HSS VI.1 (2017)

DOI: 10.1515/hssr -2017-0007

Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions

Ioana Galleron[*]

University of Grenoble-Alpes, France

Abstract

This paper looks at how digital humanities can modify our more traditional understanding and conceptualisation of literary characters. Through the analysis of cast lists from more than 880 French plays from 1630 to 1810, and the "close reading" of some sample texts, it proposes a classification of units of characterisation (caractérisèmes) that can be identified in plays. In the last part, the paper sketches a protocol for the encoding of these characters in a TEI conformant way, and discusses the advantages and the drawbacks of such an endeavour.

Keywords

Digital humanities, Theatrical studies, Characters, TEI encoding, French theatre, 17th century theatre

Since their beginning, digital humanities have fared well in the field of theatrical text analysis, where the possibilities of distant reading and the interest of new visualisations based on the computer analysis of languages and structures have repeatedly been demonstrated (see Carson, 1997; Moretti, 2011; Steggle, 2015 – to cite but a few). However, many other research questions, asking for the possibility to parse large collections of plays and to extract all relevant information, remain to be answered. To give but some examples, it would be extremely valuable to

[*] Litt&Arts, Université de Grenoble-Alpes, CS 40700, 38058 Grenoble CEDEX 9, France, ioana.galleron@gmail.com.

be able to retrace the developments of themes, plots and episodes, gags and spectacular turns of events, and many other features of the like, circulating from a play to another in time as well as in space. This is particularly the case for the study of the French theatre from 1630 to 1810, a period when plays of all genres are particularly numerous, characterized by striking family resemblances and yet individualized through specific traits. Blocks of significance are highly recurrent, giving sometimes the impression of a kaleidoscopic poetics which calls for a computer-assisted approach, in order to tame numbers and diversity, and to better observe regularities and differences.

Amongst the possible applications of such a semi-automated approach, the study of characters seems particularly promising. Dramatists of this period use a quite small and stable set of characters, and borrow characters one to another, customising them for their purposes; one can detect periods when certain types are in fashion ("coquêtes" and "petits-maîtres" from 1680 to 1720, hypocrites after *Tartuffe* and in the 1760s, etc.), and periods of decline. However, the identification of such regularities can be judged as at least incomplete, since related, to date, to the erudition of scholars and to their capacity to remember relevant information gathered through their readings. A digital approach would allow both systematicity and exhaustivity, provided plays exist in a machine-readable format, and they are properly encoded (i. e. marked up with a relevant set of tags). Getting back to the example of a study upon the "Tartuffes" of the period, spotting all relevant characters means that all plays from 1630 to 1810 have been digitally edited, and that during the editing process, the various types of characters - be they main, secondary or hardly perceptible in the background – have been encoded with regards to their moral characteristics. It is clearly not possible to rely on the identification of a character by simply looking at plays alluding to hypocrisy in their title or role names (this is far from covering the whole set of possibilities), and even searching for *hypocrites* and *hypocrisy* in full texts may leave aside interesting results.

Such a more refined digital representation of characters seems feasible. On the one hand, digital editions have now a quite long history and numerous examples of best practices, if not definitive standards: the

XML format has imposed itself as a viable solution, and the consortium for Text Encoding Initiative (TEI) provides guidelines not only for structural, but also for analytical encoding. On the other hand, several attempts, which will be further detailed in the next section, have been made to digitally encode characters, providing ideas and starting points. More generally, after a period of disinterest in characters, these 'existent[s] endowed with anthropomorphic traits' (Prince, 1987: 12) have returned to the lime-light, as the feeling prevails again that "since three or four millenniums […], no great or lasting idea has been communicated without having use of a character, be it energetic as Gilgamesh, passive as Meursault, central as Ulysses or episodic as Octave" (Colonna, 2007:141). If these developments in narratology have not led, to date, to a unified theory about characters (and more largely about narrative/ dramatic concepts), they allow at least to observe general tendencies and to build a tentative model, much needed for dealing with characters in a digital paradigm – where Golem[1] is able of working fast and parse quickly enormous volumes of data, but gets lost (and nasty!) when confronted to ambiguity.

In the following section, this paper will present the model proposed to describe theatrical characters, before applying it to a corpus of French plays from the 17th and 18th centuries so as to build up a typology. In the third section, the paper will propose technical solutions for actually encoding characters in an XML/TEI conformant way. In the final section, I will discuss the advantages and drawbacks of the proposed typology and encoding protocol.

