Le corpus EvalRefGen

Amalia Todirascu
LILPA, Université de Strasbourg
todiras@unistra.fr

Objectif

Développement d'un outil d'identification automatique de chaînes de références RefGen (Longo et Todirascu, 2010), disponible pour le français

Identification des chaînes de référence (Corblin, 1995 ; Cornish, 1995 ; Schnedecker, 1997)

Prise en compte des paramètres du genre textuel (Biber, et Conrad, 2009)

Système à base de connaissances (règles heuristiques qui définissent des contraintes linguistiques complexes entre les maillons d'une chaîne)

Integration de *RefGen* dans **un système de détection automatique des thèmes** (thèse de L.Longo)

Chaîne de référence (CR): définition

- au moins 3 expressions référentielles référant la même entité du discours (Schnedecker, 1997)
- rupture de chaine si redénomination du Np (Schnedecker, 1997)
- marqueur de continuité ou de rupture thématique dans le discours (Cornish, 1995), (Vonk *et al.*,1992), (Goutsos, 1997)

« **En entrant dans la cour de l'hôtel, Buckingham sauta à bas de son cheval, et, sans s'inquiéter de ce qu'il deviendrait, il lui jeta la bride sur le cou et s'élança vers le perron.**» (Dumas, « Les trois mousquetaires »)

Chaînes de référence et genre textuel

Le type de CR est une caractéristique du genre textuel étude sur les portraits journalistiques (Schnedecker, 2005) .

Etudes de CR dans un corpus multi-genre (Longo et Todirascu, 2009)

- longueur
- distance entre maillons
- type d'expression
- position thématique

Positionnement sur le calcul de la référence

Méthode par apprentissage automatique (Ng et Cardie, 2002 ; Hoste, 2005 ; Denis, 2007)

Pas de corpus de grande taille annoté en chaînes de référence pour le français (dernière campagne SemEval 2010); Annodis disponible depuis peu de temps

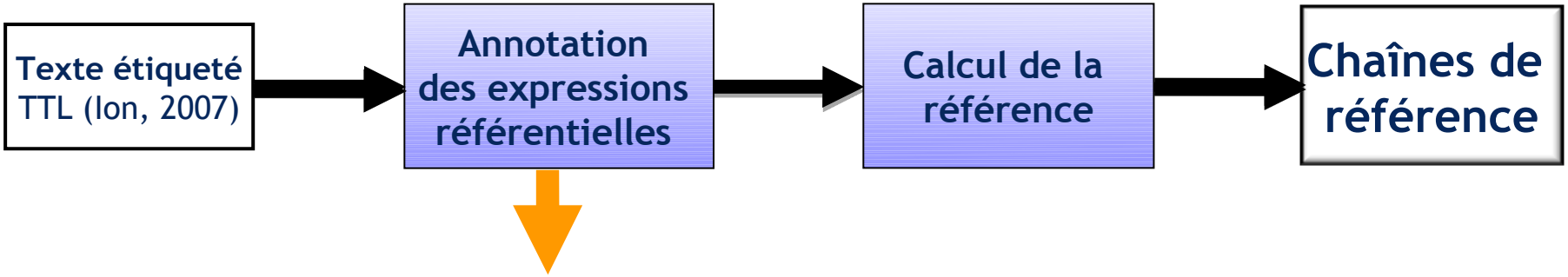
Méthode à base de connaissances (Mitkov, 2002 ; Bontcheva et al, 2002)

Résolution des anaphores pronominales

Notre méthode : RefGen (Longo et Todirascu, 2010a)

- Théorie d'optimalité(Beaver, 2004)
- Théorie de l'accessibilité (Ariel, 2001)
- Paramètres spécifiques au genre textuel (Longo et Todirascu, 2010c)

