Tourisme culturel sur Internet : Les noms propres des éditions originales de Rabelais

Denis Maurel, Nathalie Friburger, Iris Eshkol

Tourisme culturel sur Internet : Les noms propres des éditions originales de Rabelais

Denis Maurel[1], Nathalie Friburger[1], Iris Eshkol-Taravella[2]

[1]*Université François-Rabelais de Tours, Laboratoire d'informatique*

[2]*Université d'Orléans, Laboratoire ligérien de linguistique*

denis.maurel@univ-tours.fr, nathalie.friburger@univ-tours.fr, iris.eshkol@univ-orleans.fr

RÉSUMÉ. Pour exploiter le contexte culturel des œuvres présentées dans un cadre touristique de développement régional, l'idée est venue de créer un site permettant d'associer la navigation dans les œuvres via les noms propres et le déplacement touristique en Région Centre, terre de Ronsard et de Rabelais, mais aussi de Gargantua et de Pantagruel ! Le projet Renom vise à permettre une meilleure exploitation de ce patrimoine en fournissant à l'internaute des outils pour naviguer à partir des noms de personne ou de lieu, et en l'invitant à compléter sa consultation par une visite de la Touraine (par exemple le musée de La Devinière, s'il s'agit de Rabelais, mais aussi les châteaux qu'il mentionne, celui de Chinon, entre autres, ou l'Abbaye de Thélème, même si elle n'a jamais existé !).

Cet article décrit l'ajout d'une recherche de noms propres à une chaîne de transcription de textes de la Renaissance. La recherche de noms propres dans des textes de la Renaissance présente deux difficultés qui la complexifient : une grande variété d'écriture, l'orthographe n'étant pas encore fixée, et la présence d'un grand nombre de balise XML de formatage TEI, car les textes gardent trace de la présentation de l'édition originale. L'objectif est d'ajouter à ce balisage TEI un balisage supplémentaire des noms propres avec leurs extensions à gauche ou parfois à droite.

Nos programmes ont apporté une aide précieuse aux experts annotateurs et sont aujourd'hui intégrés à la chaîne de transcription.

ABSTRACT. The Region Centre-Val de Loire tries to develop tourism by the means of cultural context of regional works. One idea is the Web site creation presenting the Proper Names extracted from Regional books linked to some tourist travels. The Region Centre-Val de Loire is the demesne of Ronsard and Rabelais, but also the demesne of the giants Gargantua and Pantagruel. The Renom Project presents both the original text of Rabelais (for instance) and the travel maps of Gargantua. The web site induces the tourist to visit the Chinon Casle or the Thelem Abbey… The web site induces the tourist to visit the Chinon Castle or the Thelem Abbey… even it never existed! But also the La Devinière Rabelais Museum.

This paper deals with adding Named Entity Recognition at a transcription line of Renaissance texts. Named Entity Recognition in Renaissance texts presents two new challenges: great diversity of writing, due to word various orthographies; numerous XML-TEI tags to save the exact format of original edition. The task consisted to add Named Entity tags to this first tagging with generally their left context and sometimes right context.

Tourisme culturel

To do that, we improved the free program CasSys to parse texts with Unitex graph cascades and we built four dictionaries and eight specific cascades. The record was 80,9 % and the precision was 77,9 %. These results provided valuable assistance to annotators experts. Today, the cascades are integrated to the transcription chain.

MOTS-CLÉS : Tourisme culturel; Noms propres; textes de la Renaissance; cascades de graphe; CasSys; enrichissement de transcription.

KEYWORDS: Cultural tourism; Named entities; Renaissance texts; graph cascades; CasSys; Transcription enrichment.

1. Introduction

Dans le cadre de ce qu'on appelle aujourd'hui les *Humanités numériques*, les bibliothèques et les centres de recherche sur les documents anciens souhaitent mettre à disposition du public des textes numérisés et, souvent, enrichis. C'est le cas par exemple du site Gallica[1] de la Bibliothèque nationale de France ou du site des Bibliothèques Virtuelles Humanistes (BVH)[2] du Centre d'Études Supérieures de la Renaissance (CESR) de l'université François-Rabelais de Tours.

Pour réaliser le site des BVH, le CESR a conçu une "chaîne de transcription" pour réaliser la transcription des ouvrages originaux de la Renaissance en sa possession. Cette chaîne comprend la transcription elle-même, sa relecture, son enrichissement par des balises respectant la *Text Encoding Initiative* [3](TEI) afin d'annoter le format des ouvrages originaux dans le texte lui-même, ainsi que la correction (toujours précisée par des balises TEI) de certaines séquences présentes dans le document[4].

Pour exploiter le contexte culturel des œuvres présentées dans un cadre touristique de développement régional, l'idée est venue de créer un site permettant d'associer la navigation dans les œuvres via les noms propres et le déplacement touristique en Région Centre, terre de Ronsard et de Rabelais, mais aussi de Gargantua et de Pantagruel ! Le projet Renom[5] vise à permettre une meilleure exploitation de ce patrimoine en fournissant à l'internaute des outils pour naviguer à partir des noms de personne ou de lieu, et en l'invitant à compléter sa consultation par une visite de la Touraine (par exemple le musée de La Devinière, s'il s'agit de Rabelais, mais aussi les châteaux qu'il mentionne). Il sera même possible de « localiser » virtuellement des lieux tels que Thélème (qui n'a jamais existé) en fonction des données fournies par le

[1] http://gallica.bnf.fr/

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

[3] http://www.tei-c.org/index.xml

[4] Entre autre l'article aujourd'hui écrit *l'* qui est collé au mot dans les corpus étudiés. Un exemple est donné Tableau 1.

