Commerce Numérique: traffic signals for crossroads between cultures

Ouafae Nahli[1], Antonietta Sanna[2], Michela Bandini[2], and Federico Boschetti[1]
[1]Institute for Computational Linguistic, Italian National Research Council, Pisa, Italy
name.surname@ilc.cnr.it
[2] University of Pisa, Pisa, ITALY
antonietta.sanna@unipi.it, m.bandini9@studenti.unipi.it

***Abstract*—*Commerce* is a French literary journal - founded by Princess Margherita Caetani - which relied on the collaboration of three prestigious writers: Paul Valéry, Léon-Paul Fargue, and Valéry Larbaud. The journal is composed of twenty-nine volumes published between 1924 and 1932. Each volume includes different literary material like poems and novels, written by both well-known and unknown writers, who also translated important authors like Joyce, T.S. Eliot, Pirandello, Ungaretti, Saint-John Perse, Rilke, and Hofmannsthal.**

Considering the historical, literary, and cultural importance of the journal *Commerce*, our project "Commerce numérique" aims to digitize and to make the journal contents freely available online to both the general public and the research community.

This article describes the way in which the journal was encoded. Particular importance is also given to the encoding of poems present in *Commerce*. Some poems are in the original language and are accompanied by their French translation, other poems are in the French-translated form without the original text. In order to fully and accurately express the phenomena and their structures, we adopted some aspects of the TEI framework that will be explained in detail.

The French translation of a Moroccan Arabic poem from the 13th century is also considered. The original Arabic poem is interesting because it presents aspects of both the Moroccan dialect and the oral text. The study and the encoding of the Arabic poem in parallel to its translation highlight some important structural differences between Arabic poetry and Western poetry.

***Index Terms*—*Commerce* Journal, OCR, TEI encoding, literary journal, digital resources, Arabic poetry.**

I. Introduction

Considering the historical, literary, and cultural importance of the *Commerce* journal, the aim of our project *Commerce numérique* is to digitize the journal and to make its contents freely available online to both the general public and the research community. We, therefore, studied an appropriate approach to the encoding of the journal, with particular attention to the characteristics of each volume. The digital format makes it possible to move to other documentary scales, and it also allows to:

- make a document "intelligent" by tagging and enriching the metadata;
- isolate information units that can be used to structure a documentary set and to give it a representation of multiplicity.

This paper is an extension of our conference paper [19], where we described the stages of our work to create a digital version of the *Commerce Journal*. Here, we describe our work in more detail. The current paper underlines the significance of digital representation and of some decisions that we took based on the structure of the journal. It discusses specific issues regarding the approaches to the digitization process, the role of metadata, and the encoding of the articles contained in the journal, respecting their internal physical and semantic structure.

Section II places the journal *Commerce* within the European cultural context at the beginning of the 20th century and provides a descriptive study of *Commerce* with regard to the quantity and content of the articles. Section III describes the stages of digitization, image pre-processing, OCR, and manual correction.

Section IV discusses the encoding criteria that we adopted. In this first phase of the project, we decided to focus our attention on the physical structure of the corpus and the structural and formatting elements of each volume. The main goal of the first phase was the compliance of the digital edition with the typographical aspects of the original printed edition.

Articles vary from different topics and genres, and particular attention was devoted to encode poems. Some poems present in *Commerce* are in original form in another language and are accompanied by their French translation. Other poems are French-translated forms with no original version. Section V describes the TEI encoding of poems in more detail, providing an example for each type of poem. Particular attention was paid to the French translation of a Moroccan Arabic poem. The original Arabic poem is interesting because it presents aspects of the Moroccan dialect and aspects of the oral text. The comparative study and the encoding of the Arabic poem and its translation highlight some important structural differences between Arabic poetry and Western poetry. Both TEI encoding and linguistic analysis present a number of challenges.

In section VI, we present the results that will be made available to the scholarly community and finally a conclusion closes the article.

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

Commerce was a very important literary review founded in Paris by Princess Marguerite Caetani, wife to Prince Roffredo Caetani, after the First World War. Two important additions

from the Parisian literary scene were made in 1914. The first was that of Adrienne Monnier's bookshop, called "La Maison des Amis des Livres", where authors of contemporary French literature started to hold readings in 1917. In the same year Sylvia Beach, a young American woman who had studied modern French literature, set up a bookshop specialized in English and American books. She opened the legendary "Shakespeare and Company", a modern *Salon littéraire*, where many French authors such as André Gide, Paul Valéry, Valéry Larbaud, and Jules Romain met American and English writers like Ernest Hemingway, F. Scott Fitzgerald, Ezra Pound, and 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, and Gaston Gallimard became a model of modern literary review, which published the most important texts of European Modernism. Marguerite Caetani regularly read the NRF that she considered indisputably France's leading intellectual review. She met Paul Valéry before 1914, and probably also Adrienne Monnier, Léon-Paul Fargue, and Valéry Larbaud at the events organized at the "Shakespeare and Company" bookshop. During one of the many Sunday gatherings at Roffredo and Marguerite Caetani's home in Versailles, the idea of founding a review dedicated only to poetry, prose, and drama started to take shape. According to Marguerite, this happened in the following way:

> One day, all at once, Valéry said: "Why don't we continue our conversations, our dialogues, in published form? As a title I suggest 'Commerce', exchange of ideas. Everyone present was delighted at the idea. The directors (Larbaud, Valéry, Fargue) were appointed immediately. Adrienne Monnier and I were entrusted with 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 had received for the NRF as well as by Alexis Léger, who chose the poems that we published [9].

