Projet

Infom@gic

Extraction d'Entités Nommées par les Graphes d'Unitex

Institut d'électronique et d'informatique Gaspard-Monge

INTRODUCTION

- Grande quantité d'informations non structurées sur Internet

  -> extraction d'informations

- EN = Personnes + Organisations + Lieux

Sommaire

Présentation du projet

Définitions et outils à disposition

Les graphes d'Unitex

Balisage du corpus

Dictionnaires utilisés

Mon projet

Organisations des fichiers et lancement du programme

Visualisation et nouveau balisage

Extraction des EN

Erreurs et évaluation des résultats

Présentation du projet

- L'équipe d'Informatique Linguistique de l'IGM
- Projet Infom@gic -> laboratoire industriel de tests pour valider les meilleures technologies
- Création de graphes pour la reconnaissance et la classification des EN

Définitions et Outils à disposition

- Nom propre / EN
- Les dictionnaires de noms propres
- Variation des EN (graphique, syntaxique, lexicales ou ellipses) -> Normalisation
- Classes sémantiques :

ENAMEX :

- PERSON (*Le président de la Côte d'Ivoire, Laurent Gbagbo* ou *M. Gbagbo*)
- ORGANIZATION (*Centre national des arts et de la culture* ou *CNAC*)
- LOCATION (*Côte d'Ivoire*)
- NATIONALITY (*italien*)
- TITLE (*quotidien « Le Jour »* ou *Magazine « Abidjan Magazine »*)

NUMEX:

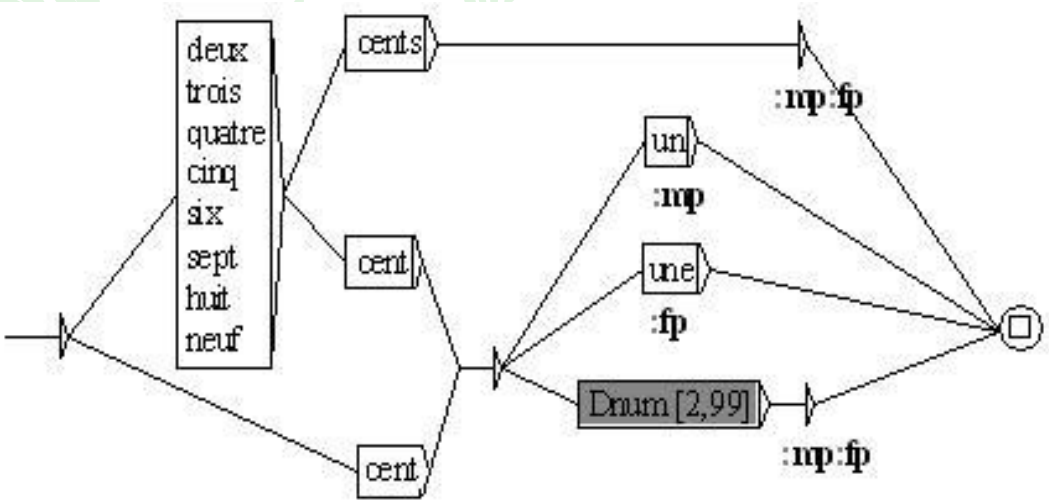
- NUMBER (*une dizaine* ou *54*)
- MESURE (*mille jeunes par an* ou *25.000 hectares* ou *1,2 millions de tonnes*)
- MONEY (*environ 727 millions de francs CFA* ou *400 millions de dollars*)
- PERCENT (*55%*)
- NUMERO (*###668* ou *page 5* ou *n°2*)

TIMEX:

- TIME (*soir* ou *à 18h* ou *17 heures 42*)
- DATE (*le 15 octobre 2001* ou *mardi 13 août dernier* ou *en 1957*)
- DUREE (*pendant près de trois heures* ou *Depuis au moins cinq mois*)
- FREQUENCE (*une fois* ou *chaque jour* ou *de temps à autre*)

Les graphes d'Unitex

- LADL (Laboratoire d'Automatique Documentaire et Linguistique)
- Environnement de développement qui permet de construire des grammaires et d'utiliser des dictionnaires



Balisage du corpus

<ENAMEX TYPE=PERSON>…</ENAMEX>

<NUMEX TYPE=NUMERO>…</NUMEX>

- Extrait du corpus interne de THALES
- Langue : français
- Taille de l'extrait : 10 965 Ko
- Sujet : Evénements politiques en Côte d'Ivoire et au Kossovo
- Caractéristiques : petits textes, dépêches d'agence, extraits de presse.

