Enrichment of Renaissance Texts with Proper Names

UDC 81'322.2:004.9

SUMMARY: The aim of the Renom project was to enrich Renaissance texts with proper names. These texts present two challenges: they exhibit great diversity due to various spellings of words and are overladen with numerous XML-TEI tags introduced to save the exact format of the original edition. The task consisted of adding Named Entity tags to this format by tagging names, that had not been already tagged, and their left, and sometimes right, context when appropriate. In order to achieve this, we have improved free, open source program CasSys to parse texts with Unitex graph cascades and we have built specific dictionaries and cascades. The evaluation showed that the slot error rate of name tagging was 6.1%. Renaissance texts enriched in this way are used in a website that unites Humanities and tourism by allowing visitors to navigate maps with names.

KEYWORDS: Named entities, Renaissance texts, Graph cascades, CasSys, Humanities and tourism

Denis Maurel
denis.maurel@univ-tours.fr,iris.
Université François-Rabelais de Tours, Laboratoire d'informatique, EA 6300

Nathalie Friburger
nathalie.friburger@univ-tours.fr
Université François-Rabelais de Tours, Laboratoire d'informatique, EA 6300

Iris Eshkol-Taravella
eshkol@univ-orleans.fr
Université François-Rabelais de Tours, Laboratoire d'informatique, EA 6300

DATE OF SUBMISSION: 4 June 2014
DATE OF ACCEPTANCE: 14 September 2014

1 Motivation

The Center of Higher Education of the Renaissance (CESR)[1] offers the Humanist Electronic Libraries (BVH) on the Web for more than ten years: a great number of Renaissance books, written by Rabelais, Ronsard and other authors are represented as scanned and transcribed books. Their transcription, defined in the TEI format, strictly follows the presentation of the scanned original: paragraphs, line breaks, abbreviations, hyphens, lettering, etc. are all preserved. Figure 1 presents an extract of the novel *Gargantua* from Rabelais as presented on the BHV website[2]: one paragraph as transcribed and scanned text. The same paragraph tagged

1 http://cesr.univ-tours.fr

2 http://www.bvh.univ-tours.fr/

in TEI-format is given bellow (--- denotes paragraph and <lb/> denotes line break).

* *
<lb/>
<hi rend="larger">E</hi>
N ceste mesmes saison Fayoles
<lb/>quart roy de Numidie envoya
<lb/>du pays de Africque a Grand-
<lb rend="hyphen"/>gousier une jument la plus enorme & la
<lb/>plus grande que feut oncques veue, &
<lb/>la plus monstreuse, Comme assez scavez,
<lb/>que Afrique aporte tousjours quelque
<lb/>chose de noveau.
* *

Rabelais wrote about Tours Region in France where the giants Grandgousier, Gargantua and Pantagruel move into. With the *Renom* project[3] we planned to develop the so-called 'literary tourism' by introducing links between names and books: the website enables visitors to navigate through the novel using the proper names and to see where the imaginary or antique places were 'located'. Figure 2 shows the imaginary Theleme Abbacy between the towns Chinon and Azay-le-Rideau, near Tours. Tourists are encouraged to visit the castles of these two towns and the Rabelais Museum near this 'place'.

EN ceste mesmes saison Fayoles
quart roy de Numidie envoya
du pays de Africque a Grand-
gousier une jument la plus enorme & la
plus grande que feut oncques veue, &
la plus monstreuse, Comme assez scavez,
que Afrique aporte tousjours quelque
chose de noveau.

Car elle estoit grande comme six Ori
flans, & avoit les pieds fenduz en doigtz,
comme le cheval de Jules Cesar, les au
reilles ainsi pendentes, comme les chievres
F iiij

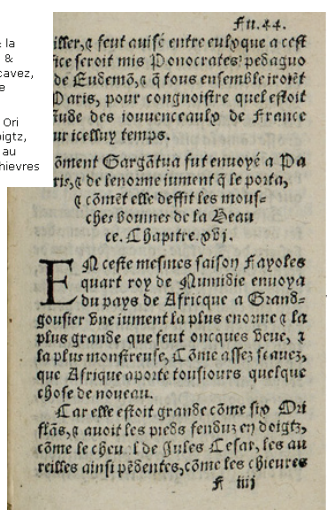
Fn.44.
iller,a feut auisé entre eulxque a ceste
ice seroit mis Ponocrates pedaguo
de Eudemõ,a q tous ensemble iroiẽt
Paris, pour congnoistre quel estoit
tude des iouuenceaulx de France
ur icelluy temps.
õment Gargãtua fut enuoyé a Pa
ris,& de lenomme iument q le porta,
q cõmẽt elle deffit les mouſ-
ches bonnes de la Beau
ce. Chapitre. xBj.
EN ceste mesmes saison Fayoles
quart roy de Numidie envoya
du pays de Africque a Grand-
gousier Bne iumment la plus enorme & la
plus grande que feut oncques Beue, & la plus monstreuſe, Cõme aſſez ſcauez,
que Afrique aporte touſiours quelque
choſe de noueau.
Car elle eſtoit grande cõme ſix Ori
flãs,a auoit les pieds fendus en doigtz,
cõme le cheu l de Jules Ceſar, les au
reilles ainſi pẽdentes,cõme les chieures
F iiij

