Les Entités Nommées : usage et degrés de précision et de désambiguïsation

Claude Martineau
Elsa Tolone
Stavroula Voyatzi
@univ-mlv.fr

26ème Colloque International sur le Lexique et la Grammaire
Bonifacio, 2-6 octobre 2007

Sommaire

- Extraction d'Information
- Extraction d'Entités Nommées
- Présentation du travail
  - Typologie
  - Méthodologie et outils
  - Ressources linguistiques
- Résultats
- Futur développement

Extraction d'Information (1/2)

- L'information d'aujourd'hui sur support informatique est :
  - massive
  - complexe et hétérogène
  - soumise à des contraintes de temps réel
- **Extraction d'Information (IE)** : conversion du texte en données structurées répondant à des questions factuelles

  ***QUI A FAIT QUOI A QUI QUAND OÙ COMMENT*...**
- **Applications** : recherche, indexation, aide à la décision, veille, question/réponse, construction de ressources …

Extraction d'Information (2/2)

Exemple

Un raid aérien a fait au moins 11 morts et 12 blessés sur le village de Menakro le mardi 12 février (Le Monde*, 2003*)

ENTITES	SEGMENTS	REPRESENTANTS
DATE	*le mardi 12 février*	12/02/03
LIEU	*sur le village de Menakro*	Menakro
FAIT	*Un raid aérien*	Attaque militaire
IMPACT	*au moins 11 morts*	Pertes humaines (Q<50)
	(au moins) 12 blessés	Dommages humains (Q<50)

Extraction d'Entités Nommées (1/3)

- La tâche d'Extraction d'Information a mis en évidence l'intérêt de reconnaître les Entités Nommées

- **Qu'est-ce que c'est une Entité Nommée (EN) ?**

  ...tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (ie. humain, économique, géographique, etc.)

  ...noms propres au sens classique, noms propres dans un sens élargi mais aussi expressions de temps et de quantité

  (MUC-7, Chinchor 1998)

Extraction d'Entités Nommées (2/3)

- **Objectif :** Repérer et catégoriser les EN dans un texte
- **Systèmes d'EEN :** plusieurs approches
  - **Symbolique**

    Règles construites à la main → Lisibilité, Évolutivité, Incomplétude
  - **Statistique**

    Connaissances acquises automatiquement par apprentissage à partir d'un corpus annoté à la main → Robustesse, Coût d'annotation
  - **Hybride**

    Symbolique + Statistique → Systèmes préférés

Extraction d'Entités Nommées (3/3)

Aujourd'hui, toutes les approches offrent des taux de reconnaissance (repérage & catégorisation élémentaire) au dessus de 90%. Cependant, l'attribution des catégories reste une tâche assez complexe.

❖ **Problèmes de catégorisation dus aux phénomènes linguistiques :**

- polysémie : ***Washington*** ***(Lieu*** *VS* ***Personne)***
- métonymie : *La* ***France*** *a signé le traité de Kyoto* ***(Lieu*** *VS* ***État)***
- référents multiples (« facettes sémantiques », selon Cruse 1986, 1995) :
Arnold Schwarzenegger *(acteur/ gouverneur de Californie/ bodybuilder)*

Désambiguïsation : tâche essentielle pour l'extraction d'EN

Présentation du travail (1/3)

- **Contexte :** projet R&D INFOM@GIC (2006-2008)
- **Objectif :** extraction et annotation fine des EN (types, sous-types, attributs)
- **Corpus brut**

  Genre : dépêches AFP et Reuters, extraits de presse

  Taille : 38 Mo, 3 173 071 occurrences de mots

  Format : TXT, disponible en HTML

  Langue : français

  Sujet : événements politiques en Côte d'Ivoire et au Kossovo (2000-2003)