Utilisation des dictionnaires après étiquetage du texte

<avoir>, <N:ms>, <A+Toponyme>

- catégories syntaxiques : nom, verbe, pronom, etc.
- catégories sémantiques : humain, prénom, toponyme, etc.

Dictionnaires utilisés

- dictionnaire des noms de **professions**
- dictionnaires des **prénoms simples**
- dictionnaire des noms de **Papes**
- dictionnaires de **toponymes**
- dictionnaire de **sigles et d'abréviations**
- dictionnaire ***ORG*** crée à la main :

  Agence Nationale de l'Aviation Civile,.N+ORG:fs

  Agence nationale de l'aviation civile,Agence Nationale de l'Aviation Civile.N+ORG:fs

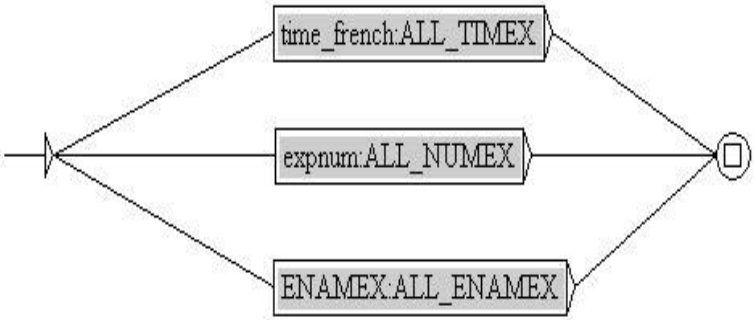
  ANAC,Agence Nationale de l'Aviation Civile.N+ORG+Sigle:fs

  A.N.A.C.,Agence Nationale de l'Aviation Civile.N+ORG+Sigle:fs
- Création du programme *sigles_fr* -> génère automatiquement les sigles d'une organisation