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

Marguerite Caetani was also able to count on the collaboration of some writers and intellectuals from other European nations, like T. S. Eliot, Giuseppe Ungaretti, Rainer Maria Rilke, Hugo von Hofmannstahl, Rudolf Kassner, and D. S. Mirskij [14] [15]. The review included literary works – excerpts from novels, short stories, poems, drama – not only from France but also from other ten different countries.

The most important and peculiar characteristic of *Commerce* was its 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 into French, as Hugo von Hofmannsthal and Giuseppe Ungaretti [6].

This characteristic makes *Commerce* a fundamental review and gives the opportunity to current researchers to better understand the European literary field of the early 20th century, where Paris was the most important city, crossed by cultures and artists from all over the world.

The *Commerce* journal is a collection of a limited number of volumes. The corpus consists of 29 volumes and of a short index of the first 16 volumes, for a total of more than 6,000 pages and about 2,000,000 tokens. Most of the 240 texts which appeared in *Commerce* had never been published before.

Commerce published a combination of contemporary and ancient texts, many of which were published in the review for the first time. Some of the authors who debuted at that time were soon to become some of the most important authors of the 20th century [14]. We can count about 140 French texts, most of which written by Paul Valéry, Leon-Paul Fargue, and Valéry Larbaud, but also by many other important French writers such as Antonin Artaud, André Breton, and Henri Michaux. Thanks to the cooperation of advisors from Germany, the magazine hosted almost 20 German works by authors like Franz Kafka, Friedrich Hölderlin, and Friedrich Nietzsche. More than 20 texts by English authors were published in the review, including some fragments from *Ulysses* by James Joyce and one text from *To the Lighthouse* by Virginia Woolf, which was published for the first time in this 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, written by authors like Søren Kierkegaard, Cheng Tcheng, Boris Pasternak, and Jose Ortega y Gasset.

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

All the volumes provided by the library of the University of Pisa and by the Camillo Caetani Foundation were scanned at the Institute for Computational Linguistics (ILC) of the National Research Council (CNR). The general digitization workflow is simple and it occurs following different phases.

1) After the acquisition process, image processing improves the quality of the digital object [11]. Page images were pre-processed to optimize character recognition by the usual page image operations, such as splitting, deskewing, dewarping, despeckling, and binarization performed by scanTailor[1].
2) The phase of Optical Character Recognition (OCR) permits an electronic conversion of digital images with typed text into the machine-encoded text. Optical Character Recognition (OCR) was performed by Tesseract[2] in a multi-language modality (French, English, Italian, German, and Spanish).
3) Proofreading of the OCR output was performed by using the CoPhiProofReader, a web application for collaborative OCR correction created at ILC-CNR. The

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

CoPhiProofReader is inspired by the WikiSource[3] correction tools, to be able to trace multiple proofreading and supervision phases. As shown in Figure 1, systems for collaborative proofreading are generally based on a comparison of image and text, page by page. However, the comparison line by line between the image box of the original printed edition and the OCR results help students and scholars to recognize the OCR errors, as demonstrated by studies on the ergonomics of proofreading [8].

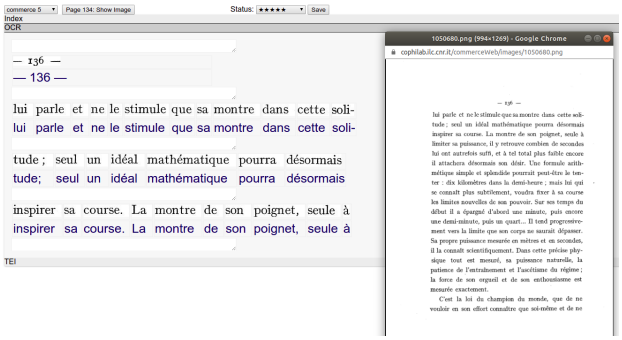

Figure 1: CoPhi ProofReader

IV. TEI-XML ENCODING

A. Encoding criteria

We chose the TEI encoding system because it is a *de facto* standard for the community of Digital Humanities, and it is extremely flexible and inclusive of most text genres. It provides elements, attributes, and other mechanisms for encoding prose, poetry, theatre, dictionaries, linguistic corpora, and other literary and academic texts. Furthermore, TEI offers a framework where it is possible to design specific customizations while remaining compliant with the guidelines. As a result, encoded texts are interoperable and can be reused, managed, and archived [7], [12], [20].

TEI is suitable to encode literary journals: good examples are *The Modernist Journals Project* (searchable database developed by Brown University and the University of Tulsa, ongoing)[4], the journal *Scandinavian-Canadian Studies / Études scandinaves au Canada*[5], and the digitized student magazines of the Victoria University of Wellington[6], in particular, *Arachne: A Literary Journal* and *Hilltop*. Further projects can be found on the TEI website[7].

Considering the limited budget of the first phase of the project and the large number of contents to be encoded, we decided to focus our attention on the minimalist structural and formatting elements of each volume. The main goal of the first phase was the compliance of the digital edition with the typographical aspects of the original printed edition. Given the large number of pages for each volume, we decided to follow a particular order focusing on one aspect at a time. After we had completed the typographical encoding, we proceeded with the encoding of the detailed structure of each volume.

We used generic tags like `<fw>` for illustrations such as asterisks to mark different sections of the text, page numbers, and numbers in the notes. The `@facs` attribute of the `<pb>` element is used to link the transcription to the facsimile of the page. We also used the `<milestone>` tag for page numbering, and we tagged every unusual space within lines in the entire journal using the tag `<space dim="vertical" unit="line" quantity="x"/>`, in which *x* refers to the number of estimated lines corresponding to vertical space. However, great care was taken to render the spatial distribution of the text on the page through encoding. This 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. Within each volume, some words were highlighted as is usually done for verses, quotes, foreign words, or keywords. Yet, in other parts of the journal, it was possible to see entire pages in *italics*, something that had to be taken into consideration given the period of time in which the journal was written. These aspects, alongside the many other typographical differences, show how many authors and different styles influenced this literary journal. Every article of the journal reflects the style of the authors but at the same time, like any other journal, *Commerce* has a specific generic structure that we tried to mark. The formal and typographical aspect is important, especially 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 [18].