[5] http://tln.li.univ-tours.fr/Tln_Renom.html

texte. Le premier site Renom (voir la Figure 1), publié le 6 février 2014, permet le lien entre le texte rabelaisien et les cartes géographiques ; le site à venir ajoutera des pointeurs vers les lieux touristiques aux alentours, en lien avec le personnage (château de Chinon…) ou l'auteur (la Devinière…).

Concrètement, les programmes développés dans le cadre du projet Renom effectuent une recherche des noms propres et des informations principales les concernant (qui sont ces personnes ou personnages ? où sont situés ces lieux ?). Cette recherche est aujourd'hui intégrée à la chaîne de transcription du CESR et permet aussi l'enrichissement des dictionnaires du Centre.

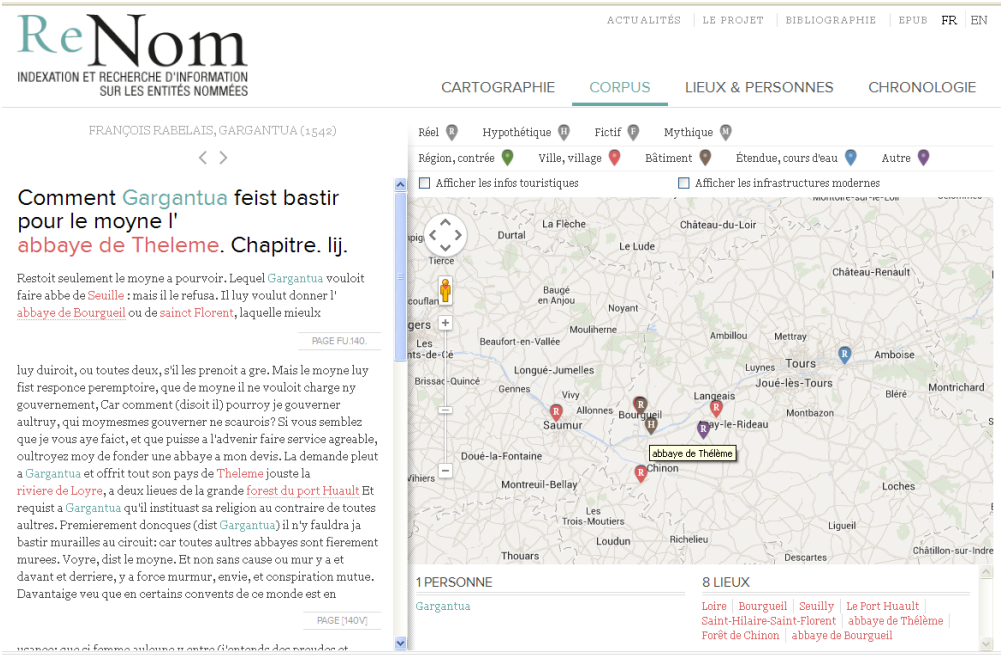

Figure 1 : Une page du site Renom

Après une présentation des données et des outils (section 2), viendra la méthodologie suivie (section 3) et, pour finir, l'évaluation des résultats obtenus (section 4).

Tourisme culturel

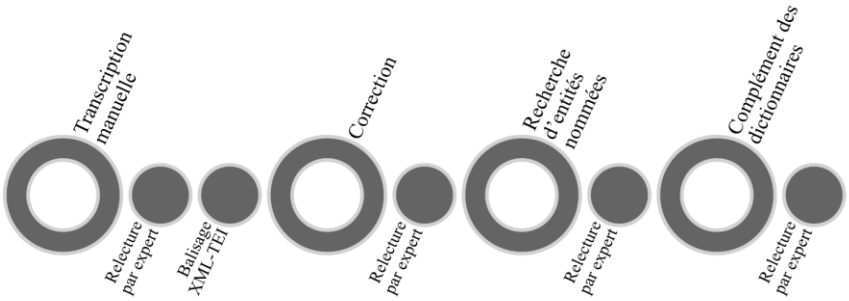

Figure 2 : La chaîne de transcription du CESR

2. Données et outils

2.1. Le corpus

Les œuvres qui composent le corpus du travail sont :

– les *Discours fantastiques*, de Justin Tonnelier (édition de 1566) ;

– le *Courtisan*, de Baldassare de Castiglione (édition de 1538) ;

– le *Voyage de Tours* (édition de 1560) et l'*Élégie sur les troubles d'Amboise* (édition de 1563) de Pierre de Ronsard ;

– *Gargantua* (édition de 1542), *Pantagruel* (édition de 1542), le *Tiers Livre* (éditions de 1546 et de 1552), le *Quart Livre* (éditions de 1548 et de 1552) et *Brève déclaration* (édition de 1552), de François Rabelais.

La chaîne de transcription du CESR produit un corpus très particulier par sa fidélité complète à la mise en page de l'édition originale, transcrite en suivant la norme XML–TEI. Le texte ne comporte pas d'accents, ou très exceptionnellement. La majuscule n'est pas toujours présente. Pour économiser de la place, les éditeurs de l'époque n'hésitaient pas à couper les mots, à les orthographier différemment suivant leur position, à les abréger ou à les finir après une parenthèse à la fin de la ligne suivante. De plus, la transcription mémorise les modifications de taille de certaines lettres ainsi que les bas de page, numéro de page, etc. Pour faciliter la lecture des ouvrages, un mot coupé en bas de page peut être réécrit en haut de la page suivante dans la marge... Tout cela est présent dans le corpus transcrit. De plus, la deuxième étape de la chaîne ajoute quelques corrections orthographiques, principalement une apostrophe pour décoller l'article du mot qu'il détermine ou l'écriture développée d'une forme abrégée. Le Tableau 1 présente un exemple pour chacune de ces situations[6].

