Обогаћивање ренесансних текстова властитим именима

УДК: 81'322.2:004.9

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

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

Ирис Ескол Таравела
iris.eshkol@univ-orleans.fr
Université d'Orléans, Laboratoire ligérien de linguistique, UMR 7270

АПСТРАКТ: Циљ пројекта Реном је да обогати ренесансне текстове властитим именима. Ови текстови представљају два изазова: велику разноврсност услед различитог записивања речи; претрпаност великим бројем XML-TEI етикета које су уведене да би се сачувао тачан изглед оригиналног издања. Наш задатак се састојао од додавања етикета именованих ентитета овом формату за имена која нису већ била обележена и за контекст са њихове леве стране, а понекад и са десне стране. У ту сврху побољшали смо бесплатан програм у отвореног кода CasSys да бисмо анализирали текстове са Unitex-овим каскадама графова и направили смо посебне речнике и каскаде. Евалуација је показала да је стопа грешке била 6,1%. Ренесансни текстови обогаћени на овај начин користе се на веб-сајту који обједињује хуманистичке науке и туризам тако што омогућавају навигацију по мапама преко имена која се на њима налазе.

КЉУЧНЕ РЕЧИ: именовани ентитети, ренесансни текстови, каскаде графова, CasSys, хуманистичке науке и туризам

ДАТУМ ПРИЈЕМА РАДА: 4. мај 2014.
ДАТУМ ПРИХВАТАЊА РАДА: 14. септембар 2014.

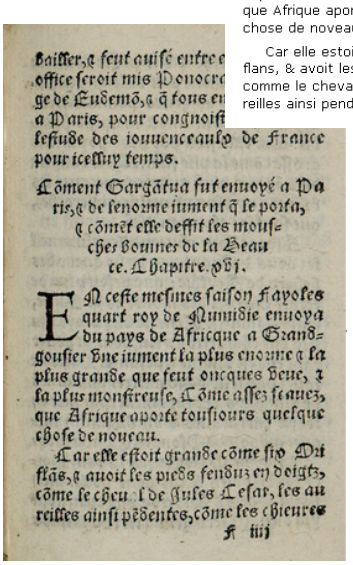
1. Мотивација

Већ више од десет година Центар за високо образовање у области ренесансе (Centre d'Études Supérieures de la Renaissance - CESR)[1] нуди на интернету Хуманистичке електронске библиотеке (BVH): многобројна дела Раблеа, Ронсара и других ренесансних писаца представљена су као скениране и транскрибоване књиге. Транскрипција користи TEI формат који строго прати изглед скенираног оригинала, дакле очувани су пасуси, подела на редове, скраћенице, цртице, изглед слова итд. Слика 1 представља један пасус (транскрибован и скениран текст) романа *Гаргантуа* од Раблеа као што је приказан на веб-сајту. Исти пасус етикетиран у TEI формату је дат у следећем примеру (--- означава пасус, а <lb/> означава нови ред).

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

```
<p>
<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
&amp; la
<lb/>plus grande que feut oncques veue, &amp;
<lb/>la plus monstreuse, Comme assez scavez,
<lb/>que Afrique aporte tousjours quelque
<lb/>chose de noveau.
</p>
```

Рабле је писао о Туренској области (у Француској) где су се доселили дивови Гаргантуа и Пантагруел Грангузије. Пројектом *Ренoм*[2] смо планирали да развијемо такозвани „литерарни туризам“ тако што ћемо повезати имена и књиге: посетиоци веб-сајта могу да се крећу у роману помоћу властитих имена и да виде где су измишљена или древна места била „лоцирана“. Слика 2 показује измишљену Телемску опатију између градова Шинон и Азе ле Ридо, у близини Тура. Туристи се подстичу да посете замкове у тим градовима и Раблеов музеј близу тог 'места'.

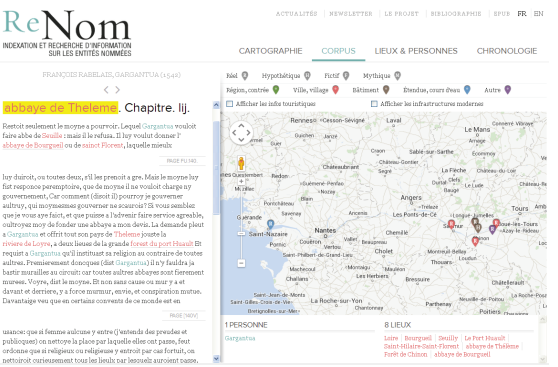


EN ceste mesmes saison Fayoles
quart roy de Numidie envoya
du pays de Africque a Grand-
gousier une jument la plus enorme & la
plus grande que feut oncques veue, &
la plus monstreuse, Comme assez scavez,
que Afrique aporte tousjours quelque
chose de noveau.