FIGURE 1. An extract from the BVH website

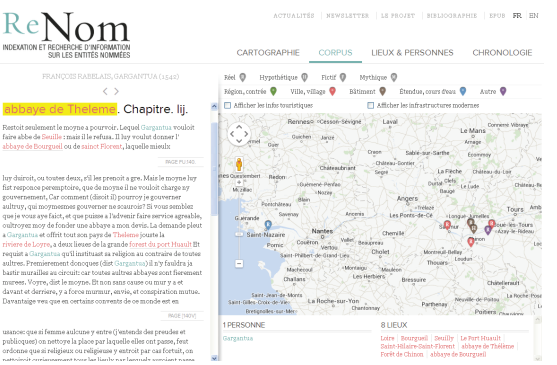

FIGURE 2. A link between a text and a map on the Renom website

The aim of this scientific project was to enrich the texts with tags indicating names of persons, personages and locations with supervised Natural Language Processing (NLP) techniques. This task, known as Named Entity Recognition (NER), has been well defined in NLP and is being tackled for two decades. In our project, the challenge was the existing TEI format, the extensive use of hyphens and the great variability in the spelling of proper names (and in the orthography of words on the whole). For instance, in the extract presented in Figure 1, the name *Grandgousier* is written *Grand-<lb rend="hyphen"/>gousier* and the name *Africa* has two orthographies, *Africque* and *Afrique*.

As introduced by MUC conferences (*Message Understanding Conferences*), the named entities include person names, location names, organization names, dates, percentages, currency (Chinchor, 1997) and sometimes titles, hours, occupations, etc. A state of the art of NER can be found in (Nadeau, Sekine, 2009); the main idea is to use *internal and external evidence* (MacDonald, 1996), i.e. the local context. For instance, in the sequence *Hugues Thierry Salel l'ainé, seigneur de Seuille*, the first name *Hugues Thierry* indicates that *Hugues Thierry Salel* is a person name (*internal evidence*) and the title *seigneur* (*lord*) proves that Seuille is a toponym (*external evidence*).

For NER task three approaches are possible: machine learning, symbolic rules and hybrid approaches. Machine learning techniques

3 http://renom.univ-tours.fr/

need training corpus, not available for this task. For this reason we used an approach based on symbolic rules (Ait-Mokhtar et Chanod, 1997; Hobbs *et al*., 1997). To facilitate the cooperation between computer scientists, linguists and Renaissance experts, we chose Unitex platform[4] (Paumier, 2003) because of its friendly interface (with use of graphs) and its free license. With Unitex, we can define cascade rule systems (CasSys menu) that benefit from all numerous functionalities of Unitex graphs. Cascades (Abney, 1991) are used in many NLP applications, such as chunking (Abney, 1996), syntactic analysis (Kokkinakis, Kokkinakis, 1999), morphological analysis (Alegria et al., 2001) and so on. Our system is inspired by (Friburger, Maurel, 2004).

The NLP community is interested in ancient languages and ancient states of modern languages (see for instance (Denooz, Rosmordus, 2009)). In old French, orthography was not fixed and a lot of name variants existed. For middle French (just before Renaissance French), specific tools and dictionaries (Souvay, 2004) or lemmatizers (Souvay, 2007) have been developed. In order to accomplish our task we also had to build specific dictionaries and cascades, which we have done in cooperation with Renaissance experts. The goal of Renom project is to help experts to tag texts precisely and to enhance dictionaries. These dictionaries contain proper names and their variants linked with unique keys (proposed by Renom and validated by experts) and locations linked with unique keys to the Geonames[5] database. This pointer is used by the website to post the map. When some too tiny locations or ancient locations were not found in Geonames, experts added new entries to Geonames.

2 Corpus presentation

The corpus contains 11 books:

–*Discours fantastiques* (edition of 1566), Justin Tonnelier;

–*Courtisan* (edition of 1538), Baldassare de Castiglione;

–*Voyage de Tours* (edition of 1560) and *Élégie sur les troubles d'Amboise* (edition of 1563), Pierre de Ronsard;

–*Gargantua* (edition of 1542), *Pantagruel* (edition of 1542), le *Tiers Livre* (editions of 1546 and 1552), le *Quart Livre* (editions of 1548 and 1552) and *Brève déclaration* (edition of 1552), François Rabelais.