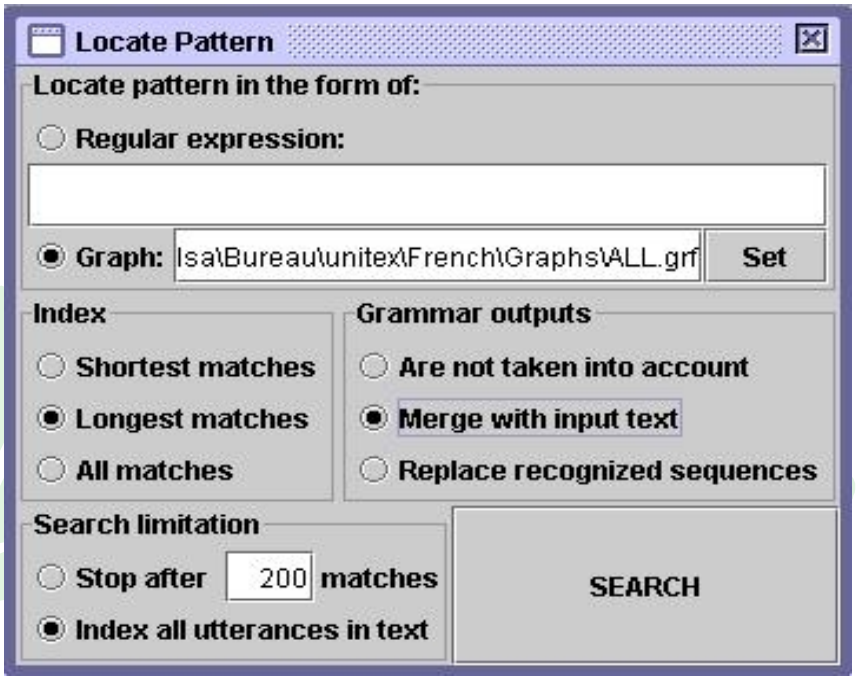
Mon projet

Organisation des fichiers

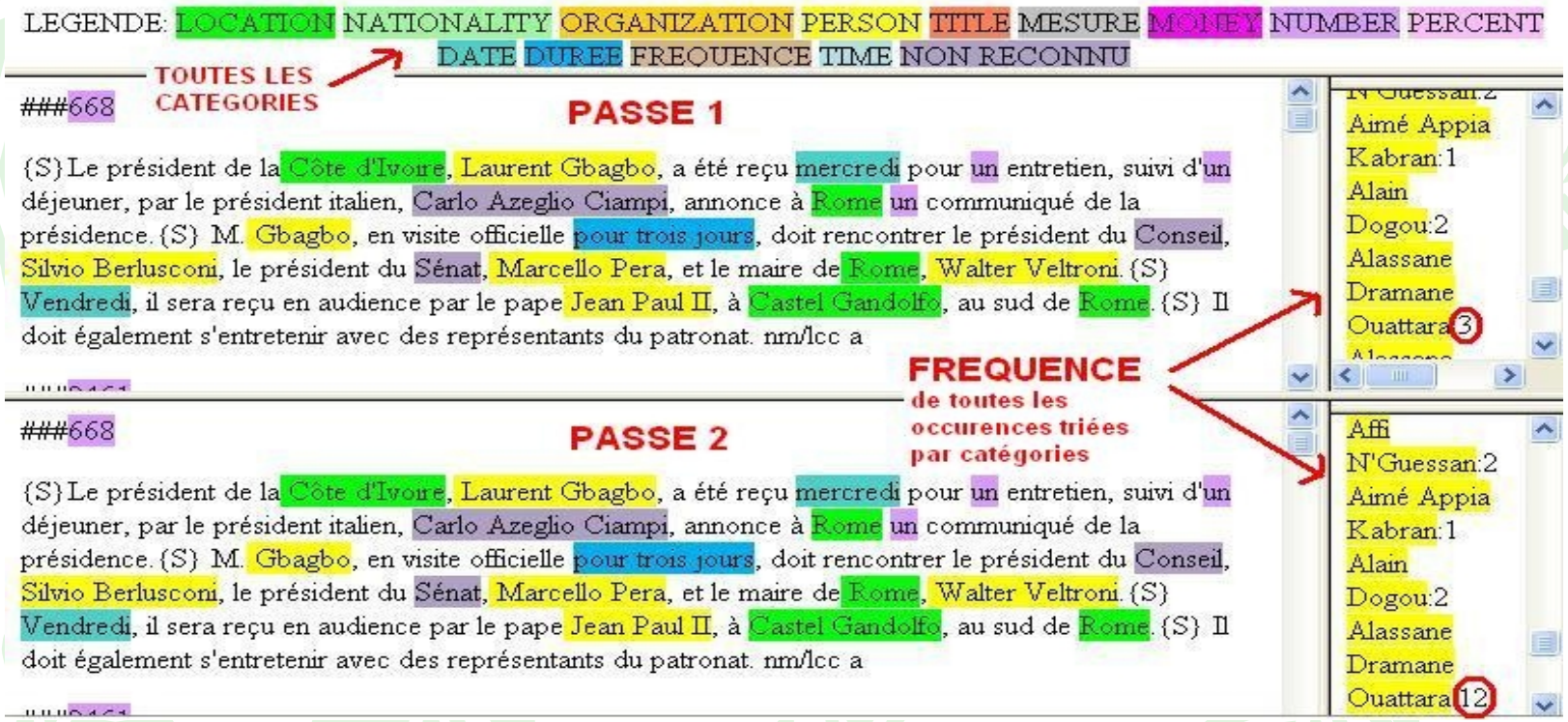
Graphe ALL.grf :



Démarche pour lancer les graphes



Visualisation et nouveau balisage



ATTRIBUTS pour différencier mots déclencheurs et EN

Titre=###668

Le Fonction=président de laNom= Côte d'Ivoire$_D$, Nom= Laurent Gbagbo, a été reçu mercredi pour un entretien, suivi d'un déjeuner, par le Fonction=président italien$_D$, Nom= Carlo Azeglio Ciampi, annonce à Nom=Rome$_D$ un communiqué de la présidence. Civilité=M. Nom= Gbagbo, en visite officielle pour trois jours, doit rencontrer le Fonction=président du Conseil, Nom= Silvio Berlusconi, le Fonction=président du Sénat, Nom= Marcello Pera, et le Fonction=maire deNom= Rome$_D$, Nom= Walter Veltroni. Vendredi, il sera reçu en audience par le pape Nom=Jean Paul II, à Nom=Castel Gandolfo$_D$, au sud de Nom=Rome$_D$. Il doit également s'entretenir avec des représentants du patronat. nm/lcc a