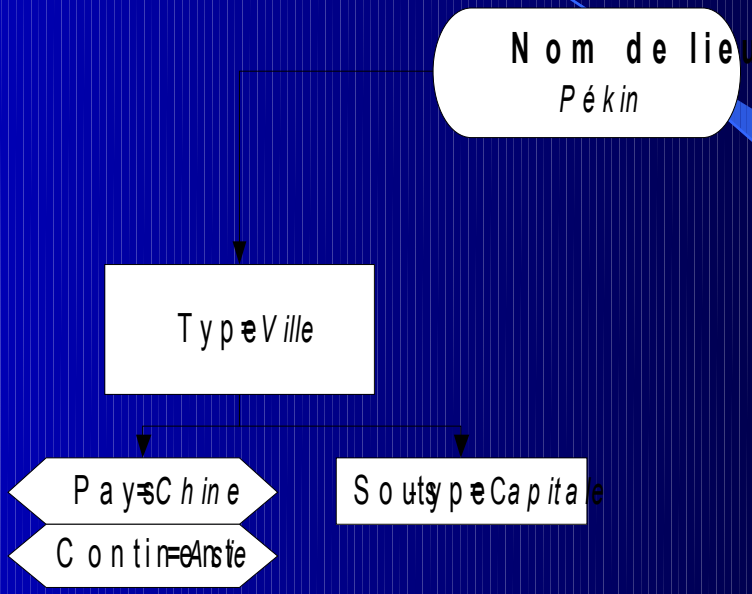
Présentation du travail (2/3)

Typologie : 9 catégories

Personnes	*Laurent Gbagbo, Pascal Affi N'Guessan*
Lieux (expressions spatiales)	*Guinée-Bissau* *axe Bouaké-Yamoussoukro*
Organisations	*Mouvement patriotique de Côte d'Ivoire, MPCI*
Faits	*25e sommet franco-britannique du Touquet*
Moyens	*Boeing 747-300*
Œuvres	*New Press, Le Nouveau Testament*
Dates & Heures (expressions temporelles)	*le 29 mars 2003, 10h00 GMT* *depuis jeudi matin*
Expressions numériques	*400 kilomètres, 50%*
Coordonnées	*01 56 40 13 72, appels.actu@rfi.fr*

Présentation du travail (3/3)

- **Vers une annotation fine des EN…**



Typologie

- **considérablement étendue par rapport à la typologie de base (MUC'98)**
- **ouverte et paramétrable pour chaque application**

Méthodologie et outils

- **Méthode symbolique**

  Grammaires locales écrites sous forme de RTN utilisant des informations morphosyntaxiques et sémantiques

- **Pour quelle raison ?**

  Réutilisation, validation et adaptation des ressources linguistiques en vue de l'extraction des EN

- **Outil utilisé**

  Unitex (Paumier 2003) : analyse des textes et traitement des ressources linguistiques

- **Étapes principales**
  - **Analyse lexicale** : consultation des dictionnaires
  - **Reconnaissance de séquences pertinentes** : application des grammaires locales
  - **Étiquetage des séquences isolées** : comparaison avec les séquences déjà reconnues

Ressources linguistiques (1/3)

Nous utilisons des dictionnaires
- spécialisés (43 921 entrées)
- généraux (1 256 951 entrées)
- construits pour le corpus (1000 entrées)

Type de dictionnaire	Auteur	Exemple	Effectifs
Prénoms	Maurel *et al.* 1996	*Caroline,.N+PR+Hum+Prénom:fs*	24 291
Toponymes	Maurel & Piton 1999	*Seine,.N+PR+Hydronyme:fs*	6 107
Pays, Capitales et Gentilés	Maurel & Piton 1999	*France,.N+PR+Toponyme+Pays+IsoFR:fs*	3 093
Adjectifs toponymiques	Maurel & Piton 1999	*parisiens,parisien.A+Toponyme+Ville:mp*	3 407
Noms de profession	Fairon 2004	*banquiers,banquier.N+Profession:mp*	4 185
Sigles et Abréviations	Maurel *et al.* 1996	*Solensi,Solidarité Enfants Sida.N+Sigle:fs*	2 838
Toponymes Africains	Trouvés sur le Web	*Assinie,.N+PR+Toponyme+Ville:fs*	400
Organisations et Abréviations	Elsa Tolone	*FMI,Fonds Monétaire International.N+Sigle+Org:ms*	500
Mots simples (DELAF)	LADL / IGM	*praesidia,praesidium.N+HumColl:mp*	984 723
Mots composés (DELACF)	LADL / IGM	*week-ends,week-end.N+Tps+weekend:mp*	**272 228**

Ressources linguistiques (2/3)

- Les entrées de catégorie *Nom* et *Adj* contenues dans les dictionnaires peuvent jouer plusieurs rôles dans la tâche d'Extraction d'EN :
  - entité nommée en tant que telle *(Napoléon, Côte d'ivoire, SNCF)*
  - mots déclencheurs présents dans le contexte immédiat d'une EN

