France's resistance to submitting its stocks to the ACA's controls or with an underrepresentation of the *control* topic in the selection of documents.

4.2.2.2 Synthesis Tables

42 A more general, comparative analysis can be provided for the positive and negative specific forms appearing in the representatives' discourse, as inspired by the synthesis tables proposed by Bergounioux et al. (1981) and Bonnafous (1981). Table 1 synthesizes the results for seven participants, a selection of shared lemmas or groups considered of interest for the study, and a set of specificities scores (in brackets) above and under the positive and negative banality thresholds.

Table 1. Comparative view of specificities scores. Selection: `said_corresp` partition.

Representative/ lemmas	`repres_aca`	`repres_cons_weu`	`repres_sac`	`repres_deleg_fr`	`repres_deleg_uk`	`repres_fr`	`repres_uk`
améri(cain)(que du nord)-états-unis			+ (2.3)			- (3.6)	
anglais-britannique-grande-bretagne - royaume-uni		- (4.3)	+ (5.9)			- (5.0)	+ (3.7)

continent_europeen- europ(e)(éen)(e occidentale)	- (2.1)	+ (2.7)	- (2.8)	+ (8.9)	- (2.4)		
contrôl(able)(e)(er)- inspection - verifi(cation) (er)	+ (31.4)		- (3.0)			- (4.5)	- (4.6)
coopér(ation)(er)	- (6.1)	+ (11.1)		+ (2.6)		- (5.8)	
finabel		- (3.2)	+ (32.5)		- (2.4)	- (2.8)	
g.e.i.p.	- (2.1)	+ (14.0)		+ (3.5)	- (2.6)	- (6.0)	- (2.5)

harmonis(ation)(er)
- norm(e)(alisation)
- règle(ment) -
standard(isation)(iser)
	- (5.6)		- (2.7)			+ (5.0)	

43 The high positive specificity score for *coopération/coopérer*[30] in the repres_cons_weu's discourse can be explained by the role of arbiter and conciliator of the WEU Council, which was intended to promote cooperation among its members in all the domains. A closer look at the contexts where this group appears shows recurrent, general patterns like *coopération en matière d'armements*, *coopération des pays européens en matière d'armements*, and *cooperation européenne en matière d'armements*[31] occurring both in the repres_cons_weu and repres_deleg_fr's discourse (which also displays a positive score but with a lower value) or more specific occurrences, such as *coopération intergouvernementale en matière de recherche, coopération intergouvernementale en matière d'études*[32] (repres_cons_weu), *coopération en matière de missiles*, or *coopération européenne dans le domaine aéronautique*[33] (repres_deleg_fr).

44 The highest positive score in the table (*finabel* for `repres_sac`) may be explained by the frequent mentions of the organization during the meetings of the SAC (acting as a link between it and the United Kingdom, not a member of Finabel),[34] as well as by the adherence of Great Britain to this organization referred to in the `repres_sac`'s discourse (see also section 4.2.2.4). The second highest positive and negative scores (*g.e.i.p.*:[35] for `repres_cons_weu` and `repres_fr`, respectively) are less clear but one can observe that the term tends to co-occur with the name of another organization (*c.d.n.a.*)[36] in the `repres_cons_weu`'s discourse, while being completely absent (0 occurrences) from the French representatives' discourse. The main reason seems to be the substance of the discussions linked to the competences of different organizations about standardization. Further examination is also needed to interpret the deficit of the group *control-inspection-verification* reflected by negative specificity scores for `repres_uk` and `repres_fr` (0 and 7 occurrences, respectively, out of a total of 85) that can be determined, as already mentioned, by the selection of documents, potentially more centered on the production and standardization of armaments than on their control.

4.2.2.3 Lexical Profile

45 Another type of analysis resulting from the encoding was an exploration of the combination of lemma and part-of-speech (POS) tagging and specificity measures, which may be related to the so-called "lexical profile" (Guyard 1981) of a participant in the institutional discourse. It consists of a list of relevant lemmas (with positive specificities above the banality threshold) and corresponding to certain parts of speech. Table 2 illustrates this type of profile for two representatives (France, United Kingdom) as individuals and three categories of POS (noun, adjective, and verb), obtained by taking into account specificity scores for the `said_who` partition.

Table 2. Lexical profiles. Selection (lemmas, `said_who` partition).