TROUVE GRACE A UN DICTIONNAIRE

GENRE TYPE et de la Fonction — **GENRE** de la Civilité

Titre=###668

Fonction=[ms+fs][profession]Le président de laNom= Côte d'Ivoire$_D$, Nom= Laurent Gbagbo, a été reçu mercredi pour un entretien, suivi d'un déjeuner, par Fonction=[ms+fs][profession]le président italien$_D$, Nom= Carlo Azeglio Ciampi, annonce à Nom=Rome$_D$ un communiqué de la présidence. Civilité=[ms]M. Nom= Gbagbo, en visite officielle pour trois jours, doit rencontrer Fonction=[ms+fs][profession]le président du Conseil, Nom= Silvio Berlusconi,Fonction=[ms+fs][profession] le président du Sénat, Marcello Pera, et Fonction=[ms+fs][profession]le maire de Rome, Nom= Walter Veltroni. Vendredi, il sera reçu en audience par le pape Nom=Jean Paul II, à Nom=Castel Gandolfo$_D$, au sud de Nom=Rome$_D$. Il doit également s'entretenir avec des représentants du patronat. nm/lcc a

ATTRIBUTS soulignés

###668

Le président de la$_D$(Côte d'Ivoire)$_D$,$_D$(Laurent)$_D$ Gbagbo*, a été reçu mercredi pour un entretien, suivi d'un déjeuner, par le président$_D$(italien)$_D$, $_D$(Carlo)$_D$ Azeg (Rome)$_D$ un communiqué de la présidence. M. Gbagbo*, en visite officielle pour trois jours, doit rencon eil,$_D$(Silvio)$_D$ Berlusconi*, le président du Sénat,$_D$(Marcello)$_D$ Pera*, et le maire de $_D$(Rome)$_D$, $_D$(W edi, il sera reçu en audience par le pape $_D$(Jean Paul)$_D$ II*, à $_D$(Castel Gandolfo)$_D$, au sud de $_D$(Ro entretenir avec des représentants du patronat. nm/lcc a

Laurent Gbagbo

civilité:
Prénom:Laurent
Nom:Gbagbo
Nationalité:
Fonction:président
Grade:
Pays:Côte d'Ivoire

MOTS DU DICTIONNAIRE mis entre D(...)D

Noms de personnes

Preuve interne et externe

- **Contexte gauche** (civilité, titre ou nom de profession) :