  Exemple: *Le groupe **Vivendi** prend le contrôle de **Neuf Télécom***

  preuve externe preuve interne

- Les traits sémantiques contenus dans les dictionnaires peuvent dans certains cas servir à attribuer une catégorie à l'EN repérée :

  Exemple: *L'**UNESCO** a fait de l'**Afrique** une priorité*

UNESCO,Organisation des Nations Unies pour l'éducation, la science et la culture.N+Sigle+Org:fs
Afrique,Afrique.N+PR+Toponyme+Supranational:fs

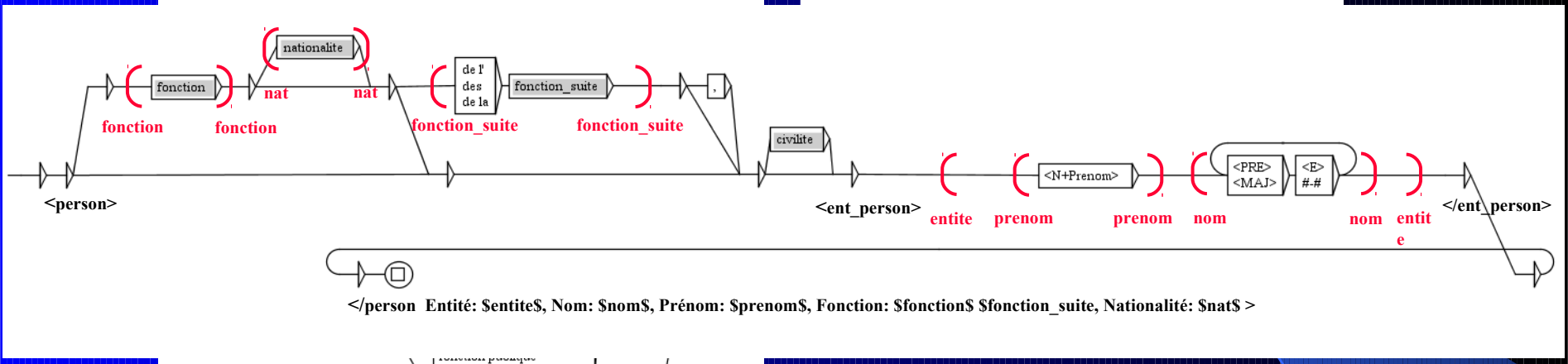
Ressources linguistiques (3/3)

- *Nous avons utilisé des grammaires locales représentées par environ 600 graphes qui font référence aux informations contenues dans les dictionnaires*
- *Ces grammaires ont été élaborées au sein de l'IGM et sont rassemblées et accessibles grâce au système Graalweb (Constant 2004)*
- *Initialement prévues pour effectuer la reconnaissance de patterns morphosyntaxiques, les grammaires ont dû être adaptées pour l'extraction d'EN*

De la reconnaissance à l'extraction

Le ministre ivoirien des Affaires Étrangères, Abou Drahamane Sangaré
a convoqué l'ambassadeur du Burkina à Abidjan, Émile Ilboudo

Attribut Attribut EN



Concordance

Le ministre ivoirien des Affaires Étrangères, Abou Drahamane Sangaré a convoqué

Le <person> _mini Drahamane Sangaré
</ent_person></perso bou, Fonction:ministre
des Affaires Étrangère n

- **Délimiter l'EN: Identifier ses bornes**
- **Attribuer une catégorie sémantique non ambiguë**
- **Associer éventuellement les attributs présents**

Nom: Drahamane Sangaré
Prénom: Abou
Fonction: ministre des Affaires Étrangères
Nationalité: ivoirien