Commerce shows many different types of literary material, mainly translated into French, from different periods, and from different non-Indoeuropean languages. Encoding allowed us not only to disseminate literary content that was in large part unknown but also to analyze the main stylistic differences between different authors and mostly between original versions and their translations. Much time was spent analyzing poems of the first 6 volumes, and we realized that the TEI guidelines made it possible to differentiate verses in poems from lines in prose, to mark some of their stylistic and fundamental aspects, and to create a link between internal and external sources.

B. Physical structure of the text

Commerce is historically a print-based journal and, like most printed journals, it uses a traditional hierarchical structure in issues, volumes, and articles. The tags that are used for the encoding have to respect its logical structure. The XML structure of each volume reflects the way in which the paper version of the volume is structured, and it also marks all the typical components of a journal.

All 29 volumes have a title page, an index, the body of the text containing various articles, and a colophon at the end. More specifically, each volume is preceded by a front matter,

[3]https://wikisource.org
[4]https://modjourn.org
[5]https://scancan.net/articles.htm
[6]http://nzetc.victoria.ac.nz/tm/scholarly/tei-corpus-VUWMagazines.html
[7]https://tei-c.org/Activities/Projects/

followed by a back matter, and divided into articles, which are subdivided into sections and paragraphs for prose, or strophes and verses for poetry.

The title page is included in the <front> tag which contains all the prefatory matter. In every volume of *Commerce*, this tag includes everything that goes from the very first page of the volume to the page that marks the beginning of the first article and embeds the two tags <titlePage> and <listBibl>, which correspond to the title page of the volume and the index. The <titlePage> element is composed of three other elements marking a lot of other information like the title of the document, the authors, and the document imprint which is when and where it was published, and by who. For this section, we also focused on only a few metadata aspects considered fundamental, such as dates, names and surnames, addresses, and locations. Each volume of *Commerce* has a page showing the index, for which we took into consideration the title itself - *Sommaire* - encoded with the tag <head>. We also encoded every single title and the authors of the articles listed in the index with <title> and <author> elements. At the very end of each volume, the colophon - a page that usually provides information about the publication of the book - is tagged <back ana="colophon">.

Articles are gathered inside the <group> element, which is designed to simplify the encoding of collections, anthologies, and cyclic works. In addition, the <group> element reflects the potentially complex internal structure of the journal. We encoded each article using the <text> tag that was opened on the first page of the article and that was closed on the last page of the article. We considered as part of an article even the blank page that may appear between two articles, aimed to make things clearer and to include every single page of the journal in a specific article.

In order to distinguish every article from the others, we used specific attributes listed as follows: @xml:id. For example, the listing 1 illustrates the @xml:id of the article *"Extraits de son LOG BOOK"* that is the second article of Volume 6 of the *Commerce* journal, and starts on page 13 and ends with the blank page 26, even if the signature of the author, Edmond Teste, is on page 25.

```
<text xml:id="a06-p13-p26"
     type="article"
     n="EDMOND TESTE,
     Extraits de son LOG BOOK"/>}
```

Listing 1: The @xml:id of *"Extraits de son LOG BOOK"*

In this case, te value of the @xml:id lies on the a06 element indicating that it is an article of the sixth volume, and the number of the first and last page of the article, p13-p26. The @type attribute shows what kind of text it is; indeed the value is article. The @n attribute specifies a label, and its value is composed by the name of the author and the title of the article as it appears in the *Sommaire*.

Articles vary from different topics and genres, and we have devoted particular attention to the coding of poems which we explain in the further section.

V. Encoding of poems

A. General encoding

Poems are composed of verses grouped in strophes according to rhymes or number of verses. The listing 2 shows an example of the encoding of a strophe taken from a poem in *Commerce 6*, page 195. In the first phase of encoding, the poem is divided into strophes (lg: line groups), and the verses are encoded using the <l> element.

The strophe has a specific attribute @xml:id, whose value is composed of: @lg followed by the number of the volume, the page number of the strophe, and the number of the strophe (in chronological order). Finally, it is always mandatory to mark the language of the strophe while encoding, using this specific attribute: @xml:lang, with the language as value. We use the @n attribute to number the verse in chronological order starting from the first verse to the very last one of the poem.

```
<lg xml:id="lg6_195_1"} xml:lang="fra">
  <l rend="hi" n="1">
    Celui qui a baisé le front
    meurtri du temps
  </l>
  <l rend="hi" n="2">
    Avec la tendresse des fils
  </l>
  <l rend="hi" n="3">
    Se rappellera plus tard le temps qui
   s'endormit
  </l>
  <l rend="hi" n="4">
    Dans la couche profonde de blé sous
    la fenêtre
  </l>
</lg>
```

Listing 2: General encoding of poems

There are two different kinds of poems in *Commerce*. Some poems are original in another language and are accompanied by their French translation. In this case, we created a link between the original poem and the corresponding French translation. We used attributes to link each verse from the original version of the poem to each verse of the translated version, as we will show better later in this paper (point *A*, section VI). Other poems are French-translated forms without the original version. We encoded them following the steps shown before, with the generic tags presented. The French translation of an original Arabic poem was particularly interesting and worth further analysis, as we will examine in-depth in point *B* of section VI.