The format is very particular: as we explained in Section 1, it respects to the utmost degree the layout of the original edition, line breaks, footers, initial letters and so on. Sometimes, transcribers added some corrections as well. Some examples of this format are:

- Initial letter (*Pantagruel*)

  <lb/><hi rend="larger">P</hi>

  Antagruel quelque jour pour se

- Transfer of the end of the first line at the end of the second one, after parenthesis, following the Latin citation on the same line – transcription is on three lines.[6]

  <item>Les hoseaulx, alias les bottes de patien
  <lb rend="hyphen"/>
  <hi rend="bottom">(ce.</hi>
  </item>
  <item><foreign xml:lang="lat">Formicarium artium</foreign>.</item>

- Footer ("Dace" truncated by the page number - 188)

  <lb rend="hyphen"/>
  bek Norwerge, Sweden, Rich, Da-
  <lb/>
  <lb/>
  <lb/>
  <pb n="188" xml:id="_Page_-188"/>
  <lb/>

4 http://www-igm.univ-mlv.fr/~unitex/
5 http://www.geonames.org/
6 The edited text is: Les hoseaulx, alias les bottes de patien Formicarium artium. (ce.

```
<lb/>
<fw place="top-left" type="pageNum">[94v]
</fw>
<lb rend="hyphen"/>ce, Gotthie, Engroneland, les
Estre-
```

- Correction (insertion of an apostrophe)

```
<item>Les aultres a sainct Jean
<choice>
<orig>dangery</orig>
<reg>d'angery</reg>
</choice>.
</item>
```

The named entity brackets have to contain all the format brackets. For instance, the last example becomes:

```
<item>Les aultres a <placeName>sainct Jean
<choice>
<orig>dangery</orig>
<reg>d'angery</reg>
</choice></placeName>.
</item>
```

3 Typology used

The CESR used TEI format for transcribed texts, so it was natural to adopt the same TEI typology. We used four types: geography (*geogName*), places (*placeName*), organizations (*orgName*) and persons (*persName*).

3.1 Geography and places

Geographic names were divided into two subtypes: geonyms (mountains, plains, plateaus, grottos…) and, hydronyms (oceans, seas, rivers, lakes, ponds…), for which a special value of the attribute *type* of the tag *geogName* was used. When possible, the geographic names were additionally specified by adding a specific internal tag *geogFeat*.

```
<geogName type="geo" key="#loc_montsinai">
<geogFeat>mont</geogFeat> Sinaï
</geogName>
<geogName type="hydro" key="#loc_loire">
<geogFeat>rivière</geogFeat> de Loyre
</geogName>
```

Place names were sometimes also subtyped (cities, countries, estates and buildings) by using the same attribute *type*.

```
<placeName type="city" key="#loc_seuilly">
Seuille</placeName>
<placeName type="country" key="#loc_france">France</
placeName>
<placeName type="building" key="#loc_
lapommardiere">mestayrie de la Pomardiere
</placeName>
```

Two locations were sometimes embedded.

```
<placeName type="building">Palais de
<placeName type="city" key="#loc_poitiers">
Poictiers</placeName>
</placeName>
<placeName key="#loc_guevede">gue de
<geogName type="hydro" key="#loc_vede">
Vede</geogName>
</placeName>
<geogName key="#loc_ilescanaries">isles de <placeName
key="#loc_canaries">Canarre
</placeName></geogName>
```

3.2 Organizations

Organizations were divided into three subtypes: people, estates and communities. The CESR chose not to link organizations with keys.

```
<orgName type="domaine">Royaulme de
<placeName type="pays" key="#loc_france">
France</placeName></orgName>
```

Organizations could also be embedded.

```
<orgName type =»domaine»>Royaulme des
<orgName type = «peuple»>Dipsodes</orgName>
</orgName>
```

When it was difficult to decide whether a name refers to a place or an organization we inserted both tags, *placeName* and *orgName*.

```
<placeName type="building"
key="#loch_coingnaufondabbaye">
<orgName type="community">abbaye de
<placeName type="city"
key="#loch_coingnaufond">
Coingnaufond</placeName>
</orgName></placeName>
```

3.3 Persons

In simplest cases just *persName* tags (with their keys) were used.

```
<persName key="#pers_aristote">Aristote
</persName>
```

We added internal tags for first names (*foreName*), surnames (*surName*) and particles (*nameLink*), where necessary.

```
<persName key="#pers_francoisconnan">
<forename>François</forename>
<nameLink>de</nameLink>
<surname>Connan</surname></persName>
```

Finally, these tags were extended with titles or civilities (*roleName*) that were subtyped: nobility role, religious role, function or occupation, honor. When the title included a place name, it was also tagged: the *lord of Essars* is a person, but *Essars* is a place:

```
le
<persName key="#pers_seigneurdesessars">
<roleName type="nobiliary">seigneur
</roleName>
<placeName key="#loc_desessars">des Essars
</placeName></persName>, &amp;amp; quelques
```

We sometimes additionally specified personal names when nicknames or roles within the family were mentioned ('the elder son' below).

```
<persName key="#pers_huguesthierrysalel">
<forename>Hugues</forename>
<forename>Thierry</forename>
<surname>Salel</surname>
<genName>l'ainé</genName>,
<roleName type="nobiliary">seigneur de
<placeName type="ville" key="#loc_seuilly">
Seuille</placeName></roleName></persName>
```

Texts contained ambiguities. If possible, the expert would choose the right interpretation. For instance below, *saint Martin de Candes* may refer to a church or a person:

```
<persName key="#pers_sainctmartindecan
dessaintmartin"><placeName key="#loc_sa
inctmartindecandessaintmartin">sainct
<lb/>Martin de <placeName type="city"
key="#loc_candessaintmartin">
Candes</placeName>
</placeName></persName>
```

4 Dictionaries

Persons	1 145
Locations	987
Organizations	57
Other words	2 622

TABLE 1. Number of dictionary entries

As we mentioned earlier, it was often necessary that our NER recognizes the context of named entities. To that end we built a variant orthography dictionary, studying contexts and using an old first name list.