Extension du contexte d'analyse

Le Quai d'Orsay se trouve dans l'impossibilité d'affirmer que…

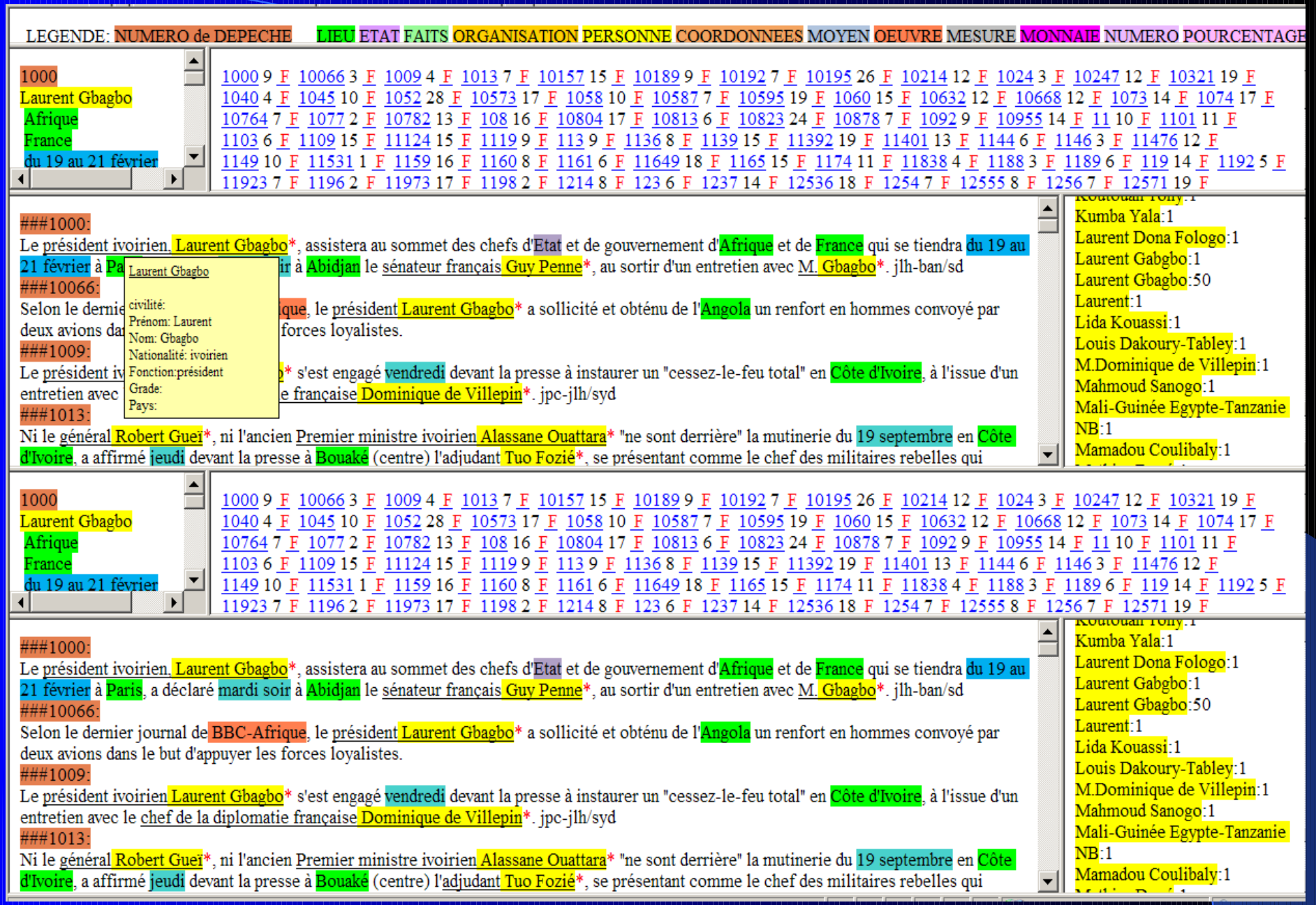
[Dictionnaire]	*Orsay,.N+PR+Toponyme+Ville:fs*
[Preuve interne]	*Quai d'Orsay :lieu_micro-toponyme*
[Preuve interne/externe]	*Quai d'Orsay_se trouve :lieu_micro-toponyme*
[Contexte éloigné]	*Quai d'Orsay_se trouve dans l'impossibilité:organisation*

- *Améliorer la catégorisation d'EN en tenant compte de contextes plus longs :*
  - *autres catégories grammaticales, notamment les verbes*
  - *unités complexes (expressions figées, constructions à Vsup)*

Résultats

- Une première évaluation sur la totalité du corpus est prévue fin 2007
- Visualisation des résultats