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

СЛИКА 1: Део BVH веб-сајта

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



СЛИКА 2: Веза између текста и мапе на веб-сајту *Ренoм*

Циљ овог научног пројекта је да обогати текстове етикетама које указују на имена особа, значајних личности и локација помоћу надгледаних техника за обраду природног језика (Natural Language Processing – NLP). Овај задатак, познатији као препознавање именованих ентитета (Named Entity Recognition – NER) добро је дефинисан у области обраде природних језика, и на њему се ради већ две деценије. У нашем пројекту, постојећи TEI формат, прекомерно коришћење цртица и велика разноврсност у правопису властитих имена (и у правопису речи у целини) представљају прави изазов. На пример, у одломку на слици 1, име Грангузије (*Grandgousier*) пише се *Grand-<lb rend="hyphen"/>gousier* а назив Африка се пише на два начина, *Africque* и *Afrique*.

Као што је успостављено на MUC конференцијама (*Message Understanding Conferences*), именовани ентитети се односе на имена особа, називе локација, називе организација, датуме, проценте, валуте (Chinchor, 1997) и понекад титуле, сате, професије итд. Достигнућа у овој области су представљена код (Nadeau, Sekine, 2009); главна идеја је да се користе унутрашњи и спољашњи докази (MacDonald, 1996), тј. локални контекст. На пример, у *Иī Тијери Салел Стариjу, īосūодар Севиље* (Hugues Thierry Salel l'ainé, seigneur de Seuille), прво име *Иī Тијери* указује да је *Иī Тијери Салел* име особе (унутрашњи доказ) а титула *īосūодар* (seigneur) указује да је *Севиља* (Seuille) топоним (спољашњи доказ).

За решавање NER задатка могућа су три приступа: машинско учење, симболична правила,

а понекад и хибридни приступ. Техникама машинског учења потребан је корпус за обучавање који није доступан за наш задатак. Из тог разлога користили смо приступ заснован на симболичним правилима (Ait-Mokhtar et Chanod, 1997; Hobbs et al., 1997). Да би се успоставила сарадња између информатичара, лингвиста и стручњака за ренесансу, изабрали смо Unitex платформу[3] (Paumier, 2003) због погодног интерфејса (са коришћењем графова) и бесплатне лиценце. Са Unitex-ом можемо да дефинишемо каскадне системе засноване на правилима (CasSys мени), који могу да користе многобројне могућности Unitex графова. Каскаде (Abney, 1991) су коришћене у многим NLP апликацијама, као што је рашчлањивање (Abney, 1996), синтаксичка анализа (Kokkinakis, 1999), морфолошка анализа (Alegria et al., 2001) итд. Наш систем је инспирисан радом (Friburger, Maurel, 2004).

NLP заједница је заинтересована за древне језике и древне облике савремених језика (погледајте на пример Denooz, Rosmordus, 2009). Старофранцуски правопис није строго утврђен и постојало је много варијаната имена. За средњефранцуски (мало пре ренесансног француског) развијени су посебни алати и речници (Souvay, 2004) као и лематизери (Souvay, 2007). За реализацију нашег задатка било је потребно да и ми направимо посебне речнике и каскаде што смо и учинили у сарадњи са стручњацима за ренесансу. Циљ овог пројекта је да помогне стручњацима да прецизно етикетирају текстове и поправе речнике. Ови речници садрже властита имена и њихове варијанте повезане јединственим кључевима (предложио их *Реном* а одобрили стручњаци), као и локације такође повезане јединственим кључевима са базом података GeoNames.[4] Веб-сајт користи овај показивач да их означи на мапи. Када премале или старе локације нису биле доступне у бази GeoNames, стручњаци су у ту базу додавали нове одреднице.

2. Презентација корпуса

Корпус се састоји од једанаест књига:

– *Фантастични говори* (*Les discours fantastiques*) (издање из 1566), Јустин Тонелие (*Justin Tonnelier*);

– *Дворјанин* (*Courtisan*) (издање из 1538), Балдасаре Кастиљоне (Baldassare de Castiglione);

– *Путовање у Тур* (*Voyage de Tours*) (издање iz 1560) и *Елегија о Амбоазовим невољама* (*Élégie sur les troubles d'Amboise*) (издање из 1563), Пјер де Ронсар (Pierre de Ronsard);

– *Гаргантуа* (издање из 1542), *Пантагруел* (издање из 1542), *Трећа књига* (le *Tiers Livre*) (издања из 1546. и 1552), *Четврта књига* (le *Quart Livre*) (едиције из 1548. и 1552) и Кратка изјава (*Brève déclaration*) (издање из 1552), Франсоа Рабле (François Rabelais).