Part of speech / Name, occurrences	**Noun**	**Adjective**	**Verb**
Chauvel (FR) (1072)	*arme, accord_d'exécution, recensement, mise, choix*	*commun, équitable, régional, secret*	*procéder*

Lloyd (UK) (845)	*pays, discussion, arrangement, gouvernement_britannique, coopération*	*bilatéral, déterminé, multilatéral, analogue, final*	*engager, associer*

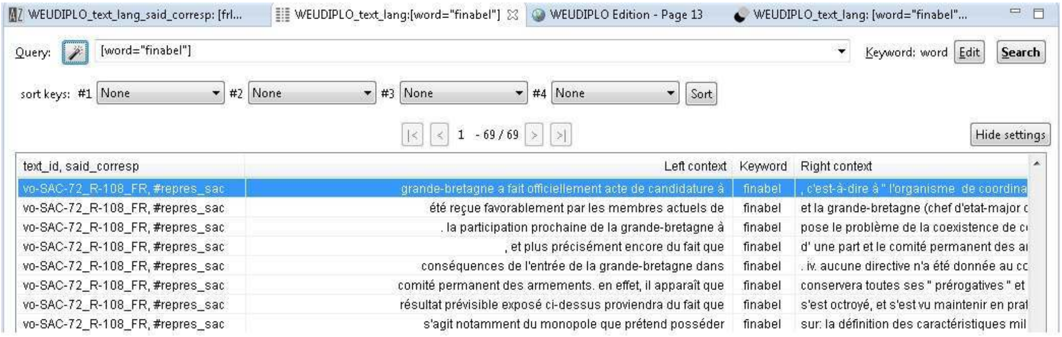
46 For the adjectives, we can point out an opposition on the axis *commun* versus *bilatéral, multilatéral*[37] manifested in contexts such as *programme (régional), intérêt, défense, études, fonds commun(e)(s)*[38] (Chauvel) versus *base, discussion, arrangements, comités directeurs bilatéra(l)(le)(ux)*[39] (Lloyd). The first profile is probably less clearly defined, but for the second one, the association of the lemmas provided for all three categories seems to convey a certain sense of action towards cooperation and dialogue.

47 Similar specificities-based analyses (not described here in detail) were performed for other categories of word properties: POS, for instance, which indicates a high positive specificity score (17.95) for the conditional verbal form in the `repres_fr` discourse (`said_corresp` partition); or other partitions, such as `said_ana` or `text_subtype`, that take into account the type of annotated discourse (oral/written) or the subtype of the document (minutes, note, study, or memorandum).

4.2.2.4 Queries, Concordances, and Co-occurrences

48 The analysis of specificities was combined with other methods, both of a quantitative and qualitative nature, for examining the documents, for instance, by querying for specific word properties (lemmas, word forms, POS, or combinations of these elements) and by means of the concordances and co-occurrences features (Textométrie 2014, sections Construire une concordance, Cooccurrences, Lexique et Index). Figure 4 presents the results of a query for "finabel" and the corresponding list of concordances that displays a left and right context, the file, and the representative's discourse containing the word (i.e., `repres_sac` with the highest positive specificity score for this unit, as shown in table 1).

Figure 4. Concordances. Extract (“finabel,” `said_corresp` partition).


WEUDIPLO_text_lang_said_corresp: [frl...
WEUDIPLO_text_lang:[word="finabel"]
WEUDIPLO Edition - Page 13
WEUDIPLO_text_lang: [word="finabel"...
Edit
Search
Sort
Hide settings

Query: [word="finabel"]
Keyword: word

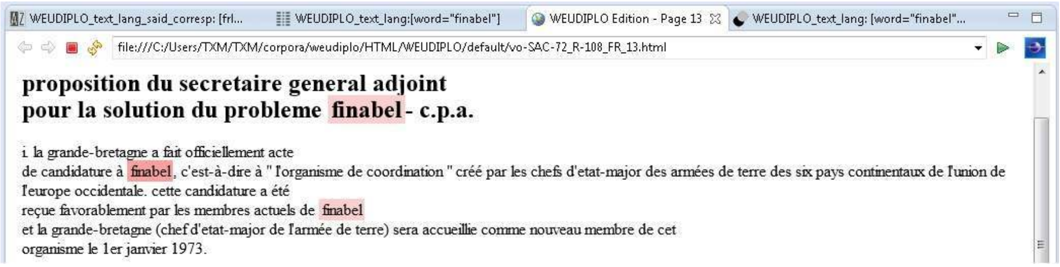
sort keys: #1 None #2 None #3 None #4 None