DEMO


LEGENDE: NUMERO de DEPECHE LIEU ETAT FAITS ORGANISATION PERSONNE COORDONNEES MOYEN OEUVRE MESURE MONNAIE NUMERO POURCENTAGE
1000
Laurent Gbagbo
Afrique
France
du 19 au 21 février
1000 9 F 10066 3 F 1009 4 F 1013 7 F 10157 15 F 10189 9 F 10192 7 F 10195 26 F 10214 12 F 1024 3 F 10247 12 F 10321 19 F
1040 4 F 1045 10 F 1052 28 F 10573 17 F 1058 10 F 10587 7 F 10595 19 F 1060 15 F 10632 12 F 10668 12 F 1073 14 F 1074 17 F
10764 7 F 1077 2 F 10782 13 F 108 16 F 10804 17 F 10813 6 F 10823 24 F 10878 7 F 1092 9 F 10955 14 F 11 10 F 1101 11 F
1103 6 F 1109 15 F 11124 15 F 1119 9 F 113 9 F 1136 8 F 1139 15 F 11392 19 F 11401 13 F 1144 6 F 1146 3 F 11476 12 F
1149 10 F 11531 1 F 1159 16 F 1160 8 F 1161 6 F 11649 18 F 1165 15 F 1174 11 F 11838 4 F 1188 3 F 1189 6 F 119 14 F 1192 5 F
11923 7 F 1196 2 F 11973 17 F 1198 2 F 1214 8 F 123 6 F 1237 14 F 12536 18 F 1254 7 F 12555 8 F 1256 7 F 12571 19 F
###1000:
Le président ivoirien Laurent Gbagbo*, assistera au sommet des chefs d'Etat et de gouvernement d'Afrique et de France qui se tiendra du 19 au 21 février à Pa ir à Abidjan le sénateur français Guy Penne*, au sortir d'un entretien avec M. Gbagbo*. jlh-ban/sd
###10066:
Selon le dernie que, le président Laurent Gbagbo* a sollicité et obtenu de l'Angola un renfort en hommes convoyé par deux avions da forces loyalistes.
###1009:
Le président iv * s'est engagé vendredi devant la presse à instaurer un "cessez-le-feu total" en Côte d'Ivoire, à l'issue d'un entretien avec e française Dominique de Villepin*. jpc-jlh/syd
###1013:
Ni le général Robert Gueï*, ni l'ancien Premier ministre ivoirien Alassane Ouattara* "ne sont derrière" la mutinerie du 19 septembre en Côte d'Ivoire, a affirmé jeudi devant la presse à Bouaké (centre) l'adjudant Tuo Fozié*, se présentant comme le chef des militaires rebelles qui
Laurent Gbagbo
civilité:
Prénom: Laurent
Nom: Gbagbo
Nationalité: ivoirien
Fonction:président
Grade:
Pays:
Kumba Yala:1
Laurent Dona Fologo:1
Laurent Gabgbo:1
Laurent Gbagbo:50
Laurent:1
Lida Kouassi:1
Louis Dakoury-Tabley:1
M.Dominique de Villepin:1
Mahmoud Sanogo:1
Mali-Guinée Egypte-Tanzanie
NB:1
Mamadou Coulibaly:1
1000
Laurent Gbagbo
Afrique
France
du 19 au 21 février
1000 9 F 10066 3 F 1009 4 F 1013 7 F 10157 15 F 10189 9 F 10192 7 F 10195 26 F 10214 12 F 1024 3 F 10247 12 F 10321 19 F
1040 4 F 1045 10 F 1052 28 F 10573 17 F 1058 10 F 10587 7 F 10595 19 F 1060 15 F 10632 12 F 10668 12 F 1073 14 F 1074 17 F
10764 7 F 1077 2 F 10782 13 F 108 16 F 10804 17 F 10813 6 F 10823 24 F 10878 7 F 1092 9 F 10955 14 F 11 10 F 1101 11 F
1103 6 F 1109 15 F 11124 15 F 1119 9 F 113 9 F 1136 8 F 1139 15 F 11392 19 F 11401 13 F 1144 6 F 1146 3 F 11476 12 F
1149 10 F 11531 1 F 1159 16 F 1160 8 F 1161 6 F 11649 18 F 1165 15 F 1174 11 F 11838 4 F 1188 3 F 1189 6 F 119 14 F 1192 5 F
11923 7 F 1196 2 F 11973 17 F 1198 2 F 1214 8 F 123 6 F 1237 14 F 12536 18 F 1254 7 F 12555 8 F 1256 7 F 12571 19 F
###1000:
Le président ivoirien Laurent Gbagbo*, assistera au sommet des chefs d'Etat et de gouvernement d'Afrique et de France qui se tiendra du 19 au 21 février à Paris, a déclaré mardi soir à Abidjan le sénateur français Guy Penne*, au sortir d'un entretien avec M. Gbagbo*. jlh-ban/sd
###10066:
Selon le dernier journal de BBC-Afrique, le président Laurent Gbagbo* a sollicité et obtenu de l'Angola un renfort en hommes convoyé par deux avions dans le but d'appuyer les forces loyalistes.
###1009:
Le président ivoirien Laurent Gbagbo* s'est engagé vendredi devant la presse à instaurer un "cessez-le-feu total" en Côte d'Ivoire, à l'issue d'un entretien avec le chef de la diplomatie française Dominique de Villepin*. jpc-jlh/syd
###1013:
Ni le général Robert Gueï*, ni l'ancien Premier ministre ivoirien Alassane Ouattara* "ne sont derrière" la mutinerie du 19 septembre en Côte d'Ivoire, a affirmé jeudi devant la presse à Bouaké (centre) l'adjudant Tuo Fozié*, se présentant comme le chef des militaires rebelles qui
Kumba Yala:1
Laurent Dona Fologo:1
Laurent Gabgbo:1
Laurent Gbagbo:50
Laurent:1
Lida Kouassi:1
Louis Dakoury-Tabley:1
M.Dominique de Villepin:1
Mahmoud Sanogo:1
Mali-Guinée Egypte-Tanzanie
NB:1
Mamadou Coulibaly:1