Формат је веома специфичан: као што смо објаснили у првом делу, поштује се целокупан изглед оригиналног издања, подела на редове, футери, почетна слова итд. Понекад су и они који су транскрибовали додавали неке корекције. Неки примери овог формата су:

- Почетно слово (Пантагруел)

  <lb/><hi rend="larger">P</hi>Antagruel quelque jour pour se

- Преношење краја првог ред на крај другог реда, после заграда, иза латинског цитата у истом реду - транскрибовано на три линије.[5]

  <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>

- Футер (реч 'Dace' је прекинута бројем странице – 188)

  <lb rend="hyphen"/>
  bek Norwerge, Sweden, Rich, Da-
  <lb/>
  <lb/>
  <lb/>
  <pb n="188" xml:id="_Page_188"/>
  <lb/>

3 http://www-igm.univ-mlv.fr/~unitex/

4 http://www.geonames.org/

5 Транскрибовани текст у оригиналу гласи: 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-
```

- Корекција (уметање апострофа)

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

Заграде именованих ентитета морају да садрже све заграде форматирања. На пример, последњи пример постаје:

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

3. Коришћена типологија

CESR је користио TEI формат за транскрибоване текстове, тако да је било природно да се усвоји иста TEI типологија. Користили смо четири типа: географска имена (*geogName*), називе места (*placeName*), називе организација (*orgName*) и имена особа (*persName*).

3.1 Географска имена и називи места

Географска имена су подељена у два подтипа, геониме (планине, равнице, висоравни, пећине…) и хидрониме (океани, мора, реке, језера, баре…) за које је коришћена посебна вредност атрибута *type* етикете *geogName*. Ако је било могуће, географске одреднице су додатно прецизиране додавањем посебне унутрашње етикете *geogFeat*.

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

Називи места су некад такође подељени у подтипове (градови, земље, имања и зграде) коришћењем истог атрибута type.

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

Две локације су понекад уметнуте једна у другу.

```
<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 Називи организација

Организације су подељене на три подтипа: људи, имања и заједнице. CESR је одабрао да не повезује организације са кључевима.

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

Организације су такође могле да буду уметнуте једна у другу.

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

Када је било тешко да одлучимо да ли се име односи на место или на организацију, уносили смо обе етикете, *placeName* и *orgName*.

```
<placeName type="building"
key="#loch_coingnaufondabbaye">
<orgName type="community">abbaye de
```

<placeName type="city"
key="#loch_coingnaufond">
*Coingnaufond</placeName></**orgName**></**placeName**>*

3.3 Имена особа

У најједноставнијим примерима коришћене су само етикете *persName* са својим кључевима .

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

Где је требало додавали смо унутрашње етикете за лична имена (*foreName*), презимена (*surName*) и партикуле (*nameLink*).

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

Коначно, ове етикете су проширене титулама или улогама (*roleName*) које су разврстане у подтипове: племићка улога, црквена улога, функција или занимање, почасни положај. Када је у титулу укључен назив места, такође је етикетиран: на пример, *їосūодар Ecapca* (seigneur des Essars) је особа, али *Ecapc* је место:

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

Понекад смо додатно спецификовали лична имена када би надимци или улоге унутар породице били поменути (у наредном примеру старији син (l'ainé)):

<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>

Текстови су били и двосмислени. Ако је било могуће, стручњаци су бирали добру интерпретацију. У следећем примеру, *saint Martin de Candes* може да се односи и на цркву и на особу:

*<**persName key="#pers_sainctmartindecandessaintmartin"**><**placeName key="#loc_sainctmartindecandessaintmartin"**>sainct <lb/>Martin de <placeName type="city"*
*key="#loc_candessaintmartin">Candes</placeName></**placeName**></**persName**>*

4. Речници

Као што смо већ поменули, често нам је било потребно да NER препозна контекст именованог ентитета. Због тога смо направили ортографски речник варијанти, проучавајући контекст и користећи листу са старим личним именима.

На пример, у ренесанси се реч 'капетан' писала *capitaine*, *capiteine* или *cappitaine*. Ми смо изабрали синхрону одредницу за лему и додали смо својства која NER може да користи (видети 6.3 и 6.4):

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

На пример, други ред садржи пет информација: облик (*capiteine*), лему (*capitaine*), врсту речи (*N*), својство (*Military*) и морфолошке кодове (*ms*).

Три CESR листе имена које се односе на особе, организације и локације смо претворили у формат Unitex речника (представљен на почетку поглавља), које смо побољшавали након анализе сваке књиге. У произведеним речницима именима, реч има за лему свој кључ. Ови кључеви се користе да повежу различите начине записивања имена, а такође и да повежу локације са базом GeoNames:

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

У претходним радовима, CESR стручњаци су изабрали да користе експлицитне кључеве, то јест да као кључеве користе саме називе (као што је *loc_ancenis* за топоним *Ancenis*) па смо и ми морали да користимо исти приступ. Табела 1 представља број уноса у речницима на крају пројекта *Реном*.