For instance, in the Renaissance, the word "captain" was written in three different ways: *capitaine*, *capiteine* or *cappitaine*. We chose the synchronic entry as the lemma and we added features that NER could use (see sections 6.3 and 6.4):

capitaine,.N+Military:ms

capiteine,capitaine.N+Military:ms
cappitaine,capitaine.N+Military:ms

For instance, the second line contains five pieces of information: form (*capiteine*), lemma (*capitaine*), part of speech (*N*), feature (*Military*) and morphology (*ms*).

We transformed these three CESR lists of names: persons, organizations, locations in the format of Unitex dictionaries (illustrated above), which were improved after each book was parsed. In produced dictionaries of names, a word has its key for lemma. These keys are used to link different orthographies and also to link locations to Geonames:

ancenis,loc_ancenis.N+id=loc:ms
ancenys,loc_ancenis.N+id=loc:ms

In their previous work, CESR experts chose to use explicit keys (as *loc_ancenis* for the toponym *Ancenis*) so we were obliged to use the same approach. Table 1 presents the number of dictionary entries at the end of *Renom* project.

5 Improvements of Unitex platform

As we mentioned earlier, we chose Unitex platform to facilitate the cooperation between computer scientists, linguists and Renaissance experts. Unitex is open-source and free (LGPL license). With Unitex, one can parse texts with his own dictionaries (see Section 4) and write linguistic rules in a form of graphs with a very friendly interface; it is also possible to build cascades of graph with CasSys.

A graph cascade is a succession of parsing graphs: the first graph parses the text, the second graph parses the text modified by the first graph and so on.

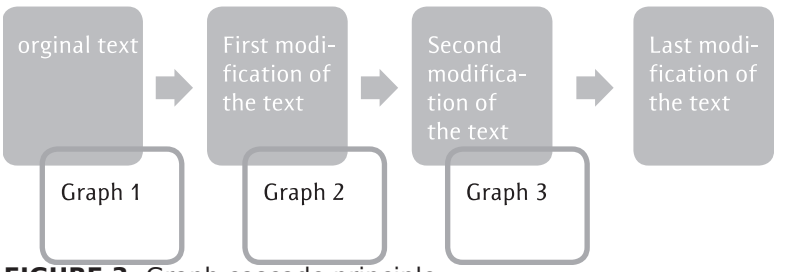

FIGURE 3. Graph cascade principle

For the *Renom* project we had to add three improvements of CasSys that will be explained in the following subsections: enabling the graph iteration until certain fixed point, and the use of Unitex morphological dictionaries and producing the more appropriate output files of the cascade.

In Unitex system, graphs parse text with possibility to merge new sequences, replace others, use variables, move sequences, and insert information from dictionaries.

5.1 Graph iteration

We added to CasSys graph iteration until certain fixed point: the iterative graph parses the text and produces a modified text, then it parses this modified text and so on until the parsing does not modify the resulting text any more (this is the fixed point). We used iterative graphs primarily to build keys of embedded types (see section 6.5).

5.2 Unitex morphological dictionaries

For a form occurring in a text a Unitex graph can extract information associated to it in dictionaries: its lemma, features or morphology codes. These dictionaries are named 'morphological dictionaries'.

We added the possibility to include this functionality into cascades. We defined our three dictionaries of names (see Section 4) as 'morphological dictionaries' which enabled us to link a name found in one of these dictionaries to its key. More information is given in Section 6.5.

5.3 File output of a cascade

The major idea when parsing a text with a cascade of graphs is to consider an already tagged piece of a text as a *multiword expression* (*MWE*) because the

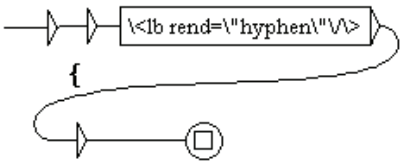

FIGURE 4. A graph recognizing XML tag *lb*[7]

other graphs of the cascade than cannot parse inside it. A sequence of characters is recognized by the Unitex system as a MWE if it is enclosed with curly brackets.

For instance, the XML tag *<lb rend="hyphen"/>* will be interpreted as a MWE if curly brackets are added before and after the tag:

{<lb rend="hyphen"/>, BaliseXML+DFlb+hyphen}

That is done by the graph presented in Figure 4. The additional information inside curly brackets (after a comma and a dot) enables the subsequent graphs of the cascade to parse this MWE with Unitex expressions as *<BaliseXML>* (for a XML tag), *<DFlb>* or *<hyphen>*, depending on the necessary degree of precision.