1 - 69 / 69

text_id, said_corresp	Left context	Keyword	Right context
vo-SAC-72_R-108_FR, #repres_sac	grande-bretagne a fait officiellement acte de candidature à	finabel	, c'est-à-dire à " l'organisme de coordina
vo-SAC-72_R-108_FR, #repres_sac	été reçue favorablement par les membres actuels de	finabel	et la grande-bretagne (chef d'etat-major c
vo-SAC-72_R-108_FR, #repres_sac	. la participation prochaine de la grande-bretagne à	finabel	pose le problème de la coexistence de c
vo-SAC-72_R-108_FR, #repres_sac	, et plus précisément encore du fait que	finabel	d' une part et le comité permanent des a
vo-SAC-72_R-108_FR, #repres_sac	conséquences de l'entrée de la grande-bretagne dans	finabel	. iv. aucune directive n'a été donnée au cc
vo-SAC-72_R-108_FR, #repres_sac	comité permanent des armements. en effet, il apparaît que	finabel	conservera toutes ses " prérogatives " et
vo-SAC-72_R-108_FR, #repres_sac	résultat prévisible exposé ci-dessus proviendra du fait que	finabel	s'est octroyé, et s'est vu maintenir en prat
vo-SAC-72_R-108_FR, #repres_sac	s'agit notamment du monopole que prétend posséder	finabel	sur: la définition des caractéristiques mil

49 A double click on a line in the concordances list provides a highlighted document view, as illustrated in figure 5.

Figure 5. Document view with highlighted items. Extract. WEU-Diplo: SAC/72/R-108[40] .


WEUDIPLO_text_lang_said_corresp: [frl...
WEUDIPLO_text_lang:[word="finabel"]
WEUDIPLO Edition - Page 13
WEUDIPLO_text_lang: [word="finabel"...

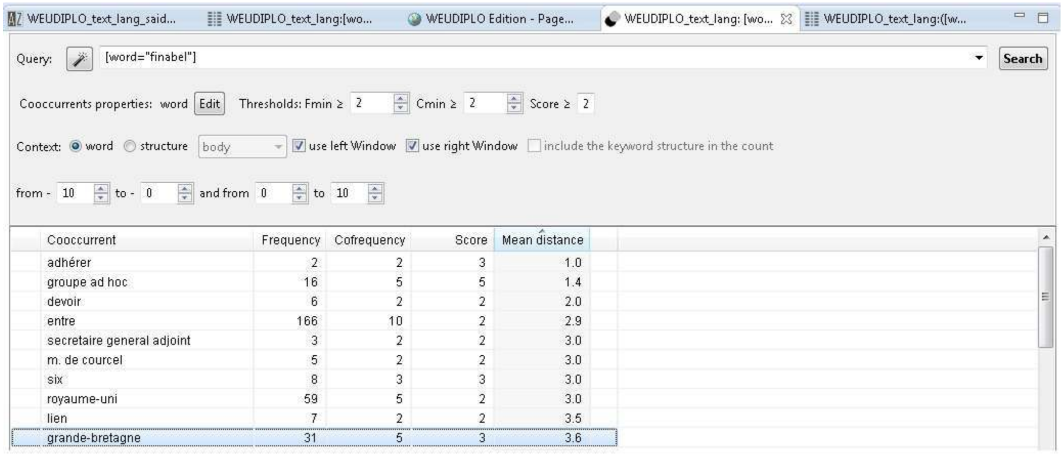
file:///C:/Users/TXM/TXM/corpora/weudiplo/HTML/WEUDIPLO/default/vo-SAC-72_R-108_FR_13.html

proposition du secretaire general adjoint
pour la solution du probleme finabel - c.p.a.

i. la grande-bretagne a fait officiellement acte
de candidature à finabel, c'est-à-dire à " l'organisme de coordination " créé par les chefs d'etat-major des armées de terre des six pays continentaux de l'union de
l'europe occidentale. cette candidature a été
reçue favorablement par les membres actuels de finabel
et la grande-bretagne (chef d'etat-major de l'armée de terre) sera accueillie comme nouveau membre de cet
organisme le 1er janvier 1973.

50 Co-occurrences may also be displayed and sorted according to the frequency, co-frequency, specificity score, or the mean distance between the keyword and the co-occurrent, as in the example presented in figure 6. Some of the co-occurents can be observed in the document view as well, figure 5 above.