The block units of a character

The starting point is the idea that a theatrical character exists as series of words, phrases and sentences, forming a defined and limited, if scattered and unevenly distributed, part of a literary text. There is no doubt that a character is more than the sum of these elements, for the reader and even more so for the spectator, who watches the impersonation of a character by an actor and is exposed as such to new elements of signification. There is also no doubt that exhaustiveness in the identification of elements contributing to the building of a character is somewhat utopian, because literary works are texts – i. e. complex interwoven objects in which each element is determined in relation to all the others. Still, it is also obvious that some parts of a play are more specifically destined to build images of fictitious beings rather than perform other functions: such are the initial statement about the name of a character, the description it gets in the cast list, portraits by other actors, information that the character gives himself (or herself) about his/ her occupation, temperament and ideas etc. These linguistic units of variable size can be identified in a reasonably consensual way, and carry with them aspects of characterisation that can be called *caractérisèmes*, by parallelism with so many *èmes* used in linguistic and literary studies. The assumption on which this paper reposes is that *caractérisèmes* constitute a finite list, likely to be grouped in (quite consensual) classes, and supporting the definition of TEI-like attributes and values.

The idea that the main frame of a character is built from an addition of various linguistic units is not new, but there is little literature reporting on attempts to systematically identify, characterise and observe the distribution of these elements and of their values in a particular literary field. So far, studies have taken a rather holistic approach, concentrating on the processes through which a character appears and persists in the reader's mind, analysing his/her function in the plot, discussing his/her significance, but paying little attention to the practicalities of the construction of such a being.

This gap has been addressed, to a certain extent, by a PhD dissertation dedicated to a formal description of literary characters using an ontology (Zöllner-Weber, 2008)[2], followed by some further developments (Zöllner-Weber, 2011). However, while this approach is in many aspects much larger than what will be attempted here, the proposed classes do not really work for many pieces of information delivered in the dramatic texts. The social position of a character (peasant or duke, doctor or king, etc.) is neither an "inside" nor an "outward" feature, as defined in Zöllner-Weber, 2008; also, the homogeneity of the "example classes" proposed in this PhD thesis is open to debate, as well as the methodology for generating the above mentioned classes.

In France, a research project conducted at Toulouse and aiming to build tools for video captures and computer-aided reading of plays, did dedicate some thought to the "digital tagging of theatrical characters" (Golopentia, 2010). Unaware of or refusing mark-up already in use in the TEI community, the members of this project took the decision to generate new sets of tags from scratch. This operation seems to be still work in progress, and it has not given place to actual encodings of characters.

Therefore, it appears that none of these initiatives has managed to establish an international standard for character encoding. This remains to be defined and negotiated, on the basis of what seems to be a shared view about how a character is built (using "units of characterisation"). In order to advance towards such a standard, a double approach has been taken here, mixing a literature review with the pragmatic analysis of electronic material at hand.

Characters in French drama: building a typology

The corpus under study is formed of 882 plays, written by approximately 217 French dramatists from 1630 to 1810. They all have been transcribed in an XML format, and enriched with more or less TEI conformant tags, either by Paul Fièvre[3], by members of OBVIL[4], or by myself[5]; all these plays have been put together on GitHub, via an organisation called Dramacode.

Confronted to the total number of plays written during this period, estimated at some 12,000 units, 882 plays may be considered as representing quite a low percentage (7.35%). However, the sample is much larger than those usually taken into account in traditional studies of literary history. It contains almost complete works by the major authors of the period (Racine, Corneille, Molière, Marivaux…), but leaves enough room for plays produced by second or third-rate dramatists. Also, it covers virtually all types of plays specific to the *Comédie-Française* (tragedy and tragi-comedies, comedies in five, three or one act, both in prose and in verse, pastorals, divertissements, etc.), but does not neglect those staged on the Italian scene in Paris, or during fairs, and, to a certain extent, in the private theatres (*théâtres de société*) that the 18th century people were fond of. This diversity of plays allows us to consider that the main French practices in terms of character building have been covered; in the meantime, analysis must be conducted with the awareness that important rhetoric and poetic differences exist from one genre to another, not to mention from one period to another or from one author to another; however, as the goal was to establish an as large classification of caractérisèmes as possible, it is not entirely illegitimate to put, so to say, everybody in the same boat, and to make no distinction between tragic, comic or operatic characters, to mention but these.