En guise de conclusion

- ***Validation et enrichissement des ressources existantes à l'IGM et adaptation en vue de l'Extraction d'Entités Nommées***

  Nouvelle version des grammaires locales sur Graalweb

- ***Le degré de précision et de désambiguïsation dépend des besoins applicatifs***

Futur développement

- ***Utilisation des données syntaxiques issues des tables de verbes du lexique-grammaire***

MERCI !

Bibliographie (1/3)

- **Allerton D.** (1987), «The linguistic and sociolinguistic status of proper names », in *Journal of Pragmatics,* vol. 11 : 61-92.
- **Bikel D. M., Miller S. Schwartz R. et Weischedel R.** (1997), «Nymble: a high-performance learning name-finder », in *Proceedings of the 5th Conference on Applied Natural language processing*, *31/03-03/04 1997*, Morgan Kaufman Publishers Inc., Washington, DC, pp. 194-201.
- **Chinchor N.** (1998), « MUC-7 Named Entity Task Definition (version 3.5) », in *Proceedings of the 7th Message Understanding Conference (MUC-7)*, *19 April-1 May* 1998, Fairfax, VA.
- **Constant M.** (2004), « GRAAL, une bibliothèque de graphes : mode d'emploi », in Muller C., Royeauté J. et Silberztein M. (éds), *Cahiers de la MSH Ledoux 1, INTEX pour la linguistique et le traitement automatique des langues*, Presse Universitaire de Franche-Comté, Besançon : 321-330.
- **Courtois B.** (1990), « Un système de dictionnaires électroniques pour les mots simples du français », in Courtois B. et Silberztein M. (éds), *Dictionnaires électroniques du français*, *Langue Française,* n° 87, Larousse, Paris : 11-22.
- **Daille B. et Morin E.** (2000), « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », in Maurel D. et Gueunthner F. (éds), *Traitement Automatique des Langues*, vol. 41/3 : 601-621.
- **Dister A et Fairon C.** (2004), « Extension des ressources lexicales grâce à un corpus dynamique », in *Lexicometrica*, Paris, version électronique : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema7/Texte-Dister.pdf.
- **Fourour N.** (2002), « Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français », in *Actes de la 9ème Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, Nancy, vol. 1 : 265-274.

Bibliographie (2/3)