B. Original poems in parallel with their translation

Special attention was devoted to encoding poems in parallel with their translation. Figure 2 illustrates a snippet of the poem *"TRAIN-STOP : NIGHT"* in Volume 5, page 128, translated in French on page 129.

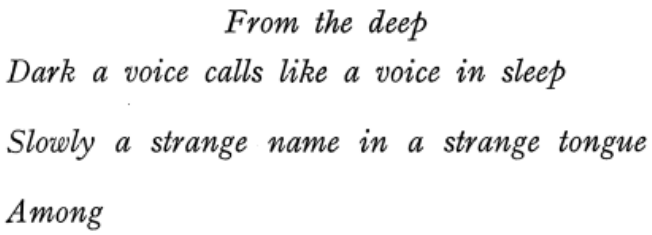

Figure 2: English Poem

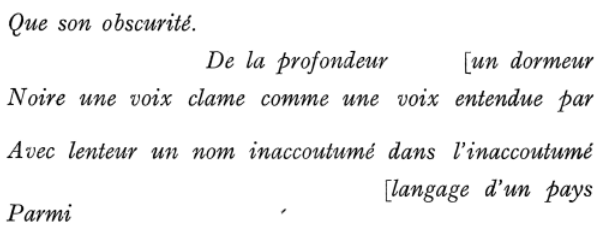

Figure 3: Parallel French Poem

As shown in the listing 3, the strophes in the original language are identified through the attribute `@xml:id`, which provides a unique identifier for the element that bears it. The attribute `@xml:id` permits to link the original poem with the corresponding translation, strophe by strophe. As we have already mentioned, texts and poems are in different languages. For this reason, it is important to specify the language through the attribute `@xml:lang`. In this case, the original language of the 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>
```

Listing 3: Encoding of the original poem

Figure 3 shows the French translation of the poem *"TRAIN-STOP : NIGHT"* in Volume 5, on page 129. It is worth noting that, in some cases, the French counterpart is split into two segments and the second segment is dislocated to the top or bottom row.

For example, the verse *Noire une voix clame comme une voix entendue par un dormeur* is divided into two segments: *Noire ... par* and *un dormeur*. The second segment is presented in the printed edition on the same line of the previous verse but separated by a bracket, which suggests the correct textual order. In order to preserve both the logical order of the segments and the original rendering of the printed edition, segments were disposed in the same order they had in the printed edition but with an ordinal indication `n` between the segments. The listing 4 shows how the segments are represented in order to preserve both the textual and the spatial order, compliant with the printed page.

The line `<l n="9">` is formed by two segments: `<seg n="1">Noire ... par</seg>` and `<seg n="2">un dormeur</seg>`, where the second segment precedes the first, and it is located on the same line of line `<l n="8">`.

Line `<l n="8">` and 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>
```

Listing 4: Encoding of the parallel French poem

In addition, each French line group (`<lg>`) was 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 corresponding to the English group identified by the attribute `xml:id="lg5_128_5"`. Therefore, it has the attribute `corresp="#lg5_128_5"`

C. French translated poem without the original text

Many poems in *Commerce* are French translations from poems in other languages (in *Commerce 6*, for example, we have poems translated from Russian, Arabic, or Turkish), and most of them are provided only in French without the original source.

For example, in the article 7 *"TROIS MYSTIQUES MUSULMANS"* of *Commerce 6*,[8] we only find the French translation of three Arabic poems by Louis Massignon. The first of these poems *"SHAYKH MIN ARDI MIKNAS" ("Shaykh*[9] *of the land of Meknes"*) is a famous Moroccan poem written in the thirteenth century, but it has been a popular song until today.

1) "SHAYKH MIN ARD MIKNAS": The poem describes the itinerant life of the Sufi[10], with a travel bag on his shoulder and dressed in rags. The author, Abū al-ḥasan aš-šuštarī, is a

[8]Commerce, Cahiers trimestriels publiés par les soins de Paul Valéry, Léon-Paul Fargue, Valéry Larbaud, CAHIERS vi, Hiver 1925, pages 151-168.

[9]The word shaykh, or sheik, in translation has taken on the modern negative meaning of despotic "oil sheik." Basically, however, the Arabic word shaykh means an old and wise person [2].

[10]A person is called a Sufi when they practice Sufism that is a mystical dimension of Islam, and emphasizes the inward search for God and shuns materialism.

Muslim writer[11] who describes life in a religious manner. Man abandons all comforts, wanders in the souks singing the praise of Allah living off charity.

The text of the poem is interesting because it presents some aspects of the Moroccan dialect and some aspects of the oral text, such as:

- Vocal communication relies on intonation as a vital element in the production of meaning [10]. Indeed, the original poem *šuwayk-un min 'arḋ-i meknās* has no punctuation marks;
- Oral style is more additive than subordinate. In fact, there are more coordinative elements (like *wa* which means "and" in Arabic) than subordinate elements (like "when", "while") [10]. In our poetry, the verses start with the connective *wa*, which links a verse with the previous one.

2) Stylistic analysis: Classical Arabic poems are characterized by many features. Some of these features are common in poems written in other languages, while others are specific to Arabic poems. They are written as a set of verses. Each verse is divided into hemistichs (i.e. half verses), which are equivalent in length [1]. The poem, *šuwayk-un min 'arḋ-i meknās* is the *zaĝal* type and it is one of the most famous compositions. The *zaĝal* is an ancient form of semi-improvised and semi-sung poem recited in a colloquial dialect that could be easily understood by ordinary people. The *zaĝal* always has an initial refrain of which a typically common form is a rhyming couplet AA, followed by an indefinite number of strophes, each of which contains a string of verses, usually (but not less than) three, rhyming together, yet differing in rhyme from one strophe to the next (BBB, CCC, DDD, etc.), followed by a final element that rhymes with the refrain but reproduces exactly half of the refrain's rhymes [16]. Our poem is composed of 17 lines, four of which are refrains. These refrains are separated between them by four lines; the first refrain is preceded by a line to present the story.