A disadvantage is that the resulting text is difficult to read. To overcome this we implemented a specific CasSys XML format that is used as an alternative cascade output:

```
<csc><form><lb rend="hyphen"/>
</form>
<code>BaliseXML</code>
<code>DFlb</code>
<code>hyphen</code></csc>
```

For that reason our cascades always come in a pair: the first one is used to parse a text while the second one transforms the specific CasSys XML format (the result of the first cascade) into the required format. The first graph of the cascade transforms all the original XML tags to MWEs.

6 Method

Our corpus contained two groups of texts. The first group already contained tags for proper names (*persName*, *geogName* and *placeName*) that were added manually by experts; our work on these texts was to add internal tags (*geoFeat*, *foreName*, *lastName*…), to search for keys in the dictionaries and to extend names to named entities by adding tags such as *roleName* (Lord, Abbey…), *genName* (elder son…) and so on. In the second group of texts only formatting was annotated, which means that we had to recognize names in these texts, before doing the same work as for the first group of texts.

We organized our work in four steps (see Figure 5). A few pre-tagged texts were parsed only from the third step.

1. Preprocessing to rebuild truncated names at the end of a line or a page;
2. Dictionary lookup and use of context rules to
3. tag names;
4. Consultation of dictionaries and application
5. of internal and expanded rules, as presented before (*firstName* versus *genName* and so on).
6. Extraction of names that are not in dictionaries.

6.1 Hyphens

As we said, a lot of hyphens occurred in texts. A first graph recognized XML tags as multiword
expression (MWE) with the part of speech (POS) *baliseXML*. The other graphs of the first cascade rebuilt words. For instance, the graph given in Figure 6 recognizes a letter larger in size than others: the *hi* tag cuts the word and the graph rebuilds it. Two *hi* tags (start and end tags) become a new MWE with POS set to *largerSup* and its new attribute memorizes