Figure 6. Co-occurents of *finabel* sorted by mean distance. Extract (`said_corresp` partition).


WEUDIPLO_text_lang_said...
WEUDIPLO_text_lang:[wo...
WEUDIPLO Edition - Page...
WEUDIPLO_text_lang: [wo...
WEUDIPLO_text_lang:([w...
Search
Edit

Query: [word="finabel"]

Cooccurrents properties: word Thresholds: Fmin ≥ 2 Cmin ≥ 2 Score ≥ 2

Context: ● word ○ structure body ☑ use left Window ☑ use right Window □ include the keyword structure in the count

from - 10 to - 0 and from 0 to 10

Cooccurrent	Frequency	Cofrequency	Score	Mean distance
adhérer	2	2	3	1.0
groupe ad hoc	16	5	5	1.4
devoir	6	2	2	2.0
entre	166	10	2	2.9
secretaire general adjoint	3	2	2	3.0
m. de courcel	5	2	2	3.0
six	8	3	3	3.0
royaume-uni	59	5	2	3.0
lien	7	2	2	3.5
grande-bretagne	31	5	3	3.6

51 As illustrated in the previous section and in figures 4, 5, and 6, Great Britain and Finabel often co-occur in the discourse. This is probably (1) because of the particular attention of the SAC to informing the British representatives about the activities of the organization and (2) because the adherence of the United Kingdom to Finabel and its consequences is often brought into discussion.

52 In order to avoid misinterpreting the specificities or to confirm some of the hypotheses suggested by this method, we often needed to combine it with co-occurrences, concordances, and visualization of the document. Therefore, co-occurrences can provide a quantitative perspective on the co-presence of some words, lemmas, or entities in the context of a given target, which in combination with concordances and document visualization may support qualitative analysis and interpretation.

4.2.3 Results Discussion

53 The TEI XML encoding and TXM analysis related to the research questions on arms design, production, and control within the WEU have enabled a set of more or less predictable results, the latter needing further examination. Among the former, we can mention those referring to the SAC and ACA roles. Arms production and control was a major part of WEU's work, despite its somewhat mixed record in this area. Protocol IV of the Modified Brussels Treaty established the ACA, while after advocating the establishment of an armaments agency in its memorandum of January 3, 1955,[41] France proposed the setting up of the SAC during a meeting of the working party on production and standardization of armaments, on the basis of Article VIII(2) of the Modified

Brussels Treaty. The SAC was subsequently created on May 7, 1955. Although the United Kingdom never opposed the SAC's activities, it actively attempted to restrict its role, both because of its resistance to any notion of supranationality[42] and because it was convinced that NATO and the organizations related to it were more effective and better positioned to achieve standardization in the field of armaments.[43] The interpretation of the less predictable results is not straightforward, since they may have been determined by an under- or overrepresentation of certain elements in the discourse, based on the selection of documents. The same could be said about the negative specificity score for the *control* group in `repres_fr`'s, but this finding is also likely to be associated with the assertion of France's resistance to submitting its nuclear stocks to the ACA's controls and the need to avoid making statements on the subject. Since the size of the corpus was relatively small, and not all the information for the documents on the selected topic and their types in the WEU archive was available, extrapolations about the TXM probabilistic model and the observed linguistic patterns at a larger scale than the pilot sample should be avoided at this stage.

54 The TEI XML combined with the TXM analysis tools can also reveal inconsistencies which may draw attention to the need for further encoding and testing additional documents. On the other hand, it is also important to take into consideration how far (or how well) the researcher/user knows the content of the documents, as a lack of context can sometimes lead to misinterpretation.

5. Conclusion

55 The goal of the pilot project has been to address research questions mainly related to the French and British positions on the topic of arms design, production, and control within the WEU from 1954 to 1982. In particular, we were interested in combining traditional, historical inquiry with TEI XML encoding and decoding in a corpus analysis phase for the identification and interpretation of linguistic patterns in the discourse of different countries and institutional representatives on armaments issues.