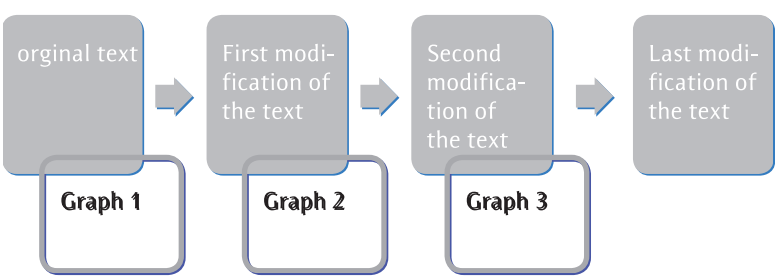
1.145	Особе
987	Локације
57	Организације
2.622	Друге речи

ТАБЕЛА 1: Број уноса у речницима

5. Побољшања Unitex платформе

Као што смо раније напоменули, изабрали смо Unitex платформу да бисмо олакшали сарадњу између информатичара, лингвиста и стручњака за ренесансу. Unitex је бесплатна платформа у отвореном приступу (LGPL лиценца). Unitex омогућава свакоме да парсира текстове сопственим речницима (видети поглавље 4), да пише лингвистичка правила у облику графова користећи се врло погодним интерфејсом, а може и да прави каскаде графова са CasSys-ом.

Каскада графова је низ графова за парсирање: први граф парсира текст, други парсира текст који је модификован првим графом и тако даље.



СЛИКА 3. Принцип рада каскаде графова

За потребе пројекта *Рεном* морали смо да унапредимо CasSys додавањем три својства која ће бити објашњена у следећим пододељцима: итерација, односно понављање примене графа до одређене фиксне тачке, омогућавање употребе Unitex-ових морфолошких речника и стварање адекватнијих излазних датотека каскаде.

У Unitex-у, графови парсирају текст са могућношћу убацивања нових секвенци, замењивања других, коришћења променљивих, померања секвенци и уметања података из речника.

5.1 Итерација графова

Додали смо у CasSys могућност итерације графова до одређене фиксне тачке: итеративни граф парсира текст и ствара модификован текст, потом парсира овај модификовани текст и тако даље све док парсирање више не модификује текст настао као резултат претходног парсирања (то је та фиксна тачка). Итеративне графове користили смо, пре свега, да направимо кључеве типова који су уметнути једни у друге (видети поглавље 6.5).

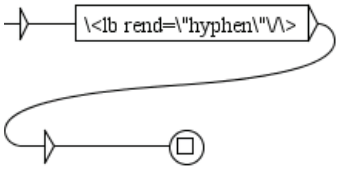
5.2 Unitex морфолошки речници

За облик који се појављује у тексту Unitex граф може да извуче информације које су му придружене у речницима: његову лему, својства или морфолошке кодове. Ови речници се зову „морфолошки речници“.

Додали смо могућност укључивања ове функције у каскаде. Дефинисали смо наша три речника имена (видети поглавље 4) као „морфолошке речнике“ и то нам је омогућило да повежемо име које смо пронашли у неком од речника са његовим кључем. Више информација дајемо у поглављу 6.5.

5.3 Излазна датотека каскаде

Основна идеја код парсирања текста каскадом графова је да се већ етикетирани део текста посматра као израз који се састоји од више речи (енгл. *multiword expression – MWE*) јер тада наредни графови каскаде више не могу да га парсирају.



СЛИКА 4. Граф који препознаје XML етикету lb[7]

,.baliseXml+nomDFlb+typeHyphen+grfbaliseXmlAttribut}

Unitex систем препознаје секвенцу карактера као MWE ако се стави између витичастих заграда.

На пример, XML етикета *<lb rend="hyphen"/>*[6] ће бити протумачена као MWE ако се стави између витичастих заграда:

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

То ради граф који је представљен на слици 4.

Додатне информације унутар витичастих заграда (после зареза и тачке) омогућавају следећим графовима каскаде да анализирају MWE са Unitex изразима као што су *<BaliseXML>* (за XML етикету), *<DFlb>* или *<hyphen>*, у зависности од потребног степена прецизности.