7 As all the other graphs, this graph adds as a feature its own name, which helps in debugging the cascade..

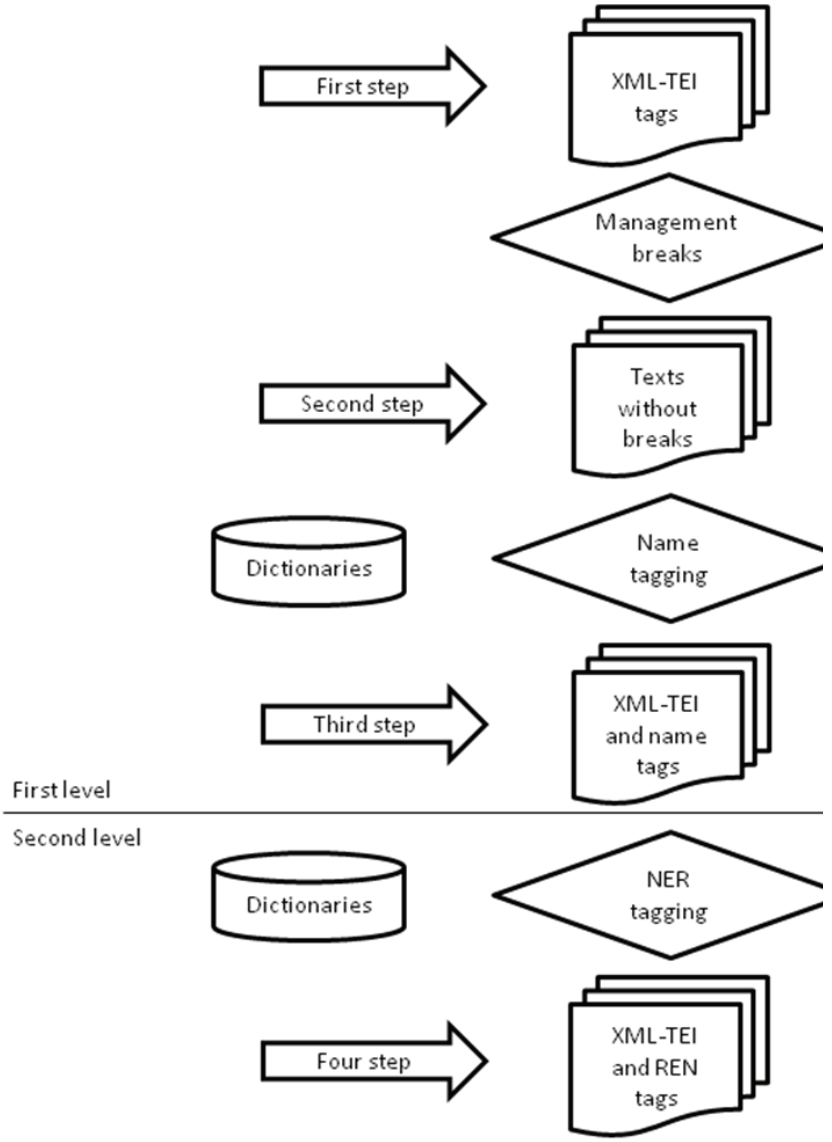

FIGURE 5. The orgnaization of work in four steps

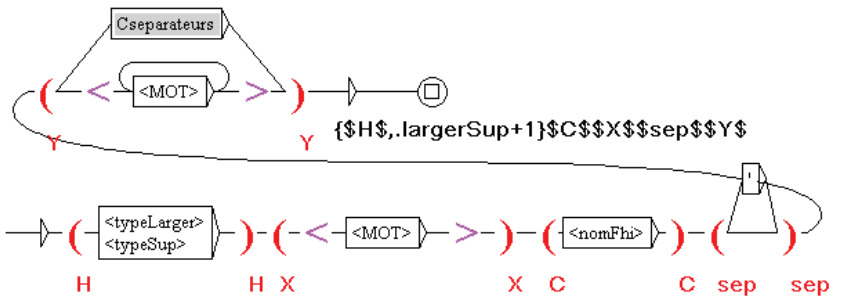

FIGURE 6. Graph for rebuilding words having the larger first letter.

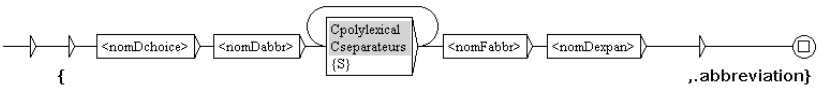

FIGURE 7. Graph for hiding abbreviations

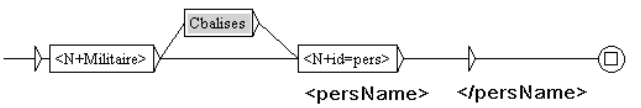

FIGURE 8. Graph for adding a tag *persName* to names found in dictionaries that occur in the military context

the scope of the highlighted sequence (here, *value="1"*).
For instance:

<hi rend="larger">E</hi>

N ceste mesme heure becomes
{<hi rend="larger" value="1"></hi>,.largerSup}
EN ceste mesme heure

6.2 Abbreviations

The graph presented in Figure 7 recognizes the XML tags *choice*, *abbr* and *expan*, and builds a MWE[8] with POS *abbreviation*.
For instance:

<choice>
<abbr>PAN.</abbr>
<expan>PANURGE</expan>
</choice>
becomes
{<choice><abbr>PAN.</abbr><expan>,.
abbreviation
}PANURGE</expan></choice>

6.3 Names recognized from dictionaries

The second step began with a dictionary lookup to tag names that were in dictionaries. Some names were ambiguous, so we used the context to disambiguate persons from locations or organizations. The graph given in Figure 8 adds a tag *persName* to personal names in the military context.
For instance:

<lb/> du capitaine Engoulevent, pour descou

becomes
*<lb/> du capitaine <**persName**>Engoulevent*
*</**persName**>, pour descou*

6.4 Names recognized only from context

After we recognized names from dictionaries, we used the same contexts to tag names that were not in the dictionaries, if the first letter was capitalized. The graph in Figure 9 tags unrecognized

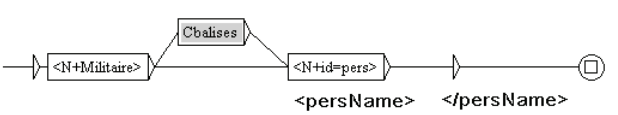

FIGURE 9. Graph for tagging unrecognized names with first letter capitalized with *persName* if occurring in the military context

8 This MWE hides the original word PAN to the parsing. Here, this name is ambiguous to the Greek mythological god Pan

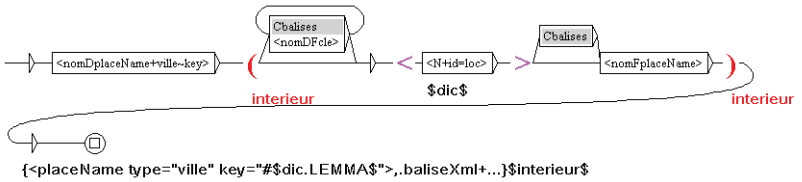

FIGURE 10. Graph that searches key in dictionaries

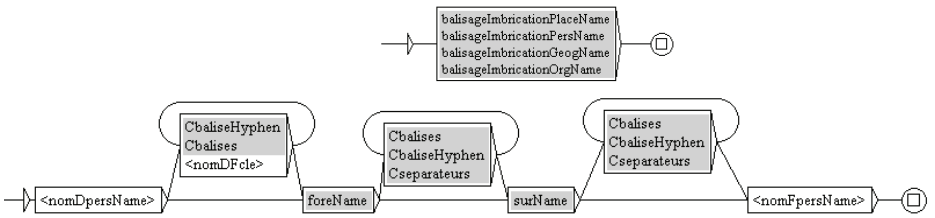

FIGURE 11. Iterative graphs for embedded names

names with *persName* in the military context.
For instance:

<lb/> du chevalereux capitaine Moses
becomes
<lb/> du chevalereux capitaine <persName>Moses</persName>

6.5 Keys

When names were identified, we searched their keys in dictionaries, if such entries existed. The graph presented in Figure 10 returns these keys in the form of the key attribute of the appropriate start tag.
For instance:

mestaiers de <placeName>Seuille</placeName> & de <placeName>Synays</placeName>.

becomes

mestaiers de <placeName ***key="#loc_seuilly"****>Seuille</placeName> & de <placeName* ***key="#loc_cinais"****>Synays</placeName>.*

If the name was not found in dictionaries, other graphs built its possible key by concatenating keys of names it consists of. The attribute *dic="no"* indicated to experts that the name (with this orthography) was not found in the dictionaries. The expert added it, with another key if it was a variant of an existing entry or with this key if it was really a new entry.