Any quick observation reveals that, even when concentrating only on explicit property ascription, as opposed to more loosely character-related information (Margolin, 2005:146), elements of characterisation appear in both types of discourse that are to be found in a play, meaning the speeches of various characters, supposed to have pragmatic consequences in the fictitious world of the play, and the stage directions of various nature, forming a kind of *dialogue* between the play writer and readers: cast lists, indications of turns of speech, setting, stage movement, and positioning precisions spread all along the play... Some of these elements are already, and very conveniently, pointed out by the TEI recommendations: the name of a character is isolated through the tag <role> as well as his/ her initial description through <roleDesc>. Others are quite invisible: there is no specific tag for identifying a portrait or isolating a set of adjectives referring to a specific *character*. They call, therefore, for a manual analysis, and for a creative use of analytical tags recommended by the TEI.

Considering this variable visibility of the characterisation elements in our texts, but also the amount of material to be treated, a combination of quantitative and qualitative approaches has been designed. An extraction of all role names (tag <role>) was performed, using an XQuery under BaseX; this was initially conducted in order to have an idea about numbers (how many characters in the considered set of plays, median and average number of characters per play, etc.), but it quickly appeared that statistics were less interesting – and more problematic considering the corpus structure – than the semantic information carried by this element. BaseX was therefore used to generate a list of role names, rather than to count values.

A second list was established, formed by the initial character descriptions to be found in the plays under analysis (<roleDesc>); this appeared, however, more complicated to deal with, as it presented some annoying holes. Indeed, in some cases a description such as *"Lysimon, ancien Ami de Pyrante"* was encoded as:

```
    <role>Lysimon</role><pc>,<pc/>
<roleDesc>ancien <choice><orig>Ami</orig><reg>ami</reg></choice>
    de <name>Pyrante</name><pc>.</pc></roleDesc>
```

In these cases, the XQuery extraction //*:roleDesc/text() retrieved *"ancien de"*; it is, of course, easy to recover the full description by adding the missing nodes in the query, but as texts have been encoded by at least 10 different people, with various points of interest, it was unsure what kind of supplementary nodes, sometimes not TEI conformant, one could meet in <roleDesc>. The decision was taken to work with a smaller but more trustworthy list built on the basis of the 35 plays by Louis de Boissy, to be found in the initial corpus. Even so, the list required some cleaning, performed in parallel with the update of orthography.

It is worth explaining that the choice of Boissy's plays for building the second list is, maybe oddly, justified by the very mediocrity of this play-writer. He produced as much for *Comédie-Française* as for *Comédie-Italienne*, but also for the popular theatres of *Foire*. He was unable to break away from or to challenge Molière's heritage, while being attentive to all the new currents traversing the French comedy. Because of these characteristics, his use of role descriptions can be deemed as quite representative not only of his era, but also of the conventions characterising the French classical period in general.

As noted above, many characterisation elements require manual analysis, impossible to conduct by a single person on the whole set of 882 plays. Four texts have therefore been selected, consisting of two comedies (*Le Tartuffe* and *Le Misanthrope* by Molière) and two tragedies (*Phèdre* by Racine and *Zaïre* by Voltaire). This selection can be criticised as being too small, unbalanced in favour of the 17th century, and too concentrated on two major genres and three well-known authors from the classic era, to the detriment of smaller, alternative and possible more popular plays, but also of other important writers such as Marivaux (for comedy) and Corneille (for tragedy), to quote but these two. Its objective was not, however, to draw definitive conclusions, but to get an overview about elements of characterisation to be found in the speeches, and about the specific classification problems they pose. It was also meant to build a comparison set, allowing us to know what caractérisèmes have been underestimated through the analysis of the role names and role descriptions.

Because the list of role names plays such an important role, it is necessary to give a more detailed description of the steps taken in order to analyse it. The first extraction brought 9,473 distinct characters, many of which shared the same name. While this similarity cannot be considered, of course, as a reliable indication of any character likeness,[6] it was of little interest, for the purpose of this study, to keep all occurrences of the same form in the list. A second extraction was therefore performed, using the <distinct-values> function under BaseX; this reduced the list by some 30% (over 6,500 results). Many of these were still redundant, however; because of differences in naming, in transcription and encoding practices between the various members of Dramacode, but also because of human mistakes, the same name could appear under more than ten variants, as it can be observed in the following example:

ERASTE
ÉRASTE
ÉRASTE,
ÉRASTE.
Éraste
Eraste
Eraste,
Eraste.
éraste
éraste,
ÉRAste
etc.