[6] C'est nous qui ajoutons les coupures de texte [...] pour rendre les exemples plus lisibles.

Tourisme culturel

Les annotations des noms propres doivent venir s'ajouter au texte balisé actuel. Pour trouver les noms propres, il faut donc un outil permettant à la fois de préserver les balises actuelles et de les prendre en compte si nécessaire dans l'annotation. Nous avons donc choisi de ne pas utiliser d'outils XML classiques, mais des cascades de graphes dont les premiers graphes traitent le balisage XML-TEI du format d'impression.

Taille de caractères, puis coupure de mot sans et avec tiret	<hi rend="larger">**E**</hi>**N ceste mesme heure Gargan** <lb rend="hyphen"/>**tua [...]** <lb/> **fut adverty [...] comment Picrocho-** <lb rend="hyphen"/>**le sestoit rempare a la Rocheclermaud**
Complément en fin de ligne suivante[7]	<l>**Clers, basauchiens mangeurs du popu-** <lb rend="hyphen"/><hi rend="bottom">**(laire.**</hi></l> <l>**Officiaulx, scribes, & pharisiens**</l>
Bas de page	<lb/>**je luy cede la mestayrie de la Pomar-** <fw place="bot-center" type="sig">M ij</fw> <pb n="180" xml:id="B360446201_B343_2_0180"/> <fw place="top-left" type="pageNum">[90v]</fw> <lb rend="hyphen"/>**diere, a perpetuite pour luy & les siens** <lb/>**possedable en franc alloy.**
Correction[8]	<lb/>**[...] sainct** <lb/>**Thomas** <choice><orig>Langloys</orig><reg>**L'angloys**</reg></choice> **voulut bien pour** <lb/>**yceulx mourir**

Tableau 1 : Exemples de balises TEI présentes dans le corpus

[7] La parenthèse apparaît, dans le texte original, en fin de ligne, mais le transcripteur l'a déplacée au début :
Clers, basauchiens mangeurs du popu-
Officiaulx, scribes, & pharisiens (laire.

[8] Dans le texte original, on avait seulement la forme *Langloys* :
[...] sainct
Thomas Langloys voulut bien pour
yceulx mourir

Tourisme culturel

Prenons l'exemple suivant où *Panarge* est corrigé par le transcripteur en *Panurge*[9] :

```
<choice>
    <sic>Panarge</sic>
    <corr>Panurge</corr>
</choice>
neust faict esvanouyr
```

Le balisage des noms propres ne devait concerner, ni *Panarge*, ni *Panurge*, mais l'ensemble de la correction :

```
<persName>
    <choice>
        <sic>Panarge</sic>
        <corr>Panurge</corr>
    </choice>
</persName>
neust faict esvanouyr
```

2.2. L'outil

Un état de l'art concernant les techniques de reconnaissance des noms propres pourra être trouvé dans (Nadeau, Sekine 2009); l'idée principale que nous utilisons est ce que MacDonald (1996) appelle *internal and external evidence*, c'est-à-dire le contexte local interne ou externe au nom propre lui-même (un prénom, une profession…). Pour traiter l'ensemble du corpus qui nous a été proposé, nous avons utilisé et amélioré le système libre CasSys (Friburger, Maurel 2004) qui permet de construire et d'utiliser des cascades de transducteurs (Abney 1991), ou plutôt de graphes, du fait de son intégration à la plateforme Unitex (Paumier 2003).

Dans une cascade de graphes (Figure 3), le premier graphe modifie par insertion ou remplacement le texte d'origine, le second modifie le texte résultant, etc. Les exemples de la section 3 illustreront le fonctionnement de CasSys. D'autres exemples sont disponibles dans (Maurel *et al.* 2011).

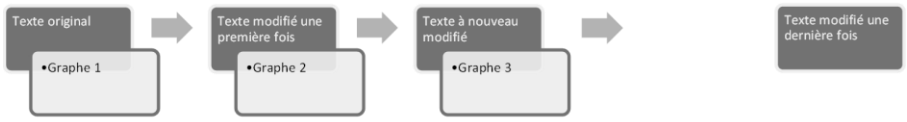

Figure 3 : Déroulement d'une cascade

[9] C'est nous qui allons à la ligne ici et dans les autres exemples.

3. Méthodologie

3.1. La gestion des multiples formes d'écriture

Pour gérer l'existence de multiples formes d'écriture des mots ou des noms propres, nous avons construit des dictionnaires (voir section 3.3.1). Pour les mots courants, c'est la *forme vedette* qui relie les différentes formes entre elles. Pour les noms propres, c'est un identifiant. Lorsque le nom propre reconnu ne figure pas dans les dictionnaires, le système en construit un.

Par exemple, la consultation d'un dictionnaire permet de savoir que le *Col Agnel* et le *col de laignel* ont le même identifiant, *loc_agnelcol*, et désignent donc un même lieu. Parfois, des personnages d'ouvrages différents portent le même nom sans désigner la même personne et ils reçoivent alors des experts des identifiants différents. Après le traitement d'un ouvrage, les noms propres non présents dans le dictionnaire et leurs identifiants construits sont soumis aux experts du CESR pour validation et intégrés aux dictionnaires pour le traitement des autres ouvrages.

Dans l'exemple ci-dessous, le graphe de la Figure 4 trouve, entre deux balises *placeName*, un nom de lieu (boîte *N+id=loc*) dans le dictionnaire, lui associe un nom de variable (*dic*) et ajoute l'identifiant, qui occupe dans le dictionnaire la position de lemme (*$dic.LEMMA$*), comme valeur de l'attribut *key*.