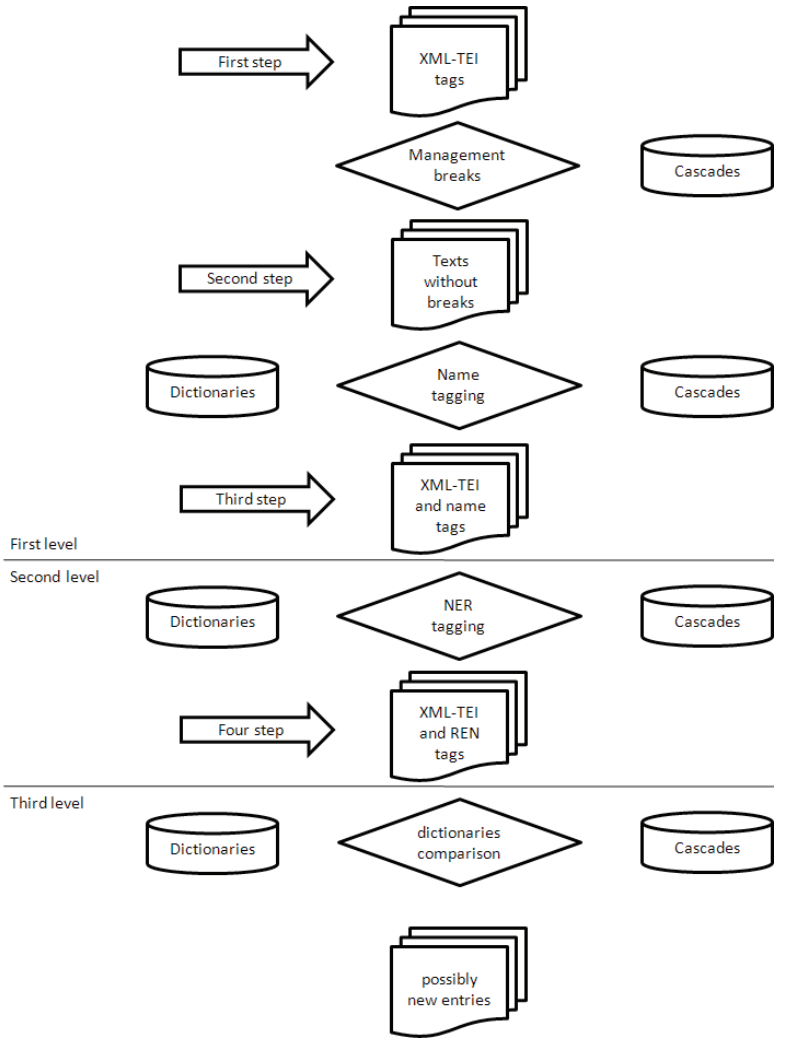
Мана је што се текст који се добија тешко чита. Због тога смо имплементирали посебан CasSys XML формат који се користи као алтернативни излаз из каскаде:

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

6. Метода

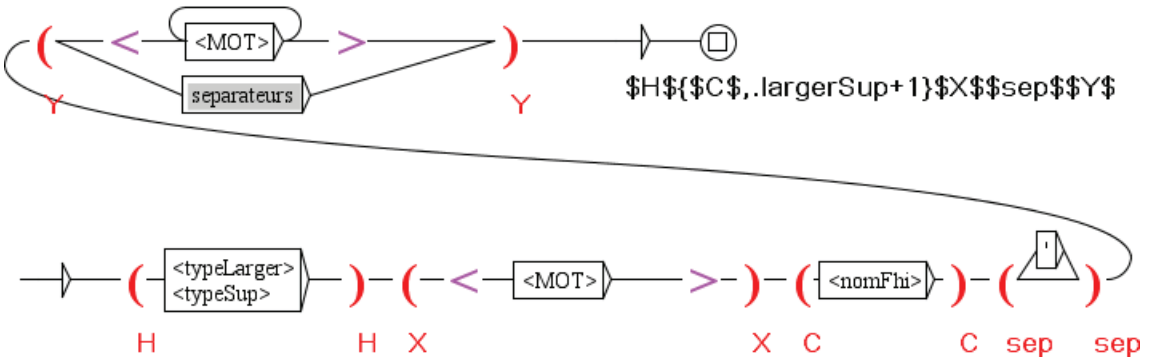
Наш корпус се састоји од две групе текстова. Прва група је већ имала етикете за властита имена (*persName*, *geogName* и *placeName*) које су стручњаци ручно додали; наш задатак на овим текстовима био је да додамо унутрашње етикете (*geoFeat*, *foreName*,

Из тог разлога наше каскаде увек долазе у пару: прва се користи за парсирање текста док друга трансформише посебан CasSys XML формате (резултат прве каскаде) у тражени формат. Први граф каскаде трансформише све оригиналне XML етикете у MWE.



СЛИКА 5. Организација рада у четири етапе

6 lb (line break) је TEI етикета која означава почетак новог реда у неком издању текста.
7 Као и сви остали графови, и овај граф додаје као једно својство своје сопствено име, што помаже у отклањању грешаке у каскади.