Whatever the typographical and orthographical choices, *the name "Éraste"* is still a single piece of information with regard to character building: the name is a mimetic, traditional and minimalist element, opening a horizon of possibilities to be further filled in by the development of the play. It was therefore legitimate to reduce the various occurrences to a conventional one, as the machine had already done for identical strings. This operation was performed manually, arriving at some 5,000 *distinct* role names. Once again, this does not mean that all Éraste in the plays have been reduced to one, but that the various typographies of "Éraste" have been considered as bringing only one piece of information (that the character has a name).

Simultaneously with this cleaning operation, and using the same excel spread sheet, an analysis was carried out of the information given about characters by their role names. To any scholar familiar with the French plays from the considered era, *simple* first names can tell a lot, as shown by the example quoted in note 6. But even without mobilising this specific knowledge it can be observed that "*Colinette*" brings geographical information (she is French) as opposed to "*Érigone*" and to "*Alzire*", who are *foreign* names or to be considered as such from the outset. "*Colinette*" brings also an indication to the genre (and maybe to the age and social extraction) of the character; in a similar way, "*Jupiter*" and "*Hermes*" give indications about the ontological status of characters, a detail of some importance when contrasted with "*Amphitryon*" or "*Sosie*". When reasoning from an encoding point of view, it is debatable that such information should systematically be tagged on the basis of the role name only: *Albione* is Roman, as the name suggests and the play confirms, but as the role description leaves this aspect implicit, it may be wiser not to "add" it, so to say, to the character at this point of the text. However, this kind of decision is to be taken in a second phase, following consultation with the scholarly community.

If first names represent an important percentage of the list, in numerous cases designation is carried in different, and sometimes more elaborated, ways. Writers may prefer a combination of first and last names, or of a civility appellation with a (first) name *("Madame Pernelle", "Monsieur Damis")*. For Hispanic characters, the triplet civility + first name + last name is in order *("Dom César d'Avalos")*. In other cases, a short role description is adjoined to the name, as in *"Blanchet le jeune"* (young Blanchet), *"L'Ombre de Clitemnestre"* (the Spectre of Clitemnestra) or *"Le vieil Horace"* (old Horatius). Last but not least, many characters do not have a name: *"la femme du procureur"* (the wife of the attorney), *"six* *violons"* (six violin players), *"les sept arts libéraux"* (the seven liberal arts) are such examples. In all these situations, information about age, social status, ontological nature, or even the moral particularities of the character ("la coquette Célimène" – Celimene the coquette) is less ambiguous than when dealing with *simple* names, and it proves the utility of concentrating on the <role> tag for a first analysis of character construction.

The perusal of the above-mentioned 5,000 role names, of different degrees of semantic richness, allows us to build a first, unordered, list of caractérisèmes. In the <castList>, a <castItem> can be defined by:

- a name;
- a profession (attorney, peasant, footman…);
- a social origin or status (marquis, count, prince, staff…);
- a geographical origin (Gascon; chief of Janissaries);
- a collective or individual aspect (choir, band of, "Calotins and Calotines");
- a role in the play (the lover);
- a linguistic/ dramatic behaviour (mute notary, group of dancing fauns);
- an age;
- moral particularities (a "précieuse")
- a religion (Muslim troop);
- political opinions (the conspirators).

The comparison of this list with what can be observed by scrutinizing role descriptions (text delimited with <roleDesc> tag) in Boissy brings no new aspects, with the notable exception of those related to the costumes and to some cases of double identity. Characters can act in disguise: lovers dress themselves as maidens to be closer of their mistresses, husbands may appear, for some reasons, as friends or brothers to their wives, Arlequin enacts an allegorical being *("le Je-ne-sais-quoi"* - I don't know what), a court buffoon, or a mythical band leader (Corésus). Role descriptions bring more systematic information about family positions, or more generally about the social relations between characters (friendship, sharing of special interests, such as fondness for a special kind of music, etc.). They can also be more specific about the roles they will play: Boissy specifies quite systematically the characters in romantic rivalry. On the whole, they do not radically change the perspective upon the already observed types of caractérisèmes, which appear to support the three major aspects of any character as identified by James Phelan, i. e. the mimetic, the thematic and the synthetic component.[7]

Do other aspects appear by using the third, qualitative, approach, consisting in the manual identification of elements of characterisation in four specific plays? This does not seem to be the case in the four plays studied to date. In the following speech by Molière – a randomly chosen example -, linguistic units contributing to the building of characters (underlined in the text) carry information that can be easily distributed in the above observed classes. Unsurprisingly, moral particularities are more frequent than observations pertaining to other categories:

DORINE.
Oh vraiment, tout cela n'est rien au prix du fils;
Et si vous l'aviez vu, vous diriez, c'est bien pis.
Nos troubles l'avaient mis sur le pied d'homme sage,
Et pour servir son Prince, il montra du courage:
Mais il est devenu comme un homme hébété,
Depuis que de Tartuffe on le voit entêté.
Il l'appelle son frère, et l'aime dans son âme
Cent fois plus qu'il ne fait mère, fils, fille, et femme.
C'est de tous ses secrets l'unique confident,
Et de ses actions le directeur prudent.
Il le choie, il l'embrasse ; et pour une maîtresse,
On ne saurait, je pense, avoir plus de tendresse.
À table, au plus haut bout, il veut qu'il soit assis,
Avec joie il l'y voit manger autant que six;
Les bons morceaux de tout, il fait qu'on les lui cède;
Et s'il vient à roter, il lui dit, Dieu vous aide.
(*C'est une servante qui parle.*)
Enfin il en est fou; c'est sont tout, son héros;
Il l'admire à tous coups, le cite à tout propos;

Ses moindres actions lui semblent des miracles,
Et tous les mots qu'il dit, sont pour lui des oracles.
Lui qui connaît sa dupe, et qui veut en jouir,
Par cent dehors fardés, a l'art de l'éblouir;
Son cagotisme en tire à toute heure des sommes,
Et prend droit de gloser sur tous tant que nous sommes.
Il n'est pas jusqu'au fat, qui lui sert de garçon,
Qui ne se mêle aussi de nous faire leçon.
Il vient nous sermonner avec des yeux farouches,
Et jeter nos rubans, notre rouge, et nos mouches.
Le traître, l'autre jour, nous rompit de ses mains,
Un mouchoir qu'il trouva dans une Fleur des Saints;
Disant que nous mêlions, par un crime effroyable,
Avec la sainteté, les parures du diable.

However, this example (supported by many others in the analysed plays) shows that caractérisèmes in the text are more complicated to deal with. Consensus about what constitutes a unit of characterisation would probably be more difficult to reach: should one underline "c'est son tout, son héros" as information about Orgon's tendency to exaggerate, or is it just a hyperbola used by Dorine to express her frustration with the place Tartuffe occupies in the family? Also, should one encode one or two pieces of information in "directeur prudent" (a social status and a moral characteristic feature, or just a social status)? It is also clear that in such cases a technical solution must be found so as to link characterisation to the characterised person (one can see that in the same discourse Dorine portraits not only Tartuffe, but also her master, and Tartuffe's servant): this complicates the encoding, which is already to be expected as creating a huge workload, considering the number of caractérisèmes to be found already in such a short part of the play.

While these questions can be left aside for a start, so as to concentrate on the encoding of the cast lists, it is clear that they ask for further consideration, combining linguistic knowledge about acts of speech and literary analysis. Indeed, already during the classic era many writers, such as Beaumarchais, exploit more largely the possibilities of characterisation from the very onset, while contemporaneous or post-modern plays offer less informative, more jocular role descriptions. Role names and role description may raise, therefore, similar problems as those related to the identification and the tagging of linguistic units of characterisation spread all along the text.

The analysis conducted with the methods described above allows us to propose a classification of caractérisèmes in two main categories, with several classes and even subclasses regrouping a limited or unlimited number of values.

The first category brings together aspects that tend to assimilate the character to a *real* person, apt to exist in the day-to-day life or in a magic/ fantastic world substituted, whether by deep conviction or in jest, by the plain reality we experience. The second looks at the character as an artificial construction in a literary text, destined to fulfil certain functions, often in accordance with a cultural tradition. This covers the above-mentioned *mimetic* and *synthetic* aspects of a character. A question to be further discussed is if a third category, covering the *thematic components,* should be created in addition. It is, however, to be expected that most of this thematic information will be already encoded as mimetic or non-mimetic, even if one can observe that the illusion of a human being created through the assignation of the trait of '*coquetry'* to a character is rather thin, and justifies the classification of such a linguistic unit found in a <roleDesc> amongst *thematic* units rather than amongst *mimetic* ones.

