Digitizing and Encoding a Multilingual Literary Review: Commerce Numérique

Antonietta Sanna, Rachele Cinerari, Federico Boschetti, Ouafae Nahli

Digitizing and Encoding a Multilingual Literary Review: *Commerce Numérique*

Antonietta Sanna
University of Pisa
Pisa, ITALY
antonietta.sanna@unipi.it

Rachele Cinerari
University of Pisa
Pisa, ITALY
rachele.cinerari@phd.unipi.it

Federico Boschetti
ILC-CNR & VeDPH
Venice, ITALY
federico.boschetti@ilc.cnr.it

Ouafae Nahli
ILC-CNR
Pisa, ITALY
ouafae.nahli@ilc.cnr.it

***Abstract*—*Commerce* was an important literary review founded in Paris by Princess Margherita Caetani, Prince Roffredo Caetani's wife. Born in America, she was polyglot and maecenas. Between 1924 and 1932 she surrounded herself with three prestigious collaborators: Paul Valéry, Léon-Paul Fargue, Valéry Larbaud. The review promoted the translation of World and European literature in French, translating some of the most important authors like Joyce, T.S. Eliot, Pirandello, Ungaretti, Saint-John Perse, Rilke, Hofmannsthal. The aim of this project is to promote by digitizing the dissemination of the review, to develop studies and research concerning the Caetani family's cultural activities in Europe.**

All the volumes of the literary review *Commerce* have been scanned, acquired by OCR and encoded in TEI-XML. The cultural value of the operation is discussed and the work-flow to create the digital textual corpus is described in detail.

***Index Terms*—Review *Commerce*, OCR, TEI encoding, literary review, digital resources**

I. Introduction

This article describes stages of our work to create a digital version of the *Commerce* Journal. Section 1 allows to place *Commerce* review in the European cultural context at the beginning of the 20th century. Section 2 provides a descriptive study of *Commerce* regarding the quantity and content of articles. Sections 3 and 4 describes the stages of digitization and TEI coding of *Commerce* review texts. Finally, a conclusion closes the article and describes the results that will be available to the scholarly community.

II. *Commerce* and the cultural context of Europe on the early 20th century

Commerce was a very important literary review founded in Paris after the First World War by Princess Marguerite Caetani, Prince Roffredo Caetani's wife . There were two great additions from the Parisian literary scene from 1914. The first was Adrienne Monnier with her bookshop named ”La Maison des Amis des Livres”. In 1917 Monnier began to hold readings by authors of contemporary French literature. In the same year, Sylvia Beach, a young American women who studied modern French literature, created a bookshop specializing in English and American books. She opened the legendary ”Shakespeare and Company”. Many French authors such as André Gide, Paul Valéry, Valéry Larbaud, Jules Romain, met in that modern Salon littéraire American and English writers such as Ernest Hemingway, F. Scott Fitzgerald, Ezra Pound, James Joyce.

After 1918 peace brought the revival of cultural life and Paris became a home for many artists. The *Nouvelle Revue Française* (NRF), founded by André Gide, Jean Schlumberger, Gaston Gallimard became a model of modern literary review, publishing the most important texts of European Modernism.

Marguerite Caetani regularly read the NRF that she considered indisputably France's leading intellectual review. She had met Paul Valéry before 1914, but she probably met Adrienne Monnier, Léon-Paul Fargue, and Valéry Larbaud at the events organized in the Shakespeare and Company bookshop.

At one of the many Sunday gatherings at Roffredo and Marguerite Caetani's home in Versailles, the idea of founding a review only dedicated to poetry, prose and drama, took shape. According to Marguerite it happened in the following way:

> One day, Valéry said out of the blue: “Why don't continue our conversations, our dialogues, in published form? As a title I suggest ‘Commerce’, exchange of ideas. Everyone present was delighted by the idea. The directors (Larbaud, Valéry, Fargue) were appointed straightaway. Adrienne Monnier and I were put in charge of getting it going and we began at once. What the result was, remains for you to judge. I was helped immensely by Paulhan who allowed me to search among the manuscripts that he received for the NRF as well as by Alexis Léger who chose the poems that we published [4].