  M. *Gbagbo*

  le ministre de l'Intérieur *Emile Boga Doudou*

  par ***le président italien*** *Carlo Azeglio Ciampi*

- **Preuve interne (prénom du dictionnaire):**

  et ***Patrick*** *Achi*

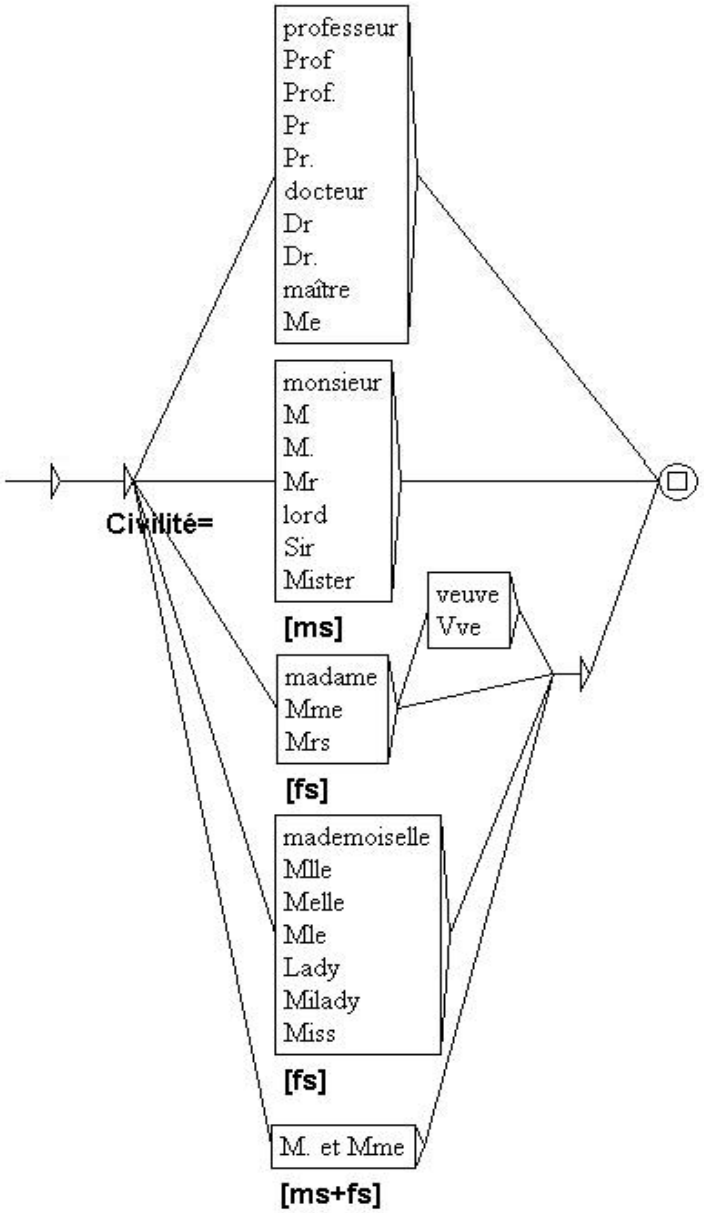
  Ahoua Stallone ***Julien***

- **Contexte droit** :

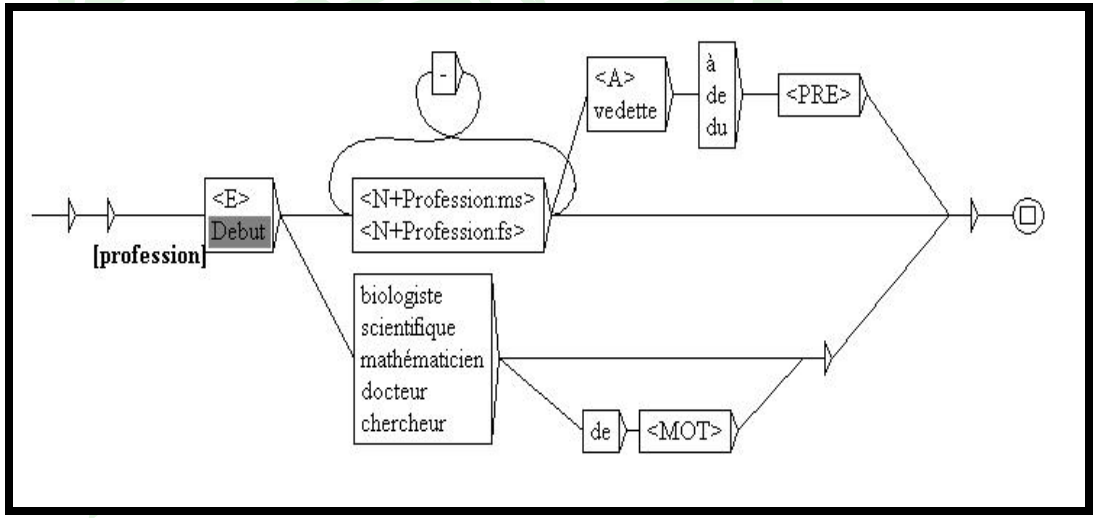
  *M. Ouattara****, président*** *du Rassemblement des républicains*

- **Noms de Papes grâce au dictionnaire** (*Jean-Paul II)*

Contextes gauche et droit

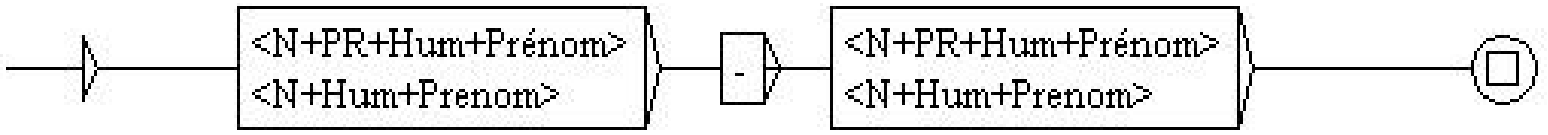


- **Civilités** (*Mme*, *Monsieur*)
- **Titres** : **politiques** (*président*, *ministre*, *député*), **militaires** (*général*, *lieutenant*), **religieux** (*cardinal*, *évêque*), **juridiques** …
- **Noms de professions** du dictionnaire (*le juge*, *l'architecte*) ou grâce au graphe *(artistre-peintre*, *caporal-chef*, *député-maire)*