- **Friburger N.** (2002), *Reconnaissance automatique des noms propres : Application à la classification automatique des textes journalistiques*, Thèse de doctorat, Université de Tours, Paris.
- **Grass T.** (2000), « Typologie et traductibilité des noms propres de l'allemand vers le français à partir d'un corpus journalistique », in Maurel D. et Gueunthner F. (éds), *Traitement Automatique des Langues*, vol. 41/3 : 643-669.
- **Gross M.** (1981),« Les bases empiriques de la notion de prédicat sémantique », in *Langages*, n° 63, Larousse, Paris : 7-52.
- **Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M. et Tyson M.** (1996), « FASTUS : a cascaded finite-state transducer for extracting information from natural-language text», in Roche E. et Schabes Y. (éds), *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, USA : 383-406.
- **Jacquemin C. et Bush C.** (2000), « Fouille du Web pour la collecte d'entités nommées », in *Actes de la 8ème Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN 2000)*, Lausanne : 187-196.
- **Kleiber G.** (1999), *Problèmes de Sémantique, la polysémie en questions*, Presses Universitaires du Septentrion, Lille (Sens et structures), 223 p.
- **LE MEUR C., GALLIANO S. et GEOFFROIS E.** (2004), « Conventions d'annotations en Entités Nommées », *ESTER*, http://www.afcp-parole.org/ester/publis.html, pp. 6-10.
- **Li H., SRIHARI R, Niu C et Li W.** (2002), « Location normalization for information extraction », in *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, Association for Computational Linguistics, Taipei, Taiwan : 1-7.
- **Maurel D. et Piton O.** (1999), « Un dictionnaire de noms propres pour Intex : Les noms propres géographiques », in *Linguisticae Investigationes*, vol. 22 : 277-287.

Bibliographie (3/3)

- **Maurel D., Belleil C., Eggert E. et Piton O.** (1996), « Le projet PROLEX : réalisation d'un dictionnaire électronique relationnel des noms propres du français », in *Proceedings of GDR-PRC Communication Homme-Machine Séminaire Lexique*, Grenoble : 164-175.
- **McDonald D.** (1996), « Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names », in Boguraev B. et Pustejovsky J. (éds), *Corpus processing for lexical acquisition* (Language, Speech and Communication), MIT Press, Cambridge, London : 21-37.
- **Paumier S.** (2003), *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.
- **Poibeau T.** (2005), « Le statut référentiel des entités nommées » in *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France.
- **Roche E. et Schabes Y.** (1997), *Finite-State Language Processing*, Roche E. et Schabes Y. (éds), MIT Press, Cambridge, Mass./London (Language, Speech and Communication), 464 p.
- **Sekine S. et Nobata C.** (1998), « An Information Extraction System and a Customization Tool », in *Proceedings of the New Challenges in Natural Language Processing and its Application, 25-26 May 1998*, Tokyo, Japan.
- **Sekine S., Sudo K. et Nobata C.** (2002), « Extended Named Entity Hierarchy», in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain : 1818-1824.
- **Tolone E.** (2006), *Rapport technique de stage en Master I d'Informatique*, Université de Marne-la-Vallée, Paris, 39 p.
- **Watrin P.** (2006), *Une approche hybride de l'extraction d'information : sous-langages et lexique-grammaire*, Thèse de doctorat, Cental, Université de Louvain-La-Neuve, Belgique.

Ressources Linguistiques (1/2)

Dictionnaire morphologique du français (Système DELA)

- **mots simples (DELAF): 984 723 entrées**

  f_fléchie,f_canonique.cat_gram+infos sém+variante:infos morph

  praesidium,praesidium.N+HumColl:ms
  praesidia,praesidium.N+HumColl:mp
  présidium,présidium.N+HumColl+praesidium:ms
  présidiums,présidium.N+HumColl+praesidium:mp

- **mots composés (DELACF): 272 228 entrées**

  f_fléchie,f_canonique.cat_gram+infos sém+variante:infos morph

  week-end,week-end.N+Tps+weekend:ms
  week-ends,week-end.N+Tps+weekend:mp

Ressources Linguistiques(2/2)

Dictionnaires spécialisés (projet PROLEX)

- **Prénoms : 24 291 entrées**
  Laurent,.*N*+*PR*+*Hum*+*Prénom*:*ms*
- **Toponymes : 6 107 entrées**
  Seine,.*N*+*PR*+*Toponyme*+*Hydronyme*:*fs*
- **Pays, Capitales et Gentilés : 3 093 entrées**
  France,.*N*+*PR*+*Toponyme*+*Pays*+*IsoFR*:*fs*
  Paris,.*N*+*PR*+*Toponyme*+*Ville*+*Cap*+*IsoFR*:*ms*:*fs*
  Français,.*N*+*PR*+*Hum*+*Toponyme*+*Pays*:*ms*:*mp*
- **Abréviations et Sigles : 2 838 entrées**
  Solensi,*Solidarité Enfants Sida*.*N*+*Sigle*:*fs*
- **Professions : 4 185 entrées**
  avocat d'affaires,.*N*+*Profession*:*ms*
  avocate d'affaires,*avocat d'affaires*.*N*+*Profession*:*fs*