3) Comparative analysis: Fundamental differences were found by comparing the Arabic original version and the French translation of *Commerce*. It should be highlighted that the use of punctuation in the two poems is distinct. The original poem, as stated elsewhere in this paper, was originally a song that was passed on orally, and spoken communication focuses more on intonation. In other words, in a written sentence we can express additional meaning with commas by placing them in different positions, whereas in oral texts the meaning is generally obtained by inflection. The Arabic original poem does not, of course, even bear the marks of punctuation. In the translation, the meaning is obtained by means of punctuation marks. We realized that the translator had decided to step away from the original structure. There are 25 verses in the French version which correspond only to 17 verses in the Arabic poem. Some French verses correspond to a single hemistich of the Arabic poem, and others include the two hemistichs of the original verse. The French translator was not so much interested in the poem being as close as possible to the original one, but rather in creating a perfect poem that could be more adequate to the readers of *Commerce*, who were probably more used to the structure of western poetry, with strophes and verse. Interesting, from a stylistic and semantic point of view, was a more modern English translation of our Arabic poem that we found: *Little shaykh from Meknes* [2], totally different from the French version in *Commerce*. The English translator tried to be as faithful as possible to the structure of the original poem, and indeed the English version does not contain verses but hemistichs.

4) TEI encoding: As we have already said, the structure of Arabic poems is totally different from that of classical ones (for example poems in English or French), because Arabic poems are usually composed by hemistichs. Indeed, when analyzing the French translation of the Arabic poem of *Commerce 6*, and comparing it with its original version, we realized that the French author had in large part changed the structure of the poem and re-adjusted it to typical classical western poems. The French author grouped all the verses in irregular strophes, the number of verses varying from a minimum of 3 to a maximum of 7. Therefore, the encoding of this particular poem in Volume 6 was different from other poems in *Commerce*. As shown in Figure 4, we still used the general tag `<lg>` for strophes and `<l>` for verses with the same attributes mentioned before. We also used the `<lg>` with the `@ana` attribute with value `couplet` to mark the refrain that was repeated 4 times and it characterizes this *zaĝal* poem. We encoded the original Arabic poem as an external source and we linked it to the French translated poem using a `@corresp` attribute. Figure 5 shows part of the Arabic poem encoded in TEI.

5) Linguistic analysis: In order to encode the linguistic analysis of the poem, we used the CoNLL-U format[12] where linguistic annotations are represented in tabular form through plain text files [5]. Figure 6 illustrates part of the translated poem *"SHAYKH MIN ARD MIKNAS"* in CoNLL-U format. The text is subdivided into word lines that register the annotation of a token by means of five fields:

1) ID: words indexed with the identifiers that take into account the physical structure of the poem. For example, the word *cheïkh* with the `ID=f01020` corresponds to the second token of the French poem, first verse;
2) FORM: contains the word form or the punctuation symbol;
3) LEMMA: contains the canonical form of the lexical entry;
4) UPOS: contains the part-of-speech tag of the word belonging to the tagset of the universal dependency grammar;[13]
5) FEATS: contains the list of morphological features belonging to the universal feature inventory;
6) MISC: the last MISC field stores additional information that does not fit into any of the other fields.

[11] aš-šuštarī is an Andalusian mystical poet. He was born in Guadix, around 1203, he lived in Morocco first, then he traveled for a long time in the East, where he died in Damietta in 1269. He left short poems in vulgar dialect, with poignant notation and modern metrics, on the basis of which melodies were immediately innovated.

[12] https://universaldependencies.org/format.html

[13] https://universaldependencies.org/u/pos/index.html

```
<lg xml:lang="fra">
 <l rend="hi" n="1" xml:id="l1.fra">Un cheïkh du pays de Meknès¶</l>
 <l rend="hi" n="2" xml:id="l2.fra">A travers les souks va chantant :¶</l>
 <lg ana="couplet" xml:id="lg.refrain.1">
 <l rend="hi" n="3" xml:id="l3.fra">« Qu'est-ce que me réclament les hommes,¶</l>
 <l rend="hi" n="4" xml:id="l4.fra">Et qu'est-ce que je leur réclame, moi ? »¶</l></lg>
</lg>
<fw type="figure">*¶</fw>
 <lg xml:lang="fra">
 <l rend="hi" n="5" xml:id="l5.fra">« Que dois-je, ami, à toutes les créatures,¶</l>
 <l rend="hi" n="6" xml:id="l6.fra"><seg n="1">Quand Lui, que nous aimons, c'est le Créateur, le Pro-¶</seg>
  <seg n="2" style="text-align:right">[vident¶</seg></l>
```