Four main classes contribute to the building of the mimetic effect in a character:

1° the ontological status, with seven proposed values: human, deity, semi-deity, animal, plant, animated object, ghost or phantom;

2° the name (which can be simple or any combination of a civility, first and last name);

3° the social status. Cast lists as well as linguistic characterisation units in text can bring information about:

a) the profession of a character. This is an open list, for the moment, as information about profession comes clearly more often in the text of the play than in the cast list. Situations have been spotted when the profession stated in the cast list is completed (or contradicted) by other occupations the character may have had or still develop. Manservants seem, in particular, to try several types of jobs before taking that of a valet, or in parallel with that.

b) his/ her family position. To date, fifteen possible values have been spotted through role description:

- father (FPF);
- mother (FPM);
- grandmother (FPGM);
- grand-father (FPGF)
- step-father or tutor (FPT);
- step-mother (FPSM);
- daughter (FPD);
- son (FPSON);
- uncle (FPU);
- aunt (FPA);
- cousin (FPC);
- wife (FPWIF);
- husband (FPH);
- widow (FPWID);
- friend (FPFR);
- neighbour (FPN).

A character may cumulate several family positions: Orgon is father to Damis and Mariane, husband to Elmire, son to M^{me} Pernelle, friend to Tartuffe. This last example shows that para-familial positions should be included as well in this category. The encoding solutions must therefore make provision for an additionable mark-up in this class.

c) his/ her geographical origin. The list of values is, again, open, even if recurrences are visible: many characters are Gascons, Normands, but also Swiss or Germans. A problematic designation is that of a character as a *Muslim*, which gives certainly information about the religion, but is also meant, in French plays, to imply an oriental origin.

d) his/her religion. Muslim or Jewish seem the most frequent values, but the list remains open. A question to be asked here is if we must converge towards a standardised taxonomy of religions (but which one?), as the above mentioned *Muslims*, for instance, are designated through a variety of appellations (eg. "*les Ismaélites*" - the Ishmaelites), or if such a normalised typology is of no use.

e) his/ her political opinions. They appear especially in plays written during the French Revolution, and have a limited number of expressions, but more plays need to be encoded before deciding if strict value definitions are to be given, or if this must remain an open category.

f) his or her position in society. For the considered period, one can class quite rigorously characters as pertaining to the upper or the lower class ("*personne de qualité*" or not), but we could also resort to the traditional classification in *"états"* (nobility, clergy, the third state). In many cases, this category is correlated to that of the profession, as usually characters are defined by their social status ('*marquis', 'baron'*) when they do not have a profession, and vice-versa. One finds, however, several cases when characters have both a social status and a profession: counts acting as ministers, for instance. The safest solution is probably to imagine ways for indicating that certain character definitions contain information about his/ her social position, without generating a closed list of values.

4° the personality details, with four other subclasses:

a) age, with five possible values: child, young, adult, old, unknown.

b) sex. Four values have been observed to date in our corpus: male, female, unknown (especially for collective characters: are "*habitants de l'île*" - inhabitants of the island - to be considered as exclusively masculine?), and indeterminate (is Polichinelle a male?).

c) physical particularities;

d) moral particularities.

According to the above-mentioned definition, the category of synthetic features covers information about traditional traits, as well as about the part a character plays in the dramatic text or performance. On the basis of the observed corpus, five classes can be defined, grouping textual clues about:

1° the kind of impersonation a character asks for: individual or collective;

2° how a character is called to perform: there are speaking, singing, dancing or mute characters;

3° the costume a character wears. This may be a surprising class under this heading, as costumes may be deemed to support the mimetic effect, through the illusion about the social or geographic extraction of a character, or about his/ her profession. Yet in the considered plays, indications about costumes are scarce and when they do appear, they rather help relating the character to a type; this is probably the case for most of the plays written within a non-realistic aesthetic horizon. The position of this class remains however debatable.

4° the kind of role it will have in the play: lover or rival, character representing the authority, or in disguise, confident, etc. Here, the main question is if we should replace these values, as stated in the text, by those established by the actantial theory (subject vs. object, adjuvant vs. opponent, etc.).

5° the type to which it appertains. In one of the Boissy's plays, the character of '*Je-ne-sais-quoi*' is indicated to be impersonated by *Arlequin*; this is sufficient to create expectations about costume and acting, or about the reactions the character will have. All Italian role names, and also some of the French ones (e.g. Crispin) go far beyond naming and they connote physical and moral particularities, as well as dramatic scripts. This is also the case for the *personae* intervening in the Roman plays, as well as for the Greek *prosopon*, for which details about the masks and the make-up, the costume and the gestures, the voice and the relationships with the other *characters* point to a whole.