СЛИКА 6. Граф за реконструисање речи чије је почетно слово веће

lastName…), да пронађемо кључеве у речницима и да проширимо имена на именоване ентитете додавањем етикета попут *roleName* (господар, опатија…), *genName* (старији син…) и тако даље. У другој групи текстова само је форматирање било анотирано, што је значило да смо прво морали да препознамо имена у овим текстовима, да би потом урадили исти посао као и са првом групом текстова.

Посао смо организовали у четири етапе (видети слику 5). Неколико већ етикетираних текстова је парсирано тек од треће етапе.

- Припремна обрада ради реконструкције одсечених имена на крају реда или стране;
- Претраживање речника и употреба контекстних правила за етикетирање имена;
- Консултовање речника и примена интерних и проширених правила, као што је раније објашњено (firstName наспрам genName и тако даље);
- Вађење имена која нису у речницима.

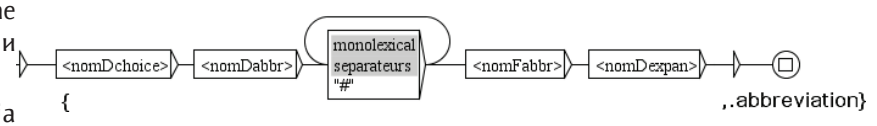
6.1 Цртица

Као што смо истакли, много цртица се појављује у текстовима. Први граф је препознао XML етикете као изразе који се састоје од више речи (MWE) и доделио им као врсту речи (POS) *baliseXML*. Остали графови прве каскаде реконструишу речи. На пример, граф на слици 6 препознаје слова већа од других: *hi*[8] етикета сече реч, а граф је реконструише. Две *hi* етикете (почетна и завршна етикета) постају нови MWE чија врста речи постаје *largerSup*, а његов нови атрибут памти опсег истакнуте секвенце (овде, *value="1"* јер је у питању слово).

На пример:
<hi rend="larger">E</hi>N ceste mesme heure
постаје
{<hi rend="larger" value="1">
</hi>,.largerSup}EN ceste mesme heure

6.2 Скраћенице

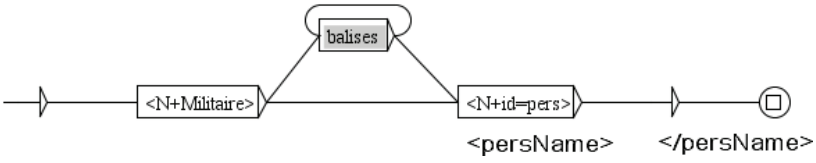
Граф представљен на слици 7 препознаје XML етикете *choice*, *abbr* and *expan*,[9] и гради *MWE*[10] са врстом речи abbreviation (скраћеница).



СЛИКА 7. Граф за сакривање скраћеница

На пример:
<choice><abbr>PAN.</abbr>
<expan>PANURGE</expan></choice>
постаје
{<choice><abbr>PAN.</abbr>
<expan>,.abbreviation}PANURGE</expan></choice>

8 hi (highlighted) је TEI етикета која служи за означавање делова текста који су графички истакнути у односу на преостали текст.
9 choice, abbr and expan су TEI етикете које се користе за скраћенице: choice окупља алтернативна кодирања за неки део текста, abbr (abbreviation) означава било коју врсту скраћенице, expan (expansion) означава текст од кога је скраћеница настала.
10 Овај MWE сакрива оригиналну реч PAN која се више неће парсирати. Овде је то важно јер је реч двозначна и представља и име грчког митског бога Пана.



СЛИКА 8. Граф за додавање етикете *persName* именима пронађеним у речницима која се појављују у војном контексту

6.3 Имена препозната из речника

Друга етапа је почела претраживањем речника да би се етикетирала имена која се налазе у речнику. Нека имена су била вишезначна, па смо користили контекст да бисмо утврдили да ли се ради о имену особе, локације или организације. Граф на слици 8 додаје етикету *persName* личним именима у војном контексту.

На пример:

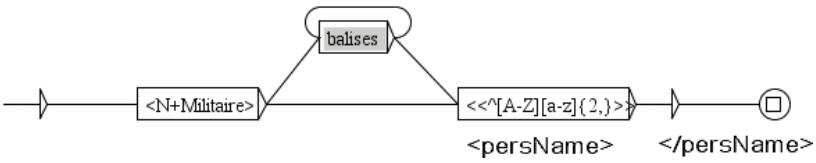
<lb/> du capitaine Engoulevent, pour descou

постаје

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

6.4 Имена препозната само на основу контекста

Након што смо препознали имена из речника, искористили смо исте контексте да етикетирамо имена која нису у речницима, уколико им је прво слово велико. Граф на слици 9 етикетира са *persName* имена која нису препозната, а јављају се у војном контексту.