To build a key was not trivial, because of embedded names. For instance, we had to add three keys for the name *château du gué de Vede*: one for the proper noun 'Vede' (the key extracted from dictionaries), one for 'the ford of Vede' and one for 'the castle of the ford of Vede':

<placeName ***key="#loc_chasteauduguedevede" dic="no"****>chasteau du <placeName* ***key="#loc_guedevede" dic="no"****>Gue de <geogName* ***key="#loc_vede"****>Vede</geogName></placeName></placeName>*

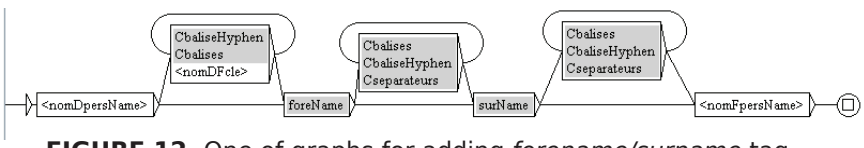

FIGURE 12. One of graphs for adding *forename/surname* tag

The graph that built keys of embedded names is applied iteratively. It calls four subgraphs, one for each type. Figure 11 presents this graph and one of its subgraphs that inserts a key in a *placeName* tag.

6.6 Tags inside of names

We inserted some tags inside name tags: tags for forenames and surnames inside *persName* tags and *geogFeat* tags inside *geogName* tags. For this task we used a Renaissance dictionary of forenames. The following instances posed some difficulties: a person with more than one forename (*Hugues Thierry Salel*), with a multi-word surname (*Jan Trivolse Guallo*) or with a particle (*Ulrich Thierry du Gallet*). The graph given in Figure 12 tags one forename and one surname (tags are in subgraphs). After application of this graph both the names and their inside tags are considered as MWE.

6.7 Extended tags

We also extended *persName* to named entities that contained tags *roleName* (Lord, Abbey, teacher...), *genName* (elder son...) or *addName* (nicknames). We added new tags to the left or to the right of an already recognized personal name and we moved *persName* tags. The graph in Figure 13 adds tags to the left of a named entity when it is preceeded by a *roleName*. In this case the key does not change.

For instance, *Epistemon* is a personal teacher (*précepteur*), so the sequence

<lb>ton precepteur
<persName key="#pers_epistemon" dic="no">
Epistemon<persName> don't

becomes

<lb>ton
<persName key="#pers_epistemon" dic="no">
<roleName type="function">precepteur
<roleName> Epistemon<persName>

#I	#D	#T	#E	#TE	#S	#R
5	19	3	3	0	136	150
SER					6,1%	
Precision					96,3%	
Recall					87,3%	
Type precision					94,1%	
Limits precision					94,1%	

TABLE 2. Evaluation

6.8 Dictionary enhancement

Finally, the two last cascades built a new file for dictionary enhancement: they erased the text, except the names that were not already in dictionaries or were in them but with different features. For instance, the last example:

<lb>ton precepteur
<persName key="#pers_epistemon" dic="no">
Epistemon<persName> don't

becomes

Epistemon #pers_epistemon

The entire list of names with the feature *dic="no"* was transmitted to experts to improve the dictionaries.

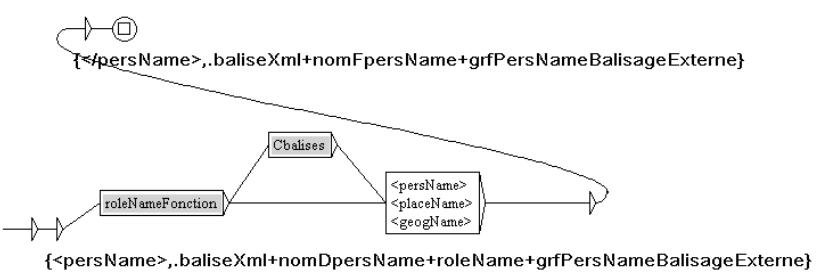

FIGURE 13. Graph for tagging a *roleName* at the left of a name

7 Evaluation

To evaluate our work, we parsed two books of Pierre de Ronsard from our corpus (*Voyage de Tours* and *Élégie sur les troubles d'Amboise*). We did not use these books to construct our cascades – for developing them we used mainly the François Rabelais' books. We computed a weighted variant of the *slot error rate* (SER) (Makhoul et al., 1999) used in French evaluation campaign. SER distinguished between three types of errors:

Insertion (I - weight 1): we tagged words that are not names.