56 Given the small scale of the corpus used in the project and the fact that it may not be a sufficiently representative sample means that a generalization of the results should not be performed without extended testing on an additional set of documents, conducting further evaluation of the probabilistic model and an estimation of the sample as compared to the collection from which it was extracted. From a methodological perspective, however, the TEI XML encoding and decoding

experiments have proved that the approach can assist qualitative and quantitative methods for the study of historical and discursive phenomena in a collection of institutional documents and a chosen theme. More precisely, the encoding may be quite helpful for researchers, even if they have no previous knowledge of the content of the documents. That is to say, if meaningful thresholds for analysis are set, one can find out the main topic of the documents, and the tags may help in discovering who the speakers are and the main "orientation" of their speech (i.e., their position) in terms of what is specific to their discourse. However, in order to enhance the reading of the specificities, combination with other methods, such as concordances, co-occurrences, and document visualization, is required.

57 Going back to the initial idea of considering a new link in the decoding–encoding chain suggested by Burnard, we see the TEI *encoding* as adding a "material" layer to the original text, which further supports both machine and human interpretation (*decoding*). In a larger sense, despite the inherent bias and limitations related to the selection and the number of documents used in the study, we have attempted to prove the "materiality" of the TEI encoding and decoding as a basis for hermeneutic inquiry in the quest for producing knowledge via digital "instruments," as Capurro and Idhe have previously stated in their accounts on a digital and material hermeneutics.

BIBLIOGRAPHY

Andrews, Tara. 2013. "The Third Way: Philology and Critical Edition in the Digital Age." *Variants: The Journal of the European Society for Textual Scholarship* 10: 61–76.

Bergounioux, Alain, Michel Launay, Josette Lefvre, René Mouriaux, and Jean-Pierre Sueur. 1981. "Le vocabulaire des confédérations syndicales ouvrières: une analyse des spécificités." *Mots* 2: 139–56. doi:10.3406/mots.1981.1025.

Bonnafous, Simone. 1981. "Le vocabulaire spécifique des motions Mitterrand, Rocard et CERES au congrès de Metz (1979)." *Mots* 3: 79–94. doi:10.3406/mots.1981.1040.

Burnard, Lou. 1998. "On the hermeneutic implications of text encoding." Accessed March 6, 2016. http://users.ox.ac.uk/~lou/wip/herman.htm.

Capurro, Rafael. 2010. "Digital Hermeneutics: An Outline." In *AI & Society* 25(1): 35–42. Accessed March 6, 2016. http://www.capurro.de/digitalhermeneutics.html. doi:10.1007/s00146-009-0255-9.

Delhauteur, Dominique. 1991. “Les activités du Conseil de l’UEO en matière de coopération dans le domaine des armements.” In dossier “notes et documents” no. 160. Brussels: GRIP (Groupe de recherche et d’information sur la paix et la securite). http://www.grip.org/fr/node/1014.

Guyard, Marie-Renée. 1981. “Spécificités d’auteurs dans *Le Surréalisme au service de la Révolution*.” *Mots* 2: 95–122. doi:10.3406/mots.1981.1023.

Hayles, Katherine. 2005. *My Mother Was a Computer: Digital Subjects and Literary Texts.* Chicago: University of Chicago Press. Excerpt: http://www.press.uchicago.edu/Misc/Chicago/321487.html.

Heiden, Serge. 2010. “The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme.” In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–398. Tokyo: Institute for Digital Enhancement of Cognitive Development, Waseda University. Accessed July 24, 2017. http://halshs.archives-ouvertes.fr/halshs-00549764/en.

Ihde, Don. 2003. “More Material Hermeneutics.” Paper presented at the meeting on “Hermeneutics and Science,” Tihany, Hungary, June 7 - 11, 2003. Accessed March 6, 2016. http://humanitieslab.stanford.edu/23/admin/download.html?attachid=178178.

Lafon, Pierre. 1980. “Sur la variabilité de la fréquence des formes dans un corpus.” *Mots* 1: 127–65. doi:10.3406/mots.1980.1008.

Manoff, Marlene. 2006. “The Materiality of Digital Collections: Theoretical and Historical Perspectives.” *portal: Libraries and the Academy* 6(3): 311–25. doi:10.1353/pla.2006.0042.

McGann, Jerome. 2001. *Radiant Textuality*. New York: Palgrave.

Nikander, Pirjo. 2008. “Constructionism and Discourse Analysis.” In *Handbook of Constructionist Research*, edited by James A. Holstein and Jaber F. Gubrium, 413–28. New York: The Guilford Press. Accessed March 6, 2016. http://www.helsinki.fi/sosiaalipsykologia/arkisto/Nikander%20ch21%202008.pdf.