Encoding the characters: some technical proposals

How to encode this typology? TEI offers detailed recommendations about name tagging, but gives no counsels about how to handle the other classes. The proposal is to resort to the *feature structure* declaration. Two main libraries can be created:

```
<fsDecl type="mimetic_characterisems">
<fsDecl type="synthetic_characterisems">
```

Within each library, one can declare as many feature names as the above-mentioned classes, with a precise or an open number of values; precise values will provide @ana attributes for the encoding process, while in other cases a string-value will be tagged as such in text. This is an example concerning the age of the characters:

```
<fDecl name="age">
  <vRange>
    <vAlt>
      <symbol value="young" xml:id="AGY"/>
      <symbol value="old" xml:id="AGO"/>
      <symbol value="child" xml:id="AGC"/>
      <symbol value="adult" xml:id="AGA"/>
      <symbol value="unknown" xml:id="AGU"/>
    </vAlt>
  </vRange>
</fDecl>
```

Feature elements have the advantage of being provided with a <resp> attribute, allowing us to indicate when a speaker does not assume the full responsibility of a characterization, even if this is not precisely the type of use for which <resp> has been imagined. Similarly, <cert> could be employed to indicate if the statement is an objective, an ironic or a *sentimental* one – unless this were too much of a tag abuse. Last but not least, the <corresp> attribute can link characterization and the characterized, when units of characterization are to be found in speeches (the link is less ambiguous in the cast lists).

Getting back to the feature names, the choice to proceed with precise values has numerous advantages, but it also poses several problems.

Should the sex of characters be encoded by using an intuitive list of values, as proposed above, or should we follow the ISO norm in the domain? Also, the proposed age values will appear difficult to assign in some situations: is a 22-year old widow young, or adult? At what age does an adult become old, and a child young? Should a value "precise_age" be added to the list of attributes, so as to cope with the situation when a character is attributed a more or less precise age *("ce sont soixante ans qui épousent vingt")*?

The social status and the personality details raise a specific problem, because the hierarchical levels provided by the feature structure declaration appear insufficient. Indeed, if the six sub-classes listed as describing the social status of a character are declared as values, there is no room for further refining the types of family position, for instance. The solution, to date, is to declare subclasses as features:

```
<fDecl name="family_position">
<fDecl name="religion">
<fDecl name="profession">
etc.
```

However, this means losing the information about the intermediary level of conceptualisation.

Discussion and conclusions

Rather than pushing further the analysis in this paper, the questions raised above need to be answered through scholarly informed discussion and pondering of advantages and disadvantages, and this can be considered as one of the aims of this paper. Based on a typology of caractérisèmes in the French classical plays, the encoding practices proposed here are limited by their national bias. Tested on the cast lists from a small amount of plays, the described tags and attributes have been found quite satisfactory, in spite of the limitations briefly described at the end of the previous section. One of their major advantages, as compared to other similar initiatives, is that they are TEI compatible. It is, however, to be expected that the proposed libraries of features will have to evolve when confronted with the entire corpus available on Dramacode, and especially with the units of characterization observable in speeches.

A question of particular interest is to what extent they can work for plays written in other contexts than the French Ancien Régime, and in other languages than French. Considering the cultural exchanges between the various European countries during the two *classical* centuries (17th and 18th), as well as the common roots of the Occidental theatre, it is not unreasonable to hope that they will prove of larger applicability and validity. This contribution seeks only to push further the reflection about character encoding, universality, robustness and adaptability remaining the major challenges of any encoding endeavour.

A richer encoding of theatrical texts may be a Sisyphean task, but it may also trigger more interest for the digital humanities amongst traditional scholars. Much time could be spared, and more accurate readings could be conducted, if semantic elements of theatrical texts were tagged in a way or another, especially in an era when crowd-sourced projects start to develop not only for the electronic transliteration of paper-written sources, but also for the interpretation of texts (see Dobson et al., 2015). Literary studies have much to expect from distant-reading, but to fully benefit from this they may need first a lot of material to be close-read and tagged; in a field developing fast, this means accepting a paradigm of *slower digital humanities.*

References

Béhar, H. (1996). *La Littérature et son Golem*, Paris: Honoré Champion.

Béhar, H. (2010). *La Littérature et son Golem*, tome 2, Paris: Classiques Ganier.