Figure 4: Encoding of the French translated poem

```
<l n="1"><seg n="1" type="hemistich" corresp="#l1.fra">¶ شَوَيْخ مِنْ أَرْضِ مِكْنَاس</seg>
    <seg n="2" type="hemistich" corresp="#l2.fra">¶ وَسْطَ الأَسْوَاق يُغَنِّي</seg></l>
<l n="2" ana="refrain" corresp="#lg.refrain.1"><seg n="1" type="hemistich" corresp="#l3.fra">
¶ أَشْ عَلَيَّا مِن النَّاس
</seg>
    <seg n="2" type="hemistich" corresp="#l4.fra">¶ وأشْ على النَّاسِ منِّي</seg></l>
<l n="3" corresp="#l5.fra"><seg n="1" type="hemistich">¶ أَشْ عليًّا يا صَاحِبِ</seg>
    <seg n="2" type="hemistich" corresp="#l5.fra">¶ مِن جَمِيع الخَلايقْ </seg></l>
<l n="4"><seg n="1" type="hemistich" corresp="#l6.fra">¶ الذِي هُوَ نَهْوَاه</seg>
    <seg n="2" type="hemistich" corresp="#l6.fra>¶ هُوَ خَالق ورَازِق</seg></l>
<l n="5" corresp="#l7.fra"><seg n="1" type="hemistich">¶ لا تقُلْ يا ابني كِلْمَة</seg>
    <seg n="1" type="hemistich">¶ إلَّا إنْ كنتَ صادِق</seg></l>
<l n="6"><seg n="1" type="hemistich" corresp="#l8.fra">¶ خُذْ كلامِي فِي قَرْطاس</seg>
    <seg n="1" type="hemistich" corresp="#l9.fra">¶ واكتُبُوا حِرْزَ عَنِّي</seg></l>
```

Figure 5: Encoding of the Arabic poem

Figure 7 illustrates part of the Arabic poem in CoNLL-U format. The text is organized using three types of lines:

- word lines that register the annotation of a token by means of five fields represented by single tab characters (see below);
- one blank line that marks hemistich boundaries;
- two blank lines that mark the verse boundaries.

Word lines contain the following fields:

1) ID: words indexed with the identifiers that take into account the physical structure of the poem. For example, the word مِن (*min* "from") with the `ID=a01.1.02.0` corresponds to the second token of the Arabic poem, first verse, first hemistich. Number 0 indicates that the token matches with only one word. The last number of the ID indicates the order number of sub-tokens that make up the token. The word الأَسوَاق, (*al='aswāq* "the souks"), corresponds to the second token of the Arabic poem in the first verse, second hemistich, and it is composed by the two sub-tokens, ال (*al* "the") and أَسوَاق (*'aswāq* "souks"), which have `ID=a01.2.02.1` and `a01.2.02.2` respectively;
2) FORM: contains the word form or punctuation symbol, i.e. the word form in the previous example *al='aswāq*;
3) LEMMA: contains the canonical form of the lexical entry. For example, the lemma of the word form in the previous example *al='aswāq* is the word form in the previous example *s=uq*;
4) UPOS: contains the part-of-speech tag of the word belonging to the tagset of the universal dependency. The lemma *s=uq* is a noun; grammar;[14]
5) FEATS: contains the list of morphological features belonging to the universal feature inventory;
6) MISC: the last MISC field stores additional information that does not fit into any other field. In our case, we inserted the word with its corresponding English lemma, also indicating whether the word was dialectal.

As seen above, the CoNLL-U format requires a tabular disposition of the linguistic analyses (i.e. each token in a row and each feature in a column), whereas XML-TEI documents have a hierarchical structure. The CoNLL-U format can be easily managed in a spreadsheet, but it also allows to map linguistic analyses on an XML-TEI document, in order to reconstruct the hierarchical structure of the poem and its translation, verses, words, and morphemes. To this purpose, a Python script reads the CoNLL-U table related to the Arabic poem and the CoNLL-U table related to the French translation, and it generates a new XML-TEI compliant document that respects the division in verses (`<l>`), hemistichs (`<seg type="hemistich">`), words (`<w>`), and morphemes (`<seg>`).

In addition, by means of a simple XSL stylesheet, the XML-TEI document is transformed into an (X)HTML document, in which the parts of speech extracted from the linguistic

[14]https://universaldependencies.org/u/pos/index.html

```
# global.columns = ID FORM LEMMA POS FEATS MISC
# sent_id = French_Poem
# text = SHAYKH MIN ARD MIKNAS (translation by Louis Massignon).
f01010  Un      Un      DET Definite=Ind|Gender=Masc|Number=Sing|PronType=Art
f01020  cheïkh  cheïkh  NOUN    Gender=Masc|Number=Sing
f01030  du      du      ADP AdpType=Prep
f01040  pays    pays    NOUN    Gender=Masc|Number=Sing
f01050  de      de      ADP AdpType=Prep
f01060  Meknès  Meknès  PROPN   Gender=Fem|Number=Sing


f02010  À       À       ADP AdpType=Prep
f02020  travers travers NOUN    Gender=Masc
f02030  les     le      DET Definite=Def|Number=Plur|PronType=Art
f02040  souks   souk    NOUN    Gender=Masc|Number=Plur
f02050  va      va      VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
f02061  chantant    chanter     VERB    VerbForm=Ger
f02062  _:      _:      PUNCT   _
```