Phillips, Nelson, Thomas B. Lawrence, and Cynthia Hardy. 2004. “Discourse and Institutions.” *Academy of Management Review* 29(4): 635–52. doi:10.5465/AMR.2004.14497617.

Pierazzo, Elena. 2014. *Digital Scholarly Editing: Theories, Models and Methods.* Surrey, England: Ashgate; HAL open archive, <hal-01182162>. Accessed March 6, 2016. http://hal.univ-grenoble-alpes.fr/hal-01182162/document.

Rémacle, Eric. 2009. “L’Union (de l’Europe) occidentale durant la guerre froide (1948–1989).” In *L’Amérique, l’Europe, l’Afrique (1945–1973) [America, Europe, Africa (1945–1973]*, edited by Eric Rémacle and Pascaline Winand, 187–234. Brussels: PIE-Peter-Lang.

Textométrie. 2014. *Manuel de TXM*, Version 0.7. Accessed March 6, 2016. http://textometrie.ens-lyon.fr/spip.php?rubrique64.

Thornborrow, Joanna Sarah. 2002 *Power Talk: Language and Interaction in Institutional Discourse.* Harlow: Longman.

Van Dijk, Teun A. 1993. "Principles of critical discourse analysis." *Discourse & Society* 4(2): 249–83. Accessed March 6, 2016. http://www.discourses.org/OldArticles/Principles%20of%20critical%20discourse%20analysis.pdf.

NOTES

[1] The exploitation of the WEU archives follows the decision C(11)05-Final of May 10, 2011, by the Permanent Council of the Western European Union, which appointed ANLux—the Archives nationales de Luxembourg—as the official depository of the WEU archives and gave the CVCE (http://cvce.eu/), a research and documentary center in European Integration Studies, the task of scientifically exploiting these holdings, including their publication in any form.

[2] The WEU was created on the October 23, 1954, with the signing of the "Modified Brussels Treaty" by France, Belgium, Luxembourg, the Netherlands, the United Kingdom—all five previously members of the Western Union created in 1948—and the Federal Republic of Germany invited to join the new organization. It became the first European Defence Organisation and its missions covered the settling of the problem of the Saar, the monitoring of German rearmament, and the promotion of the defense of Western Europe. The Treaty of Brussels contained a mutual defense clause in Article V. For more information, please consult http://www.cvce.eu/en/recherche/unit-content/-/unit/72d9869d-ff72-493e-a0e3-bedb3e671faa/1c06c877-402b-45d1-a126-792e99cf3fc3.

[3] Eric Rémacle referred to the "instrumentalisation" of WEU by these two countries, which he identified as "successive leader states" within the organization. See Rémacle 2009, 197.

[4] Related to the specific research question, that is, the study of linguistic patterns in the discourse of different country/institutional representatives as described in section 4.

[5] http://www.tei-c.org/oxgarage/.

[6] For example, CR(58)8 = CR (for *compte-rendu*, or minutes, of a Council meeting), the year in brackets, then the document number; PWG/CR/4 = PWG (we think that PWG stands for Production Working Group), CR (for *compte-rendu*) and the number of the meeting (the fourth meeting); C(80)40—we have noticed that the letter C was generally used for the final version of the reply to a recommendation or written question, followed by the date in brackets, although the meaning of

the second number is not clear. The letters WPM followed by a number in brackets were used for the various versions of replies to recommendations and/or written questions, but we think that the date was not indicated in brackets during the early years of WEU's activity, until the end of the 1960s or the early 1970s; if a draft was amended by a working group, the version number was added after a solidus, e.g., WPM(77)25/1.

7 Examples of file codes: CPA-043; ACA-200; 421.00; 200.400.11 vol 1/1.

8 Speakers frequently referred to statements by other people from the same or another delegation.

9 The terms "delegate" and "representative" were used to distinguish the collective discourse of a country (for instance, in the answer to a recommendation) from the discourse of an individual (in a Council's meeting for instance), although the individual was also speaking for a country.

10 Other meetings were often mentioned, such as a previous or forthcoming ministerial meeting or a meeting of the North Atlantic Council.

11 TEI Consortium. 2015. TEI Guidelines Version 2.8.0. Last updated on April 6. http://www.tei-c.org/Vault/P5/2.8.0/doc/tei-p5-doc/en/html/.