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

Devient :

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

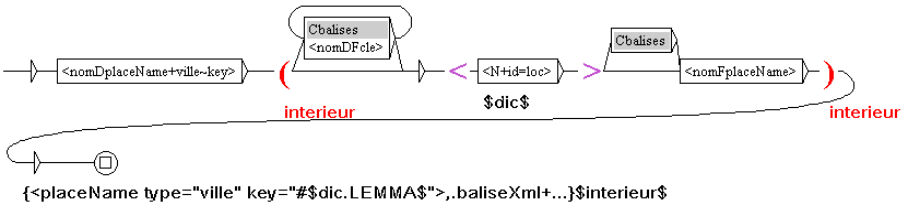

Figure 4 : Un graphe de récupération de l'identifiant dans le dictionnaire

3.2. La typologie adoptée

La typologie adoptée est celle de la TEI : les noms propres reconnus sont répartis en quatre types, les lieux géographiques (*geogName*), les lieux administratifs (*placeName*), les organisations (*orgName*) et, enfin, les personnes ou personnages (*persName*). Lorsqu'une adaptation au corpus est nécessaire, elle porte sur les attributs des balises.

Tourisme culturel

3.2.1 Les lieux géographiques et administratifs

Les lieux géographiques sont de deux types, d'une part, les géonymes : montagnes, plaines, plateaux, grottes... (*geo*) et, d'autre part, les hydronymes : océan, mer, rivière, lac, étang... (*hydro*). Les précisions géographiques sont intégrées à la balise et elles-mêmes balisées (*geogFeat*).

```
<geogName type="geo" key="#loc_montsinai">
    <geogFeat>mont</geogFeat>
    Sinaï
</geogName>

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

Les lieux administratifs peuvent éventuellement être typés (*ville*, *pays*, *batiment* ou *domaine*).

```
<placeName type="ville" key="#loc_seuilly">Seuille</placeName>

<placeName type="pays" key="#loc_france">France</placeName>

<placeName type="domaine" key="#loc_lapommardiere">
    mestayrie de la Pomardiere
</placeName>
```

Les deux types de lieux peuvent être imbriqués l'un dans l'autre.

```
<placeName type="batiment">
    Palais de
    <placeName type="ville" 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>
```

3.2.2 Les organisations

Les organisations sont partagées en trois types (*peuple*, *domaine* ou *communaute*) et n'ont pas d'identifiant associé.

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

Elles peuvent éventuellement être imbriquées.

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

Lorsqu'il est difficile de distinguer entre un lieu et une organisation, un double balisage est possible.

```
<placeName type="batiment" key="#loch_coingnaufondabbaye">
    <orgName type="communaute">
        abbaye de
        <placeName type="ville" key="#loch_coingnaufond">
            Coingnaufond
        </placeName>
    </orgName>
</placeName>
```

3.2.3 Les personnes

Le balisage le plus simple consiste à reconnaître les personnes et à leur associer leur identifiant (*key*).

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

Ce balisage peut être complété en interne par un balisage des prénoms (*forename*), des noms (*surname*) et des particules (*nameLink*).

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

Enfin, le balisage doit être étendu par des titres ou des civilités (*roleName*) qui sont typés : titres nobiliaires, militaires ou religieux, fonctions et civilités honorifiques. Lorsqu'un titre comporte un nom de lieu ou une organisation, celui-ci est aussi balisé. Enfin, les précisions familiales (*genName*) et les surnoms (*addName*) entrent aussi dans le balisage.

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

Comme pour les organisations, un double balisage est parfois possible.

```
<placeName key="#loc_saintmesmesdechinon">
    <persName key="#pers_saintmesmesdechinon">
        <roleName type="religion">sainct</roleName>
        Mesmes de
        <placeName type="ville">Chinon</placeName>
    </persName>
</placeName>
```

Tourisme culturel

3.3. Les ressources créées pour le projet

3.3.1 Les dictionnaires

Les cascades réalisées consultent quatre dictionnaires, avant le passage des graphes :

– trois dictionnaires de noms propres, gérés par le CESR, respectivement pour les personnes, les lieux et les organisations ; ces dictionnaires sont augmentés à chaque analyse par les nouveaux noms propres reconnus, après validation par les experts. Ils contiennent pour chaque entrée un nom, un identifiant, un type, éventuellement des traits (*ville*, *pays*, *batiment*...) et des variantes ;

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

– un dictionnaire créé par nous et contenant des mots déclencheurs de la reconnaissance (titres, charges, rôles, famille, prénoms...) sous de multiples formes d'écriture possibles à la Renaissance, la graphie des mots n'étant pas fixée.

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

Pour créer ce dictionnaire de mots déclencheurs, nous nous sommes servis des dictionnaires contemporains que nous possédions et des corpus annotés en noms propres pour y repérer des formes manquantes ou des variantes orthographiques.

Le Tableau 2 présente le nombre d'entrées à la livraison des cascades. Les dictionnaires sont constitués des noms propres présents dans les ouvrages et sont augmentés au fur et à mesure du traitement des documents.

Personnes	1 149
Lieux	990
Organisations	57
Déclencheurs	2 608

Tableau 2 : Taille des dictionnaires (en nombre d'entrées)

3.3.2 Le système de cascades créé

Étant donné la diversité de l'état des textes à analyser, nous avons adopté, pour optimiser la maintenance de l'ensemble, un système à trois niveaux. Le premier niveau analyse un texte où les noms propres ne sont pas déjà balisés et le transforme en un texte conforme aux textes déjà balisés par le CESR. Le deuxième niveau augmente le

balisage et le rend conforme à la typologie. Le troisième niveau, plus technique, génère une liste d'entrées à ajouter éventuellement dans les dictionnaires de noms propres (ajouts ensuite supervisés par les experts du CESR). La Figure 5 présente les trois niveaux de cette architecture.

3.4. Les trois premières étapes

3.4.1 La gestion des coupures

Comme cela a été dit à la section 2.1, les textes du corpus respectent la mise en page de l'édition originale et comporte un grand nombre de coupures de mot, ce qui rend l'utilisation des dictionnaires impossibles. La première cascade d'analyse reconnait tout d'abord les balises XML (comme expliqué à la section 0), puis reconstruit les mots coupés par des balises de format, de saut de ligne ou de bas de page. Par exemple, le graphe de la Figure 6 reconstruit les mots séparés par une balise de format (*hi*), lorsque la séparation porte sur la première lettre[10]. On englobe la balise ouvrante dans une expression polylexicale *largerSup* avec un trait indiquant le nombre de lettres pour reconstruire le texte au final ; la balise fermante est placée juste après cette dernière, avant le mot reconstitué.

Cela permet de traiter l'exemple suivant du Tableau 1.

<hi rend="larger">E</hi>N ceste mesme heure

Qui devient tout d'abord :

{<hi rend="larger">,.baliseXml+nomDhi+typeLarger}
{
{</hi>,.baliseXml+nomFhi}
,.largerSup+1}
EN ceste mesme heure

[10] Dans le système Unitex, lorsqu'on est, comme ici, en mode morphologique, l'expression <*MOT*> désigne en fait une lettre...

Tourisme culturel

Puis :