Les formes de **prénoms**

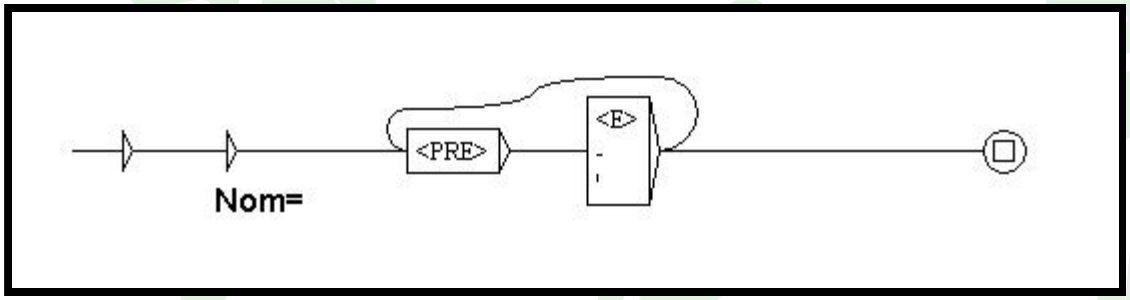
- **simples** (*Danièle*, *Louis*)
- **composés** (*Jean-Pierre*, *Charles Edouard*) :



- **composés en partie inconnus** (<PRE>)

Les formes de **patronymes**

- **simples** (*Dupont, Durand-Pérec*) :



- **composés de la particule « N' »** (*N'Guessan*, *N'dia Coffi*):

Reconnaissance des coordinations de noms de personnes

MM Bédié et Guéi

***MM. Marcello Pera**, président du Sénat, **Silvio Berlusconi**, Premier ministre italien **et Walter Veltroni**, maire de Rome.*

***MM. Ezalé** (SODEXAM) **et Abonouan** (ANAC)*

Ce qu'il reste à faire

- **Prénoms abrégés simples** (*E.* pour *Emmanuel*, *Th.* pour *Thierry*) et **prénoms composés abrégés** (*J.P.*, *J.-P.*, *J-P*, *J-P)*
- **Patronymes composés d'une particule excepté « N' » :** noms d'origine étrangère (*Mac Donnell-Douglas, O'Ryan, El Amra, Da Silva, Do Macedo)* et français *(Le Falch'un, Dupont de la Fontaine)*