Figure 6: Linguistic analysis of part of the translated poem “SHAYKH MIN ARD MIKNAS”

```
# global.columns = ID FORM LEMMA POS FEATS MISC
# sent_id = Arabic_Poem
# text = شَوَيْخ مِنْ أَرْض مِكْنَاس
a01.1.01.0  شَوَيْخ    شَوَيْخ    NOUN    Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   f01010 f01020
a01.1.02.0  مِن مِنْ ADP AdpType=Prep    f01030
a01.1.03.0  أَرْض أَرْض NOUN    Case=Gen|Definite=Ind|Gender=Fem|Number=Sing    f01040
a01.1.04.0  مِكْنَاس   مِكْنَاس   PROPN   Case=Gen|Definite=Def    f01060

a01.2.01.0  وَسَطَ وَسْطَ ADP AdpType=Prep|Case=Acc   f02010 f02020
a01.2.02.1  ال _ DET f02030
a01.2.02.2  أَسْوَاق   سُوق NOUN    Case=Gen|Definite=Ind|Gender=Fem|Number=Plur    f02040
a01.2.03.0  يُغْنِي    يُغْنِي    VERB    Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Person=3|Tense=Pres|Voice=Act   f02061


a02.1.01.0  أَشْ _ PRON    PronType=Int    f03012 f03013 f03014 f03015 f03020_Dialect
a02.1.02.1  عْلَى عْلِ ADP AdpType=Prep    f03040
a02.1.02.2  يَّا _ PRON    Case=Gen|Gender=Masc|Number=Sing|Person=1|PronType=Prs  f03030
a02.1.03.0  مِن مِن ADP AdpType=Prep
a02.1.04.1  ال _ DET f03050
a02.1.04.2  نَّاس نَّاس NOUN    Case=Gen|Definite=Ind|Gender=Fem|Number=Plur    f03061

a02.2.01.1  و و CCONJ
a02.2.01.2  أَشْ _ PRON    PronType=Int    f04021 f04022 f04023 f04024 f04030_Dialect
a02.2.02.0  عْلَى على ADP AdpType=Prep
a02.2.03.1  ال _ DET
a02.2.03.2  نَّاسِ نَّاس NOUN    Case=Gen|Definite=Ind|Gender=Fem|Number=Plur
a02.2.04.1  مِنْ مِنْ ADP AdpType=Prep
a02.2.04.2  ي ي PRON    Case=Gen|Gender=Masc|Number=Sing|Person=1|PronType=Prs
```

Figure 7: Linguistic analysis of part of the Arabic poem in CoNLL-U format

analyses are exploited to classify the `<span>` of texts and consequently to colour them according to their morphological category, as shown in Figure 8. The linguistic analyses are associated with each word (or morpheme) through the attribute `@ana`, and the correspondence between the original Arabic word (or morpheme) and the French translation is indicated by the `@corresp` attribute. Figure 8 also shows that by moving the mouse over the Arabic words (or morphemes), the corresponding French translation is highlighted (the `mouseover` event is managed through JavaScript[15]).

For example, the morpheme أَسوَاق with `ID=a01.02.02.2` (i.e. the second morpheme of the second token of the second hemistich of the first verse of the Arabic poem) corresponds to morpheme *souks* with `ID=f02040` (the fourth morpheme of the second line of the French translated poem). The listing 5 represents the analysis of the Arabic morpheme.

```
<seg xml:id="a01.02.02.2"
          n="2" type="morpheme" ana="NOUN"
          Case="Gen"|
          Definite="Ind"|
          Gender="Fem"|
          Number="Plur"
          Gloss="souks"
          corresp="\#f02040"
</seg>
```

Listing 5: Analysis of the morpheme أَسوَاق

[15]A demo of the Arabic poem in parallel with the French translation with highlighted correspondences is available at `http://cophilab.ilc.cnr.it/commerce/demo1.html`

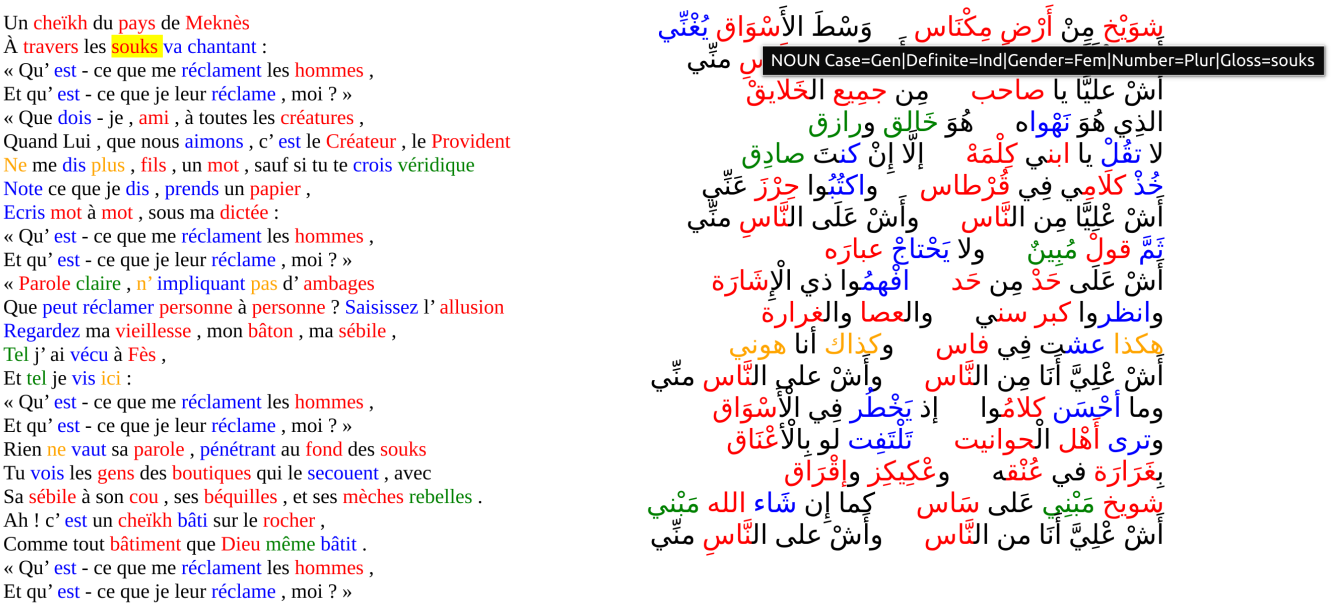

Figure 8: Interactive visualization of correspondences

It is worth noting that the author presents himself as *śuwayh* "little shaykh". At that time he was a well-known orator. This fact denotes a humility that the French translator was unable to catch; or, he may have had a manuscript where there was written *śayh* instead of *śuwayh*.[16]

VI. Current results

All the 29 volumes and the index of the review were scanned, acquired by OCR, and manually corrected by students with the supervision of ILC researchers.