СЛИКА 9. Граф за етикетирање са *persName* непрепознатих имена са првим великим словом ако се јављају у војном контексту

На пример:

<lb/> du chevalereux capitaine Moses

постаје

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

6.5 Кључеви

Када смо идентификовали имена, тражили смо њихове кључеве у речницима, уколико су такви уноси постојали. Граф на слици 10 враћа ове кључеве у облику атрибута *key* одговарајуће почетне етикете.

На пример:

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

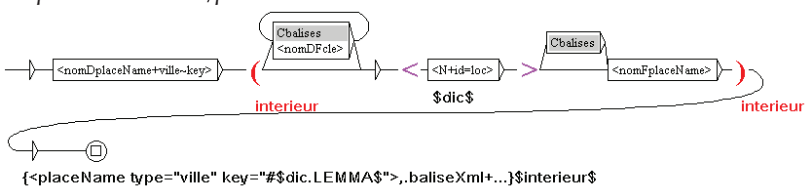
постаје

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

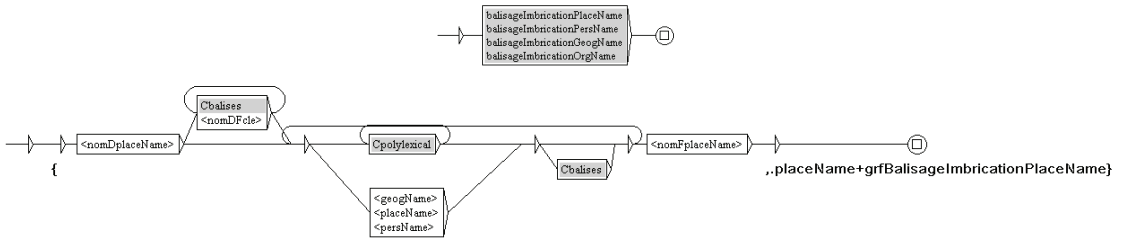
Ако име није пронађено у речницима, други графови граде његов могући кључ спајањем кључева имена од којих се оно састоји. Атрибут *dic="no"* је стручњацима указивао да име (са оваквим правописом) није пронађено у речницима. Стручњак би га додао, са другим кључем ако је то варијантни облик неког постојећег уноса или са овим кључем ако је то заиста нови унос.

Прављење кључа није било једноставно због имена која су уметнута једна у друга. На пример, морали смо да додамо три кључа за име *château du gué de Vede*: један за властиту именицу *Vede* (кључ је извађен из речника), један за *gué de Vede* и један за *château du gué de Vede*:

*<placeName **key="#loc_chasteauduguedevede"** **dic="no"**>chasteau du*



СЛИКА10. Граф које тражи кључеве у речнику



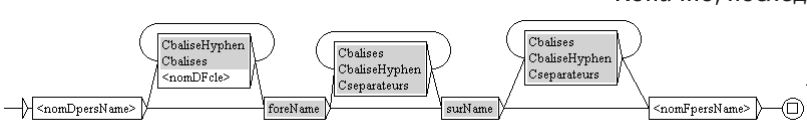
СЛИКА 11. Итеративни граф за уметнута имена

*<placeName **key=”#loc_guedevede” dic=”no”**>*
*Gue de<geogName **key=”#loc_vede”**>Vede</geogName>*
</placeName></placeName>

Граф који прави кључеве уметнутих имена примењује се итеративно; позива четири подграфа, по један за сваки тип. Слика 11 представља овај граф и један од његових подграфова који уноси кључ у етикету *placeName*.

6.6 Етикете унутар имена

Унутар етикета имена смо унели неке етикете: етикете за лична имена и презимена унутар *persName* и *geogFeat* етикете унутар *geogName*. За овај задатак користили смо речник ренесансних имена. Следећи случајеви су нам задавали потешкоће: особа са више личних имена (*Hugues Thierry Salel*), са презименима која се састоје од више речи (*Jan Trivolse Guallo*) или имају партикуле (*Ulrich Thierry du Gallet*). Граф на слици 12 етикетира једно лично име и једно презиме (етикете су у подграфовима). После примене овог графа и имена и њихове унутрашње етикете се посматрају као MWE.



СЛИКА 12. Један од графова за додавање етикета за лична имена и презимена

6.7 Проширене етикете

Исто тако смо проширили *persName* етикете на именоване ентитете који садрже етикете *roleName* (господар, опатија, учитељ…), *genName* (старији син…) или *addName* (надимци). Додали смо нове етикете са леве или са десне стране већ препознатог личног имена и померили смо *persName* етикете. Граф на слици 13 додаје етикете са леве стране именованог ентитета када му претходи roleName. У овом случају кључ се не мења.