Brown, M., Dobson, T., Grue, D., Ruecker, S. (2013). “Challenging New Views on Familiar Plotlines: A Discussion of the Use of XML in the Development of a Scholarly tool for Literary Pedagogy”. *Literary and Linguistic Computing.* 28(2). 199-208. doi: 10.1093/llc/fqt016.

Carson, C. (1997). “Drama and Theatre Studies in the Multimedia Age: Reviewing the Situation”, *Literary and Linguistic Computing*, 12, 4. 269-275

Colonna, V. (2007). “A quoi sert un personnage”. *La Fabrique du personnage.* Edited by Françoise Lavocat, Claude Murcia and Régis Salado, Paris: Honoré Champion éditeur. 141-158

Dobson, T., Ruecker, S., Brown, M., Grue, D. (2015) “Neither Bicycles Nor Sheep. Crowdsourcing Semantic Encoding for Elements of Plot”.

https://www.id.iit.edu/wp-content/uploads/2015/06/Neither-Bicycles-Nor-Sheep-Crowdsourcing-Semantic-Encoding-for-Elements-of-Plot.pdf

Elson, D. K., McKeown, K. R. (2009). "A Tool for Deep Semantic Encoding of Narrative Texts". *Proceedings of the ACL-IJCNLP. Software demonstrations*. 9-12

Golopentia, S. (2010). "Le personnage dans DRAMA". *La Notation informatique du personnage théâtral.* Edited by Monique Martinez Thomas. Carnières-Morlanwelz: Lansman Editeur. 95-127

Jouve, V. (1992). "Pour une analyse de l'effet-personnage". *Littérature*, 85. 103-111. doi : 10.3406/litt.1992.260.

Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S., Peinado, F. (2010). "Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case". *Proceedings of the Seventh International conference on Language Resources and Evaluation*. Valetta: European Language Resources Association (ELRA). 1996-2001

Ma, Y., Audibert, L., Nazarenko, A. (2009). "Ontologies étendues pour l'annotation sémantique". *20es Journées Francophones d'Ingénierie des Connaissances*. Tunisie. 205-216, hal-00378594.

Margolin, U. (2005). "Character". *Routledge Encyclopaedia of Narrative Theory*. Edited by David Herlan, D., Manfred J., Ryan, M.-L. London and New York: Routledge. 143-151

Moretti, F. (2011). "Network theory, plot analysis". *New Left Review*. 68 (March-April). 80-102.

Phelan, J. (1989). *Reading people, reading plots. Character, Progression and the Interpretation of Narratives*. Chicago and London: The University of Chicago Press.

Prince, G. (1987). *Dictionary of Narratology*. Lincoln: University of Nebraska Press.

Steggle, M. (2015). *Digital Humanities and the Lost Drama of Early Modern England: ten case studies*. London: Ashgate.

Zöllner-Weber, A. (2009). *Noctua literaria – A Computer Aided Approach for the Formal Description of Literary Characters Using an Ontology*. PhD Thesis, Universität Bielefeld. urn:nbn:de:hbz:36113097.

Zöllner-Weber, A. (2011). "Text Encoding and Ontology - Enlarging an Ontology by Semi-automatic Generated Instances". *Literary and Linguistic Computing* 26(3). 365-370.

[1] This is how Pr. Henri Béhar calls the computers employed to assist linguistic and literary studies (Béhar, 1996 ; Béhar, 2010).

[2] For a theoretical reflection upon the problems of encoding a text on the basis of ontology, see Ma et al., 2009.

[3] See his website, http://theatre-classique.fr/.

[4] Observatoire de la vie littéraire, more particularly the team involved in "Projet Molière", http://obvil.paris-sorbonne.fr/projets/projet-moliere.

[5] See the integrality of the plays by Louis de Boissy (1694-1758), http://www.licorn-research.fr/Boissy.html.

[6] *Even though one can consider, for example, that many Angéliques from our corpus share the same specificities, such as youth, beauty, marital status - single, and probably sweetness and timidity.*

[7] The 'synthetic component' is maybe somewhat clearer designated as *"effet-personnel"* – the staff effect - in Jouve, 1992.

Biographical note.

Ioana Galleron is a PhD in French languages and literatures, specialised in the study of drama from the 17th and 18th centuries, and of the literature of worldliness. Her latest book is dedicated to the *Comédie de moeurs sous l'Ancien régime*. She is also engaged in several digital or traditional editions (works by Michel Baron, Ph. Néricault Destouches, etc.), as well as in projects dedicated to the understanding of quality representations in SSH research.