12 En. "PRESENT: Prof. Dr. L. ERHARD: Federal Republic of Germany; M. A. de STAERCKE: Belgium…." Union de l'Europe occidentale. Commission intérimaire. Groupe de travail sur la production et la standardisation des armements. Troisième séance plénière tenue au Palais de Chaillot le 21 janvier 1955 à 15 heures 30. Paris: 22.01.1955. PWG/CR/2. Pp. 1–3; Annexe A; Annexe B. Archives nationales de Luxembourg (ANLux). http://anlux.lu/. Western European Union Archives. Armament Bodies. CPA/SAC. Comité permanent des armements. File CPA-033. Volume 1/1.

13 En. "Mr. Selwyn LLOYD considered that … ." Conseil de l'Union de l'Europe occidentale. Extract from minutes of the 108th meeting Ministers of WEU Council held on 5 March 1958. Rome. II. Echanges de vues sur la coopération en matière de recherche, développement et production d'armements. CR (58)8. Pp. 5–9. Archives nationales de Luxembourg (ANLux). http://www.anlux.lu. Western European Union Archives. Secretariat General/Council's Archives. 1954–1987. Subject dealt with by various WEU ORGANS. Year: 1958, 01/06/1957–30/04/1958. File 442.00. Volume 1/4.

14 Agence pour le contrôle des armements. Note. 04.1966. 6 p. Archives nationales de Luxembourg (ANLux). http://anlux.lu/. Western European Union Archives. Armament Bodies. ACA. Agency for the Control of Armaments. Year: 1957, 01/01/1957–31/12/1963. File ACA-035. Volume 1/1.

15 Example 4, En. “One year after, under analogous conditions, Minister LUNS had to assess as a president a position along the same lines … .” (Our translation, document for which we could not find an English equivalent in the WEU archives.)

16 See note 14 for the document reference. The ellipsis indicates missing paragraphs, omitted here for concision.

17 Distinguished via the values of the `@ana` attribute, as related to written versus oral discourse.

18 Agence pour le contrôle des armements. Division III. Note à l’intention du Directeur de l’ACA. Contrôle sur place des armes atomiques se trouvant dans des dépôts britanniques sur le continent européen. 22 juin 1962. Archives nationales de Luxembourg (ANLux). http://anlux.lu/. Western European Union Archives. Armament Bodies. ACA. Agency for the Control of Armaments. Year: 1957, 01/01/1957–31/12/1978. File ACA-218. Volume 1/1. The ellipsis indicates missing paragraphs or fragments, omitted here for concision. As far as the intervention of Mr. Heath is concerned, it is not specified anywhere that he was actually speaking English, but since it was a working language, it seems logical that he was expressing himself in his native language. Thus, this could be a translation, most probably from an internal translation service.

19 https://gate.ac.uk/.

20 http://textometrie.ens-lyon.fr/?lang=en.

21 All the examples of analysis presented in the paper will consequently be displayed in lowercase.

22 A subcorpus based on the `@lang="fr"` property (an attribute of the `<text>` element) was created in TXM for the analysis of the documents’ content, excluding the data from the `<teiHeader>`. The whole corpus (`<teiHeader>` included) comprised 7,015 items and 105,897 occurrences.

23 In TXM, it is called the *banality threshold*, fixed by default at the value of +/- 2.0 for positive and negative specificities scores, respectively. In figure 3, the banality thresholds are rendered by (red) horizontal lines.

24 En. *material, industry-industrial.*

25 En. *harmonization/harmonize-norm/normalization-rule/regulation-standard/standardization/standardize.*

26 En. *abc/atomic/nuclear arms/armament.*

27 See note 41 for the document reference.

[28] The size of the `repres_fr` part in `said_corresp` partition decreased from 14,338 to 10,269 words after excluding PWG/A/2.

[29] In fact, it increased slightly, from -4.5257 (7 out of a total of 85 occurrences in the whole corpus) to -4.7338 (3 out of 81 occurrences).

[30] En. *cooperation/cooperate*.

[31] En. *cooperation on armament matters, cooperation on armament matters among European countries, European cooperation on armament matters.*

[32] En. *intergovernmental cooperation on research matters, intergovernmental cooperation on study matters.*

[33] En. *cooperation on missiles matters, European cooperation in the aeronautic field.*

[34] Finabel is a Land forces organization created in 1953 with five initial members: France, Italy, Belgium, Luxembourg, and the Netherlands. Germany became a member in 1956.

[35] *Groupe européen indépendant de programme*. En. *Independent European Programme Group* (IEPG).