На пример, *Epistemon* је лични учитељ (précepteur), па секвенца

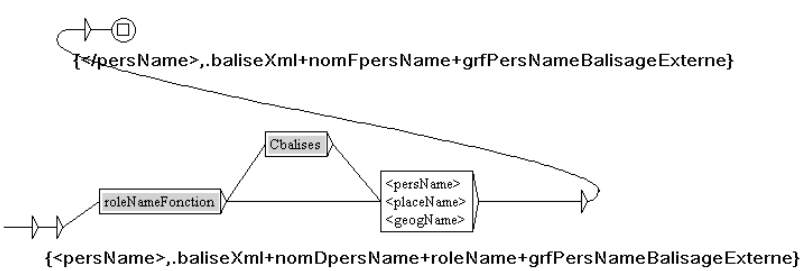
<lb>ton precepteur<persName key=”#pers_epistemon” dic=”no”>Epistemon<persName> don't

постаје

<lb>ton<persName key=”#pers_epistemon” dic=”no”>
<roleName type=”function”>precepteur<roleName>Episte
mon<persName>

6.8 Побољшање речника

Коначно, последње две каскаде су направиле нову датотеку за унапређење речника: обрисале су сав текст, осим имена која нису већ била у речницима или су била у њима али са другачијим својствима. На



СЛИКА 13. Граф за етикетирање са *roleName* са леве стране имена

пример, последњи пример:

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

постаје

Epistemon #pers_epistemon

Целокупна листа имена са карактеристикама dic="no" је прослеђена стручњацима да побољшају речнике.

7. Евалуација

Да бисмо оценили наш посао анализирали смо две књиге Пјера де Ронсара из нашег корпуса (*Путовање у Тур и Елеији о Амбоазовим невољама*). Ове књиге нисмо користили за прављење наших каскаде – за њихово развијање углавном смо користили Раблеове књиге. Израчунали смо пондерисани облик стопе грешке слота (*slot error rate* - SER)[11] (Makhoul et al., 1999) коришћен у француској кампањи евалуације. SER прави разлику између три типа грешака:

- уметање (I - тежина 1): етикетирали смо речи које нису имена.
- избацивање (D - тежина 1): пропустили смо да етикетирамо име.
- етикете са граничним грешкама: погрешан тип (T - тежина 0.5), етикета изван или унутар властитог имена (E - тежина 0.5) или оба (TE - тежина 1).

Ако је #R збир ентитета референтних текстова, SER се израчунава формулом:

$$SER = \frac{\#1+\#D+0,5 * \#E+\#TE}{\#R}$$

Са овако пребројаним вредностима, ако је #S збир уочених ентитета, можемо да израчунамо

#I	#D	#T	#E	#TE	#S	#R
5	19	3	3	0	136	150
SER					6,1%	
Прецизност					96,3%	
Одзив					87,3%	
Прецизност препознавања типа					94,1%	
Прецизност одређивања међе					94,1%	

ТАБЕЛА 2. Евалуација

и прецизност и одзив у нашем раду:

$$Precision = \frac{\#S - \#1}{\#S} \quad \text{и} \quad Recall = \frac{\#S - \#1}{\#R}$$

Етикетирани текстови су у потпуности надгледани, па су граничне грешке мање важне. Ипак, можемо да израчунамо и прецизност тачног препознавања типова и прецизност тачног одређивања међа именованих ентитета:

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

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

Резултати су представљени у табели 2.

Стручњаци CESR-а су желели да прочитају цео корпус пре објављивања на сајту. SER од 6,1% указује на то да је њихов рад заиста унапређен. Значајан број грешака избацивања кореспондира са великим бројем имена без контекста која нису представљена у речницима. На крају су стручњаци унапредили речнике.

8. Закључак

У овом раду смо представили NER задатак примењен на ренесансне текстове у XML-TEI формату. Формат корпуса и значајне варијације вокабулара су захтевале да корпус третирамо на другачији начин него савремене текстове. Користили смо речнике и каскаде засноване на правилима и добили смо 6,1% SER-а. Наш систем ће ући у производну линију транскрибованих текстова.

Најзначајнији Раблеови текстови су сада доступни на веб-сајту пројекта *Рeном* који има претраживач за имена. Користи кључеве да повеже варијантне облике имена и да их повеже са местима на мапи.

11 Ова мера директно комбинује различите врсте грешака. Свака етикета може да има један или више 'слотова'. Именовани ентитети, према MUC имају два 'слота': тип и опсег..

9. Захвалност

Овај пројекат подржава Истраживачки програм региона Центар. Аутори захваљују CESR, нарочито Сандин Брej, Мари-Лис Демоне, Жорж Фен и Мари Оливрон.

Референце:

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.