This is the origin of a brilliant chapter of a literary adventure in which Marguerite Caetani assumed the leading role between 1924 and 1932 to promote the translation like new European language. The *traduction d'auteur* (Author's translation) was a new form of creation destined to improve exchange between different cultures.

Marguerite Caetani could also count on the collaboration of some writers and intellectuals from other European nations: T. S. Eliot, Giuseppe Ungaretti, Rainer Maria Rilke, Hugo von Hofmannstahl, Rudolf Kassner, D. S. Mirskij [9] [10]. The review included literary works – excerpts from novels, short stories, poems, drama – not only from France but also from ten different countries.

III. *Commerce* corpus: quality and quantity

The most important and peculiar characteristic of *Commerce* was the multicultural and international vocation, and the publication of foreign literary works translated into French. Poems were translated by poets, prose texts by novelists. Some authors also self-translated some of their works in French, for example Hugo von Hofmannstahl and Giuseppe Ungaretti [2]. This characteristic made *Commerce* a fundamental review and gives the opportunity to nowadays researchers to better understand the European literary field of the early 20th Century, where Paris was a fundamental city, crossed by cultures and artists from all over the world.

About 240 texts were published on *Commerce*, most of them previously unpublished. We can count about 140 French texts, most of which written by Paul Valéry, Leon-Paul Fargue and Valéry Larbaud, but also from many other important French writers as for example Antonin Artaud, André Breton and Henri Michaux.

Thanks to the cooperation of advisors from Germany, the magazine hosted almost 20 German works including some by Franz Kafka, Friedrich Hölderlin and Friedrich Nietzsche. More than 20 texts by English authors were published on the review, including some fragments from *Ulysses* from James Joyce and one text from *To the Lighthouse* by Virginia Woolf, that were first published in the magazine. The review also hosted works by Italian authors – including Giacomo Leopardi and Giuseppe Ungaretti – and some literary works from China, Spain, Belgium, Denmark, Greece and Russia including authors like Søren Kierkegaard, Cheng Tcheng, Boris Pasternak, Jose Ortega y Gasset.

Commerce published a combination of contemporary and ancient texts, many of these were published on the review for the first time and some authors debuting at the time would soon become some of the most important authors of the 20th Century [9].

IV. Corpus digitization: Image pre-processing, OCR and manual correction

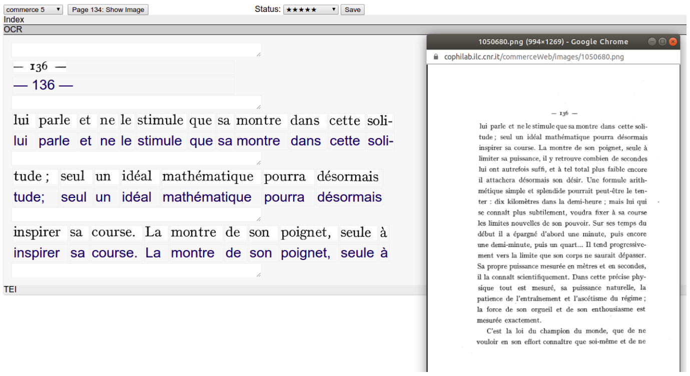
The corpus consists of 29 volumes and a short index of the first 16 volumes, for a total amount of more than 6,000 pages and about 2,000,000 tokens.

All the volumes provided by the library of the University of Pisa and Camillo Caetani Foundation have been scanned at the CNR-ILC. Page images have been pre-processed in order to optimize the character recognition by the usual page image operations, such as splitting, deskewing, dewarping, despeckling, binarization, performed through scanTailor [1].

The Optical Character Recognition (OCR) has been executed through Tesseract[2] in multi-language modality (mainly French, English, Italian, German, and Spanish).

Proof reading has been performed by using the CoPhiProofReader, a web application for collaborative OCR correction created at the CNR-ILC, as shown in Figure 1.