```
<csc>
    <form><hi rend="larger"><\form>
    <code>baliseXml<\code>
    <code>nomDhi<\code>
    <code>typeLarger<\code>
<\csc>
<csc>
    <form>
        <csc>
            <form></hi><\form>
            <code>baliseXml<\code>
            <code>nomFhi<\code>
        <\csc>
    <\form>
    <code>largerSup<\code>
    <code>1<\code>
<\csc>EN ceste mesme heure
```

La cascade de synthèse qui suit reconstruit le texte en ajoutant un attribut *value* dans les balises TEI de coupure pour mémoriser le nombre de lettres à décaler lors de la reconstruction :

```
<hi rend="larger" value="1"></hi>EN ceste mesme heure
```

Les graphes de bas de page et de correction sont légèrement plus complexes, puisqu'il s'agit non seulement de reconstruire (pour les bas de page) ou de choisir le bon mot (pour les corrections), mais aussi de cacher du texte pour la suite de l'analyse. Notons qu'une correction peut se trouver coupée elle aussi par un saut de ligne ou même un bas de page ! Le graphe de la Figure 7, utilisé pour l'expansion des abréviations, ne laisse visible que la forme corrigée, la forme d'origine étant cachée.

Tourisme culturel

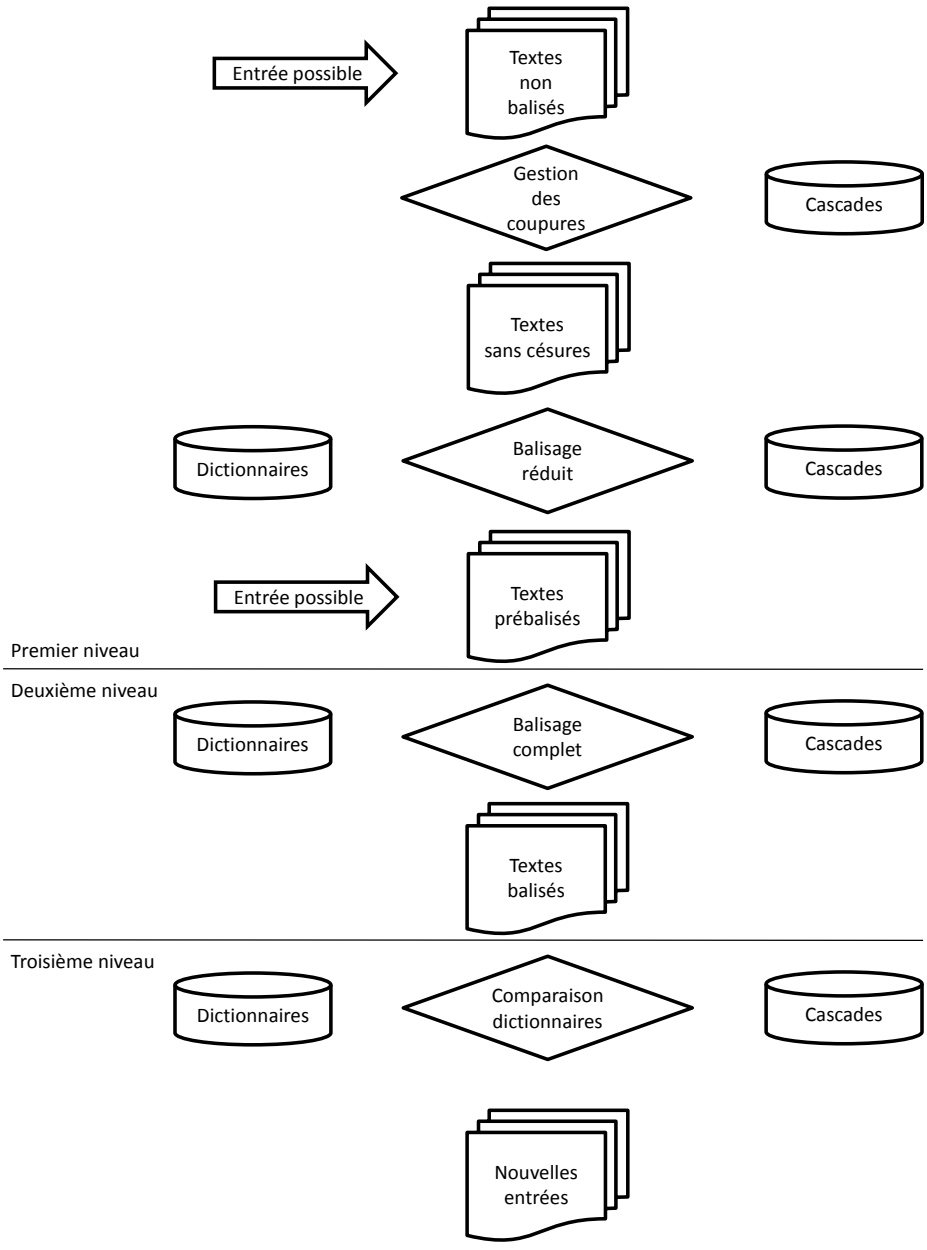

Figure 5 : Les trois niveaux de l''architecture du projet Renom

Tourisme culturel

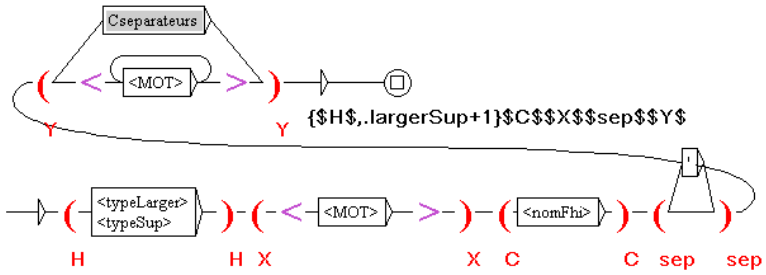

Figure 6 : Le graphe de gestion de la balise *hi*

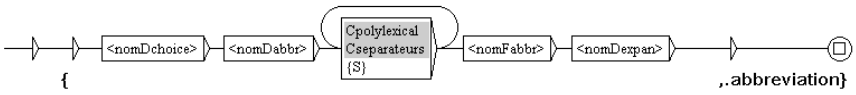

Figure 7 : Un graphe pour gérer les abréviations

Ce graphe agit donc par exemple comme suit :

<choice>
<abbr>PAN.</abbr>
<expan>PANURGE</expan>
</choice>

Devient :

{<choice>,.baliseXml+nomDchoice}
{<abbr>,.baliseXml+nomDabbr}PAN.{</abbr>,.baliseXml+nomFabbr}
{<expan>,.baliseXml+nomDexpan}
PANURGE
{</expan>,.baliseXml+nomFexpan}
{</choice>,.baliseXml+nomFchoice}

Et :

{
{<choice>,.baliseXml+nomDchoice}
{<abbr>,.baliseXml+nomDabbr}PAN.{</abbr>,.baliseXml+nomFabbr}
{<expan>,.baliseXml+nomDexpan}
,.abbreviation}
PANURGE
{</expan>,.baliseXml+nomFexpan}
{</choice>,.baliseXml+nomFchoice}

3.4.2 Le balisage réduit

La deuxième cascade d'analyse commence par l'utilisation des dictionnaires pour reconnaître et baliser les noms propres déjà connus, comme *Gargantua* dans l'exemple ci-dessus. Cependant les graphes de balisage utilise aussi le contexte, car plusieurs

entrées des dictionnaires sont ambiguës, lieu, personne ou organisation. Le graphe de la Figure 8 balise un *persName* dans le contexte d'un déclencheur de type *Militaire* :

<lb/> cens chevaulx legiers soubz la conduicte
<lb/> du capitaine Engoulevent, pour descou
<lb rend="hyphen"/>vrir le pays, & scavoir si embuche aulcune

<lb/> cens chevaulx legiers soubz la conduicte
<lb/> du capitaine <persName>Engoulevent</persName>, pour descou
<lb rend="hyphen"/>vrir le pays, & scavoir si embuche aulcune

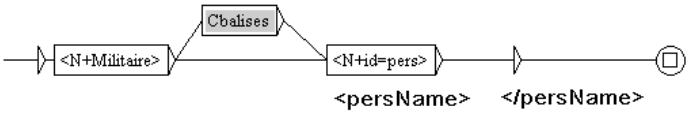

Figure 8 : Un graphe contextuel pour étiqueter les mots des dictionnaires

Lorsque le contexte ne permet pas de désambiguïser, un double balisage est inséré, comme expliqué dans la section 3.2.3. Après le balisage des mots du dictionnaire, des graphes contextuels insèrent des balises lorsqu'un mot commençant par une majuscule suit le contexte. Ces graphes utilisent des contextes similaires aux précédents. Le graphe de la Figure 9 reprend le même contexte que celui de la Figure 8 pour l'appliquer aux mots commençant par une majuscule et non déjà balisé, comme :