[36] *Conférence des directeurs nationaux des armements*. En. *Conference of National Armaments Directors* (CNAD).

[37] En. *common, joint, mutual* versus *bilateral, multilateral.*

[38] En. *joint regional programme, common interest, common defence, common studies, mutual funds.*

[39] En. *bilateral basis, bilateral discussion, bilateral arrangements, bilateral steering committees.*

[40] Compte rendu de la 108ème Réunion du Comité Permanent des Armements, tenue à Paris, le 29 septembre 1972. Union de l'Europe occidentale. Comité permanent des armements. Compte-rendu de la 108ème réunion du Comité permanent des armements tenue à Paris le 29 septembre 1972. Paris: 18.10.1972. SAC (72)R/108. pp.[s.p]; 3–4; annexe; pp.1–5. Archives nationales de Luxembourg (ANLux). http://www.anlux.lu. Western European Union Archives. Secretariat-General/Council's Archives. 1954–1987. Organs of the Western European Union. Year: 1967, 16/03/1956–30/04/1967. File 250.10. Volume 2/2.

[41] Western European Union. Interim Commission. Working Party on Production and Standardisation of Armaments. *Secretary General's note*. Paris: 17.01.1955. PWG/A/2.18 p. Archives nationales de Luxembourg (ANLux). http://anlux.lu/. Western European Union Archives. Armament Bodies. CPA/SAC. Comité permanent des armements. File CPA-032. Volume 1/1; Delhauteur 1991, p. 6.

[42] Western European Union. Interim Commission. Working Party on Production and Standardisation of Armaments. Statements made by the delegations at the first and second sessions of the Working Group. Paris: 21.01.1955. PWG/A/6. Annex A and Annex F. Archives nationales de Luxembourg (ANLux). http://anlux.lu/. Western European Union Archives. Armament Bodies. CPA/SAC. Comité permanent des armements. File CPA-032. Volume 1/1.

[43] The National Archives of the UK (TNA). Foreign Office, Western Organisations and Co-ordination Department and Foreign and Commonwealth Office, Western Organisations Department: Registered Files (W and WD Series). Western European Union (WEU). Future of Standing Armaments Committee of Western European Union. 01/01/1975–31/12/1975, FCO 41/1749 (Former Reference Dep: WDU 11/1 PART B).

ABSTRACT

By combining traditional historical enquiry with TEI XML encoding and decoding in a corpus analysis phase, the project aims at addressing research questions mainly related to the French and British positions on the topics of armament design and production and of armament control within the Western European Union (WEU) from 1954 to 1982. The paper focuses on the annotation of speakers (different countries and institutional representatives) and their discourse in a selection of institutional documents (minutes, notes, studies, memoranda) (encoding phase) and the identification of linguistic patterns on armament issues in their discourse, as well as the interpretation of results (decoding phase).

From a larger perspective, the study considers the TEI encoding as adding to the original text a "material" layer that further supports both machine and human interpretation (decoding). In this sense, this study may move closer to the concept of "material hermeneutics," by understanding code, and digital technology in general, as an instrument we can use in hermeneutic ways to produce knowledge.

INDEX

Keywords: WEU and armament digital edition, text encoding, text analysis, digital hermeneutics, materiality

AUTHORS

FLORENTINA ARMASELU

Florentina Armaselu is involved in text and technologies research at the University of Luxembourg. She obtained a PhD in comparative literature (2010) and a MSc in computational linguistics (2003) at the University of Montreal, Canada. Her current research focuses on digital editions, text encoding and text analysis.

VERÓNICA MARTINS

Verónica Martins specialized in European Studies and her current research focuses on the Mediterranean region and security questions, mainly counter-terrorism. She attended the College of Europe where she got a MA in European Studies. In 2012, she obtained her PhD in Political Science, Europe specialisation, in a joint international supervision in Sciences-Po Paris and University of Minho. She worked as a researcher at the Centre Virtuel de la Connaissance sur l'Europe on a project on Franco-British relations within the Western European Union.

CATHERINE EMMA JONES

Dr Catherine (Kate) Jones is interested in the development of useful and usable systems and data analysis to form meaningful narratives. She is based at the University of Luxembourg. She studied for an MSc in Geographical Information Systems at University College London in 2002 and went on to complete a Knowledge Transfer Partnership, PhD and Post Doc at the same university, with a focus on interdisciplinary research and mapping technologies.