The CoPhiProofReader is inspired to the WikiSource[3] correction tools, in order to take trace of multiple proof-reading and supervision phases.


commerce 5
Page 134: Show Image
Status: ★★★★★
Save
Index
OCR
TEI
1050680.png (994×1269) - Google Chrome
cophilab.ilc.cnr.it/commerceWeb/images/1050680.png

— 136 —

lui parle et ne le stimule que sa montre dans cette soli-

tude; seul un idéal mathématique pourra désormais

inspirer sa course. La montre de son poignet, seule à

— 136 —

lui parle et ne le stimule que sa montre dans cette solitude; seul un idéal mathématique pourra désormais inspirer sa course. La montre de son poignet, seule à limiter sa puissance, il y retrouve combien de secondes lui ont autrefois suffi, et à tel total plus faible encore il attachera désormais son désir. Une formule arithmétique simple et splendide pourrait peut-être le tenter : dix kilomètres dans la demi-heure ; mais lui qui se connaît plus subtilement, voudra fixer à sa course les limites nouvelles de son pouvoir. Sur ses temps du début il a épargné d'abord une minute, puis encore une demi-minute, puis un quart... Il tend progressivement vers la limite que son corps ne saurait dépasser. Sa propre puissance mesurée en mètres et en secondes, il la connaît scientifiquement. Dans cette précise physique tout est mesuré, sa puissance naturelle, la patience de l'entraînement et l'ascétisme du régime ; la force de son orgueil et de son enthousiasme est mesurée exactement.

C'est la loi du champion du monde, que de ne vouloir en son effort connaître que soi-même et de ne

Fig. 1. CoPhi ProofReader

Usually systems for collaborative proof-reading are based on comparison of image and text page by page. But the comparison line by line between the image box of the original printed edition and the OCR result facilitates students and scholars to recognize the OCR errors, as demonstrated by studies on the ergonomics of the proof-reading [6].

V. TEI-XML encoding

A. Encoding criteria

We chose the TEI encoding because it is a *de facto* standard for the community of Digital Humanists and it is extremely flexible and inclusive of most text genres. It provides elements, attributes and other mechanisms for coding prose, poetry, theater, dictionaries, linguistic corpora and other literary and academic texts. Furthermore, the TEI offers a framework where it is possible to design specific customisations while remaining compliant with the guidelines. As results, encoded texts are interoperable and can be reused, managed and archived in the future [3], [7], [12].

Commerce is historically a print-based journal and, like most print journals, it uses a traditional hierarchical structure in issues, volumes and articles.

Considering the limited budget of the first phase of the project and the large amount of contents to encode, we decided to focus our attention on the minimalist structural and formatting elements of each volume. Indeed, the main goal of the first phase is the compliance of the digital edition to the typographical aspects of the original printed edition.

Great care has been taken to render the spatial distribution of the text on the page through encoding. It is not only a typographical choice related to the *mise en page* (layout) of the magazine, but mainly a fundamental feature for poetic, dramatic and literary texts in general. The formal and typographical aspect is important, mainly considering that many of the texts date back to the so-called modernist period; a period – dated approximately between 1918 and 1940 – rich in formal, linguistic and stylistic experimentation, during which a dense network of exchanges, influences and contacts was formed. Literary magazines played a key role in this process [5].

Each volume is preceded by a front matter and followed by a back matter, and is divided in articles, which are subdivided

[1] http://scantailor.org
[2] https://github.com/tesseract-ocr/tesseract
[3] https://wikisource.org

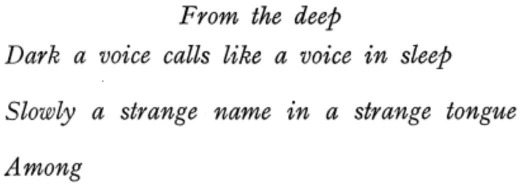
in sections and paragraphs for prose or strophes and verses for poetry. Articles are gathered inside the <group> element, which is provided to simplify the encoding of collections, anthologies, and cyclic works. In addition, the <group> element reflects the potentially complex internal structure of the journal.

Font styles (e.g. italics) are encoded but their semantics will be specified in a second stage of the work, in order to distinguish foreign words, quotations, key-words, etc.

The bibliographical references of *Commerce* have been recorded on Zotero[4], in order to make them available in different standard formats, such as BibTeX.