<lb/> du chevalereux capitaine Moses

Qui devient :

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

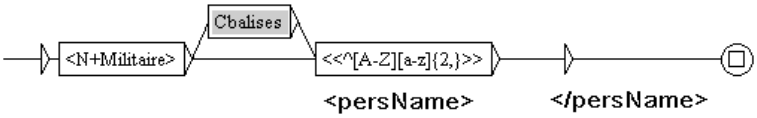

Figure 9 : Un graphe contextuel pour étiqueter les mots commençant par une majuscule

La cascade de synthèse complète la reconnaissance en reconstruisant le texte d'origine, augmenté de balises *persName*, *geogName*, *placeName* et *orgName*. Ces dernières comprennent des attributs uniquement si ceux-ci proviennent directement de la consultation du dictionnaire.

3.4.3 Le balisage complet

La troisième cascade d'analyse opère sur le texte prébalisé avec les cinq balises *name*[11], *persName*, *geogName*, *placeName* et *orgName*. Comme cela a été expliqué, ce texte est le résultat de la deuxième cascade ou provient d'un étiquetage manuel effectué avant le projet par le CESR. Dans un premier temps, les graphes préparent l'analyse du texte en travaillant sur les balises XML-TEI. À chaque fois que cela est possible, ces graphes sont bien sûr mutualisés et sont appelés par plusieurs cascades. Ensuite, la consultation des dictionnaires complète éventuellement les attributs manquants dans les balises (identifiant ou type) et, parfois, les réordonne.

Puis cette cascade gère, d'une part, le balisage interne et les imbrications, et, d'autre part, l'extension du balisage. Enfin, elle crée les identifiants des noms propres reconnus non présents dans les dictionnaires.

Le balisage interne concerne deux types de balises, les *persName*, avec en particulier l'indication du prénom et du nom, et les *geogName*, avec l'introduction de la balise *geogFeat*.

Pour la reconnaissance des prénoms, nous disposons d'une liste dans le dictionnaire de la Renaissance. Il est possible qu'une personne porte plusieurs prénoms (*hugues thierry salel*[12]) ou un nom polylexical (*Jan Trivolse Guallo*) et certains noms comportent une particule à baliser elle aussi (*Ulrich Thierry du Gallet*). Le graphe de la Figure 10 balise un prénom et un nom (dans les sous-graphes *foreName* et *surName*), toujours en prévoyant des balises optionnelles.

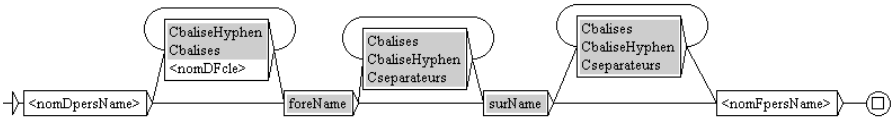

Figure 10 : Le graphe *prénoms-noms*

Pour ce qui est des *geogName*, l'insertion de la balise *geogFeat* est plus simple, puisque nous avons une liste des expansions possibles et que celles-ci se situe toujours à gauche du nom propre. Une fois le balisage interne effectué, les noms propres sont eux aussi considérés comme un seul bloc polylexical, afin de pouvoir gérer les imbrications.

---

[11] Ces balises placées manuellement par le CESR sont aussitôt transformées en balise *persName*.

[12] Ce nom est en minuscule dans le texte.

Dans une imbrication, chaque élément a un identifiant. Le système d'identifiants peut donc se révéler assez complexe à mettre en place. Par exemple, on souhaite construire trois identifiants pour le château du gué de Vede :