In addition, all the bibliographical references of *Commerce* were recorded on Zotero[17], in order to make them available in different standard formats, such as BibTeX.

At this stage of the work, the first 8 volumes were encoded according to the TEI guidelines, mainly focusing on multilingualism, parallel texts, poetry, and the spatial distribution of the text. In the near future, we intend to finish encoding the remaining volumes. The digital editions produced must then be subjected to a final check to identify any inconsistencies and to mark any errors. Afterward, we will 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. The volumes will be published online at the end of the project.

In particular, we intend to focus on multilingualism and on translation practices, and to study the journal in the cultural context of the time, tracing the different contacts made possible thanks to the journal, and following the development of the texts that appeared for the first time in *Commerce* and then spread on a global scale.

VII. Conclusion

Commerce is one of the journals that has allowed writers, poets, playwrights, and intellectuals to publish their works, often for the first time, for a wide audience, and to read texts originally written in other languages. Thanks to its international and transcultural vocation, the review has been essential and necessary for the work of many scholars. *Commerce* has fundamental importance to reconstructing the cultural landscape of the early 20th century made up of exchanges, contaminations, contacts, circulation of texts, translations, and self-translations.

Considering the historical, literary, and cultural importance of the *Commerce* journal, the aim of our project *Commerce numérique* was to digitize and to make this important literary document available online both to the general public and to the research community through fully-searchable digital editions.

The process that we followed involved the following steps: 1) digital acquisition of the Journal; 2) OCR and manual correction of the text; 3) encoding of the structures, contents and phenomena conveyed by the text.

We used the guidelines of the Text Encoding Initiative (TEI) to prepare a digital representation of the journal. TEI is an XML language designed to suggest standard encoding of textual phenomena.

During the encoding phase, we focused on the physical structure of the corpus and the structural and formatting elements of each volume. We defined the criteria that permitted obtaining a digital edition conformant to the typographical aspects of the original printed edition.

We also focussed on some stylistic and semantic aspects of the poems present in the journal. We described the TEI encoding of the poems in further detail. In particular, we illustrated the French translation of a Moroccan Arabic poem. TEI encoding and the linguistic analysis of the original poem, characterized by both dialectal and oral text aspects, represent a major challenge. We encoded verses, hemistichs, tokens, and other significant features. In addition, linguistic analysis was performed using the CoNLL-U format, which is an optimal

[16]The English translator in [2] translated as "little shaykh".
[17]https://zotero.org

method to facilitate the specialist's work. Subsequently, the analyses were linked to the textual content by exploiting the expressiveness of TEI encoding.

Encoding, linguistic analyses, and parallelism of the two texts allowed us to conduct a comparative study between the original and the translated poem, and to highlight important structural differences between Arabic poetry and Western poetry. The 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. One of the main issues of the TEI encoding is to find the best trade-off between the readability of the encoded document and the overload of information that insists on the annotated text. Thus, thematic keywords and semantic relations will be annotated in stand-off by using Euporia, an annotation tool developed at the CNR-ILC, based on Domain-Specific Languages that optimize the compactness of the annotation, even if they can be easily converted in XML-TEI, as discussed in [4].

REFERENCES

[1] A. Almuhareb, I. Alkharashi, L. AL Saud and H. Altuwaijri. "Recognition of Classical Arabic Poems". Proceedings of the Second Workshop on Computational Linguistics for Literature. Association for Computational Linguistics. Pages 9–16. 2013.

[2] l. M.Alvarez, "Abu al-ḫasan al-šuštarī: Songs of love and devotion". New Jersey: Paulist Press. 2009.

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

[4] L. Bambaci and F. Boschetti, "Encoding the Critical Apparatus by Domain Specific Languages: The Case of the Hebrew Book of Qohelet". Proceedings of AIUCD2020. Pages 7-13, 2020.

[5] S. Buchholz and E. Marsi, "CoNLL-X shared task on Multilingual Dependency Parsing". Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), pages 149–164. New York City. 2006.

[6] E. Conti, "Ungaretti, mediatore culturale di ≪Commerce≫" in Intersezioni 1/2002, pages. 89-108, Bologna, Edizioni Il Mulino. 2002.

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

[8] 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.

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

[10] M. Doss, "Some Remarks on the Oral Factor in Arabic Linguistics". Studia Orientalia Electronica, 75. Pages 49-62. 2014.

[11] M. Georgieva. "Digitization Workflows: Streamlining the Digitization Process and Distinguishing the Peculiarities in Capturing Various Archival Materials". Against the Grain, 31(1). Pages 61-65. 2019

[12] 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. Pages. 88-110. 2010.

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

[14] S. Levie, "La rivista ≪Commerce≫ e il ruolo di Marguerite Caetani nella letteratura europea, 1924-1932", Roma, Fondazione Camillo Caetani. 1985.

[15] S. Levie (edited by), "La rivista ≪Commerce≫ e Marguerite Caetani", 5 voll., Roma, Edizioni di Storia e Letteratura. 2012-2016.

[16] J. T. Monroe, "Which Came First, the Zajal or the Muwaššaa? Some Evidence for the Oral Origins of Hispano-Arabic Strophic Poetry". Oral Tradition, 4/1-2. Pages 38-64. 1989.

[17] E. Rabate, "La Revue Commerce: L'esprit Classique Moderne (1924-1932)", Paris, Classique Garnier. 2012.

[18] 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.

[19] A. Sanna, R. Cinerari, F. Boschetti and O. Nahli, "Digitizing and Encoding a Multilingual Literary Review: Commerce Numérique". 6th IEEE Congress on Information Science and Technology (CiSt). Pages 204-207. 2020.

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