Deletion (D - weight 1): we failed to tag a name.

Tags with border errors: bad type (T - weight 0.5), tag outside or inside the proper name (E - weight 0.5) or both (TE - weight 1).

If #R is the sum of the entities of the reference texts, the SER is computed by:

$$SER = \frac{\#I+\#D+0{,}5*\#T+0{,}5\ \#E+\#TE}{\#R}$$

With these counts, if #S is the sum of the detected entities, we can also compute the precision and the recall of our work:

$$Precision = \frac{\#S - \#I}{\#S} \quad \text{and} \quad Recall = \frac{\#S - \#I}{\#R}$$

The tagged texts are entirely supervised, so border errors are less important. But we can also compute the type precision (correct recognition of types) and the limit precision (the boundaries of a named entity):

$$Type\ precision = \frac{\#S-\#I-\ \#T-\#TE}{\#S} \text{ and}$$

$$Limit\ precision = \frac{\#S-\#I-\ \#E-\#TE}{\#S}$$

Results are presented in Table 2.

The experts of CESR wanted to read the whole corpus before its publication on the website. The SER of 6.1% indicates that their work was really improved. The significant number of deletion errors corresponds to a lot of names without context that were not represented in dictionaries. At the end, the experts enhanced them.

8. Conclusion

In this paper we presented the NER task applied to XML-TEI encoded Renaissance texts. The format of the corpus and the significant variations of vocabulary asked for the specific treatment as compared to contemporary texts. We used dictionaries and rule-based cascades and we obtained 6.1% of SER. Our system will enter the production line of transcribed texts.

The most important texts of Rabelais are now online at the *Renom* website powered by a search engine for names. It uses keys to link name variants and to connect them to map locations.

Acknowledgment

This work is supported by Région Centre research program. Authors thank the CESR, particularly Sandrine Breuil, Marie-Luce Demonet, Jorge Fins and Marie Olivron.

References

Abney Steven. "Parsing By Chunks". In *Principle-Based Parsing*, edited by Robert C. Berwick, 257-278. Dordrecht: Kluwer Academic Publishers, 1991.

Abney Steven. "Partial Parsing via Finite-State Cascades". *Natural Language Engineering*, Vol. 2, Issue 4 (1996): 337-344.

Ait-Mokhtar S. and Jean-Pierre Chanod. "Incremental Finite-State Parsing". In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, program chair Ralph Grishman, 72-79. Stroudsburg: Association for Computational Linguistics, 1997.

Alegria I., M. J. Aranzabe , N. Ezeiza, A. Ezeiza and R. Urizar. "Using Finite State Technology in Natural Language Processing of Basque". In Implementation and Application of Automata: 6th International Conference, CIAA 2001 Pretoria, South Africa, July 23–25, edited by Bruce W. Watson, Derick Wood, 1-12. Berlin: Springer, 2002.

Chinchor Nancy. "Muc-7 Named Entity Task Definition", Version 3.5, 17 September 1997, http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

Denooz J and S. Rosmorduc (eds.). *Langues Anciennes*, TAL, Vol. 50, No. 2(2009), http://www.informatik.uni-trier.de/~ley/db/journals/tal/tal50.html#McGillivrayPR09

Friburger N. and D. Maurel." Finite-state transducer cascade to extract named entities in texts". *Theoretical Computer Science*, Vol. 313, Issue 1 (2004): 94-104.

Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. " FASTUS: A cascaded finite-state transducer for extracting information from natural-language text". In *Finite-State Language Processing*, 383-406. MA: MIT Press, 1997.

Kokkinakis D. and S. J Kokkinakis . "A Cascaded Finite-State Parser for Syntactic Analysis of Swedish". In *EACL'99: 9th Conference of the European Chapter of the Association-for-Computational-Linguistics, Bergen, June 8-12*, 245-258. Bergen: Bergen University Fund, 1999.

MacDonald D. "Internal and external evidence in the identification and semantic categorisation of Proper Names". In *Corpus Processing for Lexical Acquisition*, 21-39, MA: Massachussetts Institute of Technology, 1996.

Makhoul J., Kubala J., Schwartz R., Weischedel R. "Performance measures for information extraction". In Proceedings of *DARPA Broadcast News Workshop*, 249-252. San Francisco: Morgan Kaufmann, 1999.

Nadeau David and Sekine Satoshi. "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, Vol. 30, Issue 1 (2007): 3-26.

Paumier S. "De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique". Thèse de Doctorat en Informatique, Université de Marne-la-Vallée. 2003.

Souvay G. "Vers un Dictionnaire électronique du Moyen Français ". In *Actes du Colloque Euralex 2004, European Association for Lexicographycongress Lorient, France, 6-10 juillet, vol. 2*, 671-678. Lorient : Universite de Bretagne Sud, 2004.

Souvay G. "LGeRM : un outil d'aide à lemmatisation du français médiéval". Paper presented at the 18th International Conference on HistoricalLinguistics ICHL 2007, Université du Québec À Montréal. Canada. 6-11 août. 2007.