```
<placeName key="#loc_chasteauduguedevede" dic="non">
    chasteau du
    <placeName key="#loc_guedevede" dic="non">
        Gue de
        <geogName key="#loc_vede">
            Vede
        </geogName>
    </placeName>
</placeName>
```

Le graphe des imbrications est itératif, afin de reconnaître plusieurs niveaux d'imbrications. Il s'agit en fait d'un graphe qui appelle quatre sous-graphes, chacun consacré à un des quatre types possibles. La Figure 11 présente ce graphe itératif et le sous-graphe qui gère une insertion dans un *placeName*.

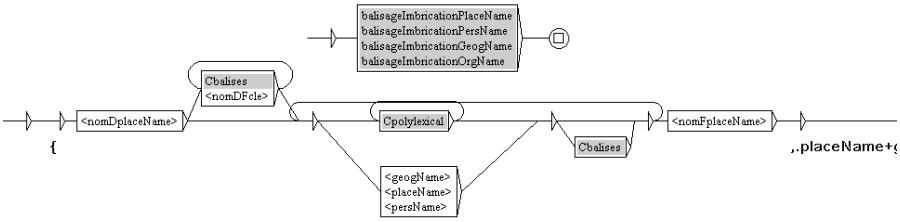

Figure 11 : Le graphe des insertions et le sous-graphe pour un *placeName*

Comme cela a été annoncé à la section 3.2.3, l'extension du balisage correspond à des titres nobiliaires, militaires ou religieux, à des fonctions et civilités honorifiques, ainsi qu'à des précisions familiales et des surnoms. Les noms propres reconnus étant à ce moment considérés comme des expressions polylexicales, il s'agit d'ajouter des balises à gauche et à droite, une balise englobante *persName* et des balises *roleName*, *genName* ou *addName*. Ces balises sont directement insérées sous la forme présentée en section **Erreur ! Source du renvoi introuvable.**. Le graphe de la Figure 12 étend le balisage à gauche d'un *persName*, comme dans l'exemple qui suit.