Organisations

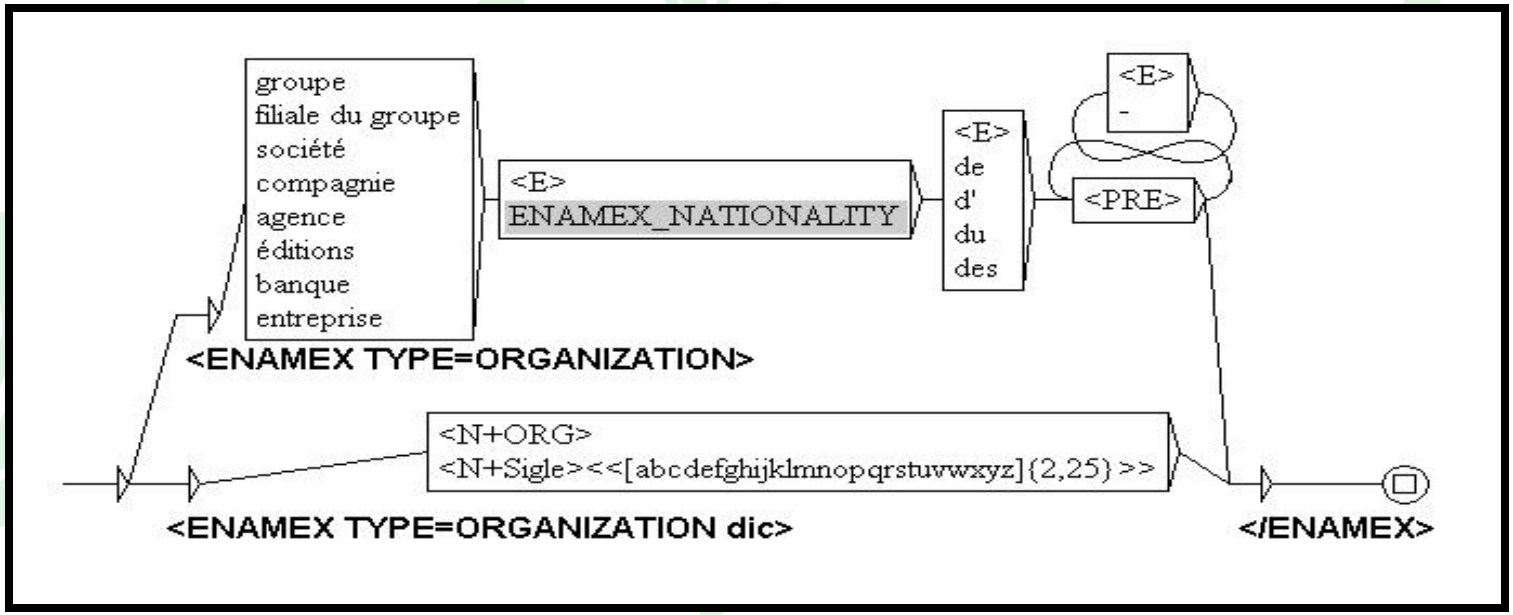
- **Preuve externe :**

Banque sud-africaine *APSA*
compagnie belge *SN Brussels Airlines*
société *Air Afrique*

- **Dictionnaires**

Ce qu'il reste à faire:

- **Preuve interne** (***Organisation*** *mondiale de la santé,* ***Banque*** *de France)*
- **Coordinations de noms**



Lieux

- **Preuve externe :**
  *dans le **département** de Bouna*
  *de la **Vallée** du Bandaman*
  *à l'**aéroport** d'Abidjan*
- **Dictionnaires :** <N+PR+Toponyme-Hum>

Ce qu'il reste à faire:

- **Preuve interne** (*Chaumont-**sur**-Loire, Main **Street**, Yosemite National **Park**)*
- **Dictionnaire des villes d'Afrique**

Les autres EN

- **Les nationalités**

  <A+Toponyme>+<N+PR+Toponyme+Hum>

- **Les titres**

  *la **chanson** « Hip hop »*

  *Un **film** « Prévention génocide »*

  *Le **journal** Fraternité Matin*

- **Les expressions numériques, de dates et de temps**

  longtemps avant que -> longtemps avant

  alors que -> non reconnu

  *très **tard** -> **très tard***

  ***pour la saison 2001**-2002 -> pour la saison 2001-2002*

  ***pour de mois** d'août -> pour de mois d'août*

  ***le 27 août** dernier -> le 27 août dernier*

Les erreurs

- Erreurs de catégorisation (mots ambigus) :
  France Télévision v.s. ***France Galle***
- Erreurs de sous-reconnaissance :
  Valéry Giscard *d'Estaing* v.s. ***André Wiltzer*** *d'Haironville*
- Titres des articles de journaux en majuscules
- Erreurs de frappe (***journal Fraternité*** *matin****, le 18 juillet 2O02)***

Evaluation = Rappel + Précision

Rappel = nombre de réponses pertinentes du système / nombre de réponses idéal

Précision = nombre de réponses pertinentes du système / nombre de réponses fournies par le système

CONCLUSION

- Grammaire des noms de personnes
- Dictionnaires car peu de preuves
- Limites des méthodes linguistiques :
  - Incomplétude des grammaires locales
  - Ambiguïtés
  - Absence de contextes
- Information importante sur le sens et le contenu des textes

Bibliographie

- *RECONNAISSANCE AUTOMATIQUE DES NOMS PROPRES* : *Application à la classification automatique de textes journalistiques*, thèse de Doctorat en Informatique de Nathalie Friburger (Tours, 2002)
- *CasSys : Manuel d'installation et d'utilisation avec Unitex* de Nathalie Friburger (Tours)
- *Unitex 1.2 :Manuel d'utilisation de Sébastien Paumier* (Université de Marne-la-Vallée, Décembre 2005).

- *Equipe d'Informatique Linguistique de l'IGM* : http://infolingu.univ-mlv.fr/
- *LADL* : http://ladl.univ-mlv.fr/French/
- Manuel d'installation d'Unitex :http://www-igm.univ-mlv.fr/~unitex
- *GRAALWEB* : http://igm.univmlv.fr/~mconstan/library/index_graalweb.html
- *TEI* : http://www.tei-c.org
- *MUC* : http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html
- *GLOSSANET* : http://glossa.fltr.ucl.ac.be