B. Parallel texts

Special attention has been devoted to encode poems in parallel with their translation. Figure 2 illustrates a snippet of the poem *"TRAIN-STOP : NIGHT"* in Volume 5, page 128 which is translated in French at page 129.



From the deep
Dark a voice calls like a voice in sleep
Slowly a strange name in a strange tongue
Among

Fig. 2. English Poem

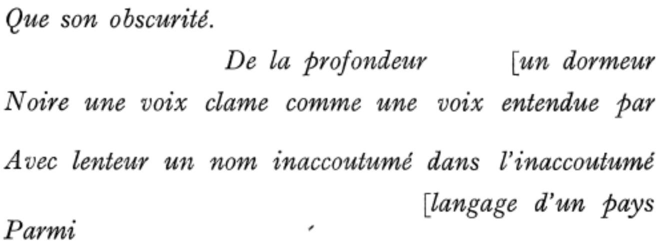
In the first phase of encoding, the poem in the original language is divided in strophes (`lg`: line groups) identified through the attribute `@xml:id` which provides a unique identifier for the element bearing it. The attribute `@xml:id` permits to link the original poem with the corresponding translation, strophe by strophe.

The identifier number is composed by the volume number, page and a progressive number. For example above, the verse 8 *From the deep* and the verse 9 *Dark a voice calls like a voice in sleep* belong to the fifth group of the poem in Volume 5, page 128, and recognized thanks to the identifier `@xml:id="lg5_128_5"`.

As we have already mentioned, texts and poems are in different languages. For this reason, it is important to indicate it through the attribute `@xml:lang`. In the example mentioned before, the original language of poem is tagged by `xml:lang="eng"`:

```
<lg xml:id="lg5_128_5" xml:lang="eng">
<space dim="horizontal" unit="character"
quantity="15"/><l rend="hi" n="8">From the
deep¶</l> <l rend="hi" n="9">Dark a voice
calls like a voice in sleep¶</l></lg>
```

Figure 3 shows the French translation of the poem *"TRAIN-STOP : NIGHT"* in Volume 5, in page 129. It is interesting to note that, in some cases, the French counterpart is split into two segments and the second segment is dislocated to the top row or the bottom row. For example, the verse *Noire une voix clame comme une voix entendue par un dormeur* is divided in two segments: *Noire ... par* and *un dormeur*. The second segment is presented on the printed edition on the same line of the previous verse, but separated by a bracket, which suggests the correct textual order.



Que son obscurité.
De la profondeur [un dormeur
Noire une voix clame comme une voix entendue par
Avec lenteur un nom inaccoutumé dans l'inaccoutumé
[langage d'un pays
Parmi

Fig. 3. Parallel French Poem

In order to preserve both the logical order of the segments and the original rendering of the printed edition, segments have been disposed in the same order they have on the printed edition but with an ordinal indication `n` between the segments.

Below, an example shows how the segments are represented in order to preserve both the textual and spatial order, which is compliant to the printed page. The line `<l n="9">` is composed with two segments `<seg n="1">Noire ... par</seg` and `<seg n="2">un dormeur</seg>` and we note that the second precede the first segment and it is located on the same line of the line `<l n="8">`.

The line `<l n="8">` and the segment `<seg n="2">` of the line `<l n="9">` are separated by a horizontal space estimated equal to 8 characters.

```
<lg corresp="#lg5_128_5" xml:lang="fra">
<space dim="horizontal" unit="character"
quantity="19"/><l n="8">De la
profondeur</l><space dim="horizontal"
unit="character" quantity="8"/><l
n="9"><seg n="2">[un dormeur¶</seg>
<seg n="1">Noire une voix clame comme
une voix entendue par</seg></l></lg>
```

In addition, each French line group (`lg`) has been linked to the original group through the attribute `corresp`. For example, the verses *De la profondeur* and *Noire clame comme une voix entendue par un dormeur* form the group which correspond to the English group identified by the attribute `xml:id="lg5_128_5"`. So, it has the attribute `corresp="#lg5_128_5"`

VI. Current results

All 29 volumes and the index of the review have been scanned, acquired by OCR and manually corrected by students with the supervision of ILC researchers. All bibliographical metadata have been entered on the Zotero platform, in order

[4] https://zotero.org

to be exported in standard formats (e.g. in BibTeX) in the next phases of the project.

At this stage of the work the first 5 volumes have been encoded according to the TEI guidelines, mainly focusing on multilingualism, parallel texts, poetry, spatial distribution of text. In the next future, we plan to finish encoding the remaining volumes. Then, the digital editions produced must be subjected to a final check to identify any inconsistencies and marking errors. Whereupon, we proceed with the generation of e-books in .epub and .mobi format in order to make them readable and navigable on a variety of devices.

We also intend, focusing in particular on multilingualism and translation practices, to study the journal in the cultural context of the time, tracing the different contacts made possible thanks to the journal, as well as the evolution of the texts that appeared for the first time on *Commerce* and then spread on a global scale.

VII. Conclusion

The international and transcultural vocation of the review has made it and still makes it fundamental and necessary for the work of many scholars. *Commerce* is one of the journals that – together with others both from France and other countries – has allowed writers, poets, playwrights and intellectuals to publish their works for a wide audience, often for the first time, and to read texts originally written in other languages.

Commerce has a fundamental importance to reconstruct the cultural landscape of the early 20th century made up of exchanges, contaminations, contacts, circulation of texts, translations and self-translations. The encoding of *Commerce* allows to protect an important literary document and at the same time to make it available for a worldwide audience through fully searchable digital editions.

Next steps are in the direction of the semantic web: one of the main objectives is to make the entire journal navigable both through thematic keywords and semantic relations.

References

[1] A.S. Armani, “Un anneau de corail, lettere di Paul Valéry a Marguerite e Goffredo Caetani”, Roma, Bulzoni Editore, 1986.

[2] E. Conti, “Ungaretti, mediatore culturale di Commerce” in Intersezioni 1/2002, pp. 89-108, Bologna, Edizioni Il Mulino, 2002.

[3] M. Dalmau, and M. Schlosser, ”Challenges of serials text encoding in the spirit of scholarly communication”, Library Hi Tech, Vol. 28 No. 3, pp. 345–359, 2010.

[4] L. Dennett, ”An American Princess. The remarkable life if Marguerite Chapin Caetani”, Montreal & Kingston, London, Chicago, McGill-Queen’s University Press, 2016.

[5] A. Sanna, “Tra modernismo ed europeismo: La Nouvelle Revue Française e Commerce“, in R. Donnarumma, S. Grazzini eds,“La rete dei modernismi europei. Riviste Letterarie e Canone (1918-1940)”, Perugia, Morlacchi Editore, 2016.

[6] F. De Simone, B. Balbi, V. Broscitto, S. Collina, R. Montanari, F. Boschetti and A. Fahad Khan, ”The Impact of Human Factors on Digitization: An Eye-tracking Study of OCR Proofreading Strategies”, Proceedings of COGNITIVE, The Tenth International Conference on Advance Cognitive Technologies and Applications, 2018.

[7] M. Holmes and L. Romary, ”Encoding models for scholarly literature”. Ioannis Iglezakis, Tatiana-Eleni Synodinou, Sarantos Kapidakis. Publishing and digital libraries: Legal and organizational issues, IGI Global, pp. 88-110, 2010.

[8] S. Levie, “Commerce 1924-1932: une revue internationale moderniste”, Roma, Fondazione Camillo Caetani, 1989.

[9] S. Levie, “La rivista Commerce e il ruolo di Marguerite Caetani nella letteratura europea, 1924-1932”, Roma, Fondazione Camillo Caetani, 1985.

[10] S. Levie (edited by), “La rivista Commerce e Marguerite Caetani”, 5 voll., Roma, Edizioni di Storia e Letteratura, 2012-2016.

[11] E. Rabate, ”La Revue Commerce: L’esprit Classique Moderne (1924-1932)”, Paris, Classique Garnier, 2012.

[12] J. Unsworth, “Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI.” Journal of the Text Encoding Initiative 1, 2011.