```
tu as
<lb>ton precepteur
<persName key="#pers_epistemon" dic="non">
    Epistemon
<persName>
dont
<choice>
    <orig>lun<orig>
    <reg>l'un<reg>
<choice>
par
<lb>vives &amp;amp; vocables instructions
```

Tourisme culturel

tu as
<lb>ton
<persName key="#pers_epistemon" dic="non">
<roleName type="fonction">
precepteur
<roleName>
Epistemon
<persName>
dont
<choice>
<orig>lun<orig>
<reg>l'un<reg>
<choice>
par
<lb>vives & vocables instructions

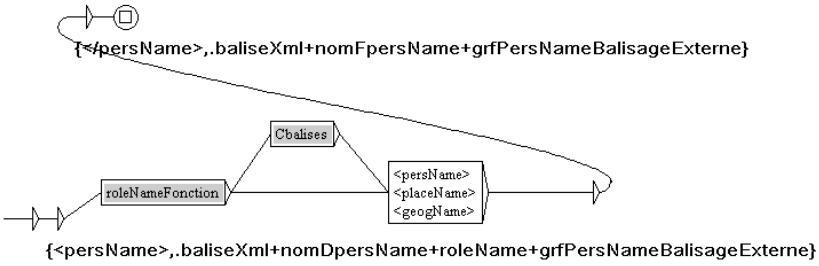

Figure 12 : Graphe d'extension du balisage à gauche d'un *persName*

4. Évaluation

L'évaluation de notre système a été réalisée sur les deux ouvrages de Pierre de Ronsard présents dans notre corpus, le *Voyage de Tours* et l'*Élégie sur les troubles d'Amboise*, documents qui n'ont pas été utilisés pour construire les cascades. Rappelons que l'essentiel de notre corpus était composé d'ouvrages de François Rabelais. Pour cette évaluation, nous avons utilisé les mesures classiques du rappel et de la précision (Figure 13).

$$Rappel = \frac{\text{nombre de noms propres correctement détectés}}{\text{nombre de noms propres réels}} = 80{,}9\%$$

$$Précision = \frac{\text{nombre de noms propres correctement détectés}}{\text{nombre de noms propres détectés}} = 77{,}9\%$$

Figure 13 : Rappel et précision

Donnons quelques exemples d'erreurs:

– Nom propre non reconnue : *Iesus* apparaît sans contexte ci-dessous et n'était pas présent dans le dictionnaire :

```
<l>Du sainct nom, ou Iesus en la croix attaché,</l>
```

– Erreur de reconnaissance : *Sinon* est un nom de personne dans notre dictionnaire de la Renaissance ; pour éviter un trop grand nombre d'erreurs, les mots ambigus avec un mot du vocabulaire commun ne sont reconnus qu'avec une majuscule initiale. Mais ici, il apparait en début de ligne, avec une majuscule, car il s'agit de vers (balise *l*) :

```
<l>Sa puissance est cruelle, &amp;amp; n'a point d'autre
    <lb/><space unit="mm" quantity="4"/>
    <choice><orig>ieu</orig><reg>jeu</reg></choice>,</l>
<l><persName key="#persm_sinon">Sinon</persName>
    de rebrusler nos coeurs à petit feu,</l>
```

– Erreur de type : *Antioche* et *Sydon* sont étiquetés *persName* au lieu de *placeName* ; l'erreur provient de la présence de *Antioche* uniquement comme *persName* dans le dictionnaire (erreur corrigée) :

```
<l>Gaigner la
    <placeName type="pays" key="#loc_palestine">Palestine</placeName>
    , &amp;amp; toute l'Idumee,</l>
<l>Tyr,
    <persName key="#pers_sydon" dic="non">Sydon</persName>,
    <persName key="#persf_antioche">Antioche</persName>
     &amp;amp; la ville nommee</l>
```

– Erreur de frontière : l'extension du balisage au mot *vallee* n'a pas été réalisée, à cause d'une mauvaise description de la correction (erreur corrigée) :

```
<l rend="positif">Quel passe-temps prens-tu d'habiter la
    <lb/><space unit="mm" quantity="6"/>
    <choice><orig>uallee</orig><reg>vallee</reg></choice></l>
<l>De
    <placeName type="batiment" key="#loc_bourgueil">Bourgueil</placeName>,
    où [...]</l>
```

5. Conclusion

Nous avons présenté l'ajout d'un module à la chaîne de transcription des textes de la Renaissance utilisée au CESR, module chargé de la recherche des noms propres dans des textes de la Renaissance, balisés dans un format XML-TEI conservant la présentation exacte des éditions originales. La particularité de ce corpus nous a conduits à adopter une méthodologie adaptée. Pour cela nous avons conçu des dictionnaires de noms propres et de mots déclencheurs contenant de multiples formes d'écriture et des cascades de graphes Unitex permettant de traiter d'une part des textes sans aucun repérage manuel de noms propres et des textes où les noms propres sont déjà partiellement marqués. Après chaque traitement d'un ouvrage, les dictionnaires de noms propres sont augmentés après supervision d'un expert. Les experts du CESR

ont considéré nos résultats comme tout-à-fait satisfaisants et permettant un précieux gain de temps dans l'annotation. Notre système est donc aujourd'hui opérationnel chez eux et sert à l'enrichissement des textes qu'ils transcrivent.

Le site internet créé permet aux touristes visitant un château ou un hameau cité dans les œuvres de Rabelais de retrouver dans le texte original les mentions qu'en fait l'auteur. Inversement, les cartes générées sur ce site à partir des toponymes du texte les invitent à se déplacer en Touraine sur les traces de Gargantua.

Ce travail a été financé par le programme de recherche d'intérêt régional de la Région Centre. Les auteurs remercient les collègues du CESR, particulièrement Sandrine Breuil, Marie-Luce Demonet, Jorge Fins et Marie Olivron.

6. Références

Abney S. Parsing By Chunks. In *Principle-Based Parsing*, pp. 257-278, Kluwer Academic Publishers. 1991.

Friburger N., Maurel D. Finite-state transducer cascade to extract named entities in texts, *Theoretical Computer Science*, 313:94-104. 2004.

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

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1):69-96.

Nadeau N., Sekine S. *A survey of named entity recognition and classification*, Satoshi Sekine and Elisabete Ranchhod, ed., John Benjamins publishing company, 3-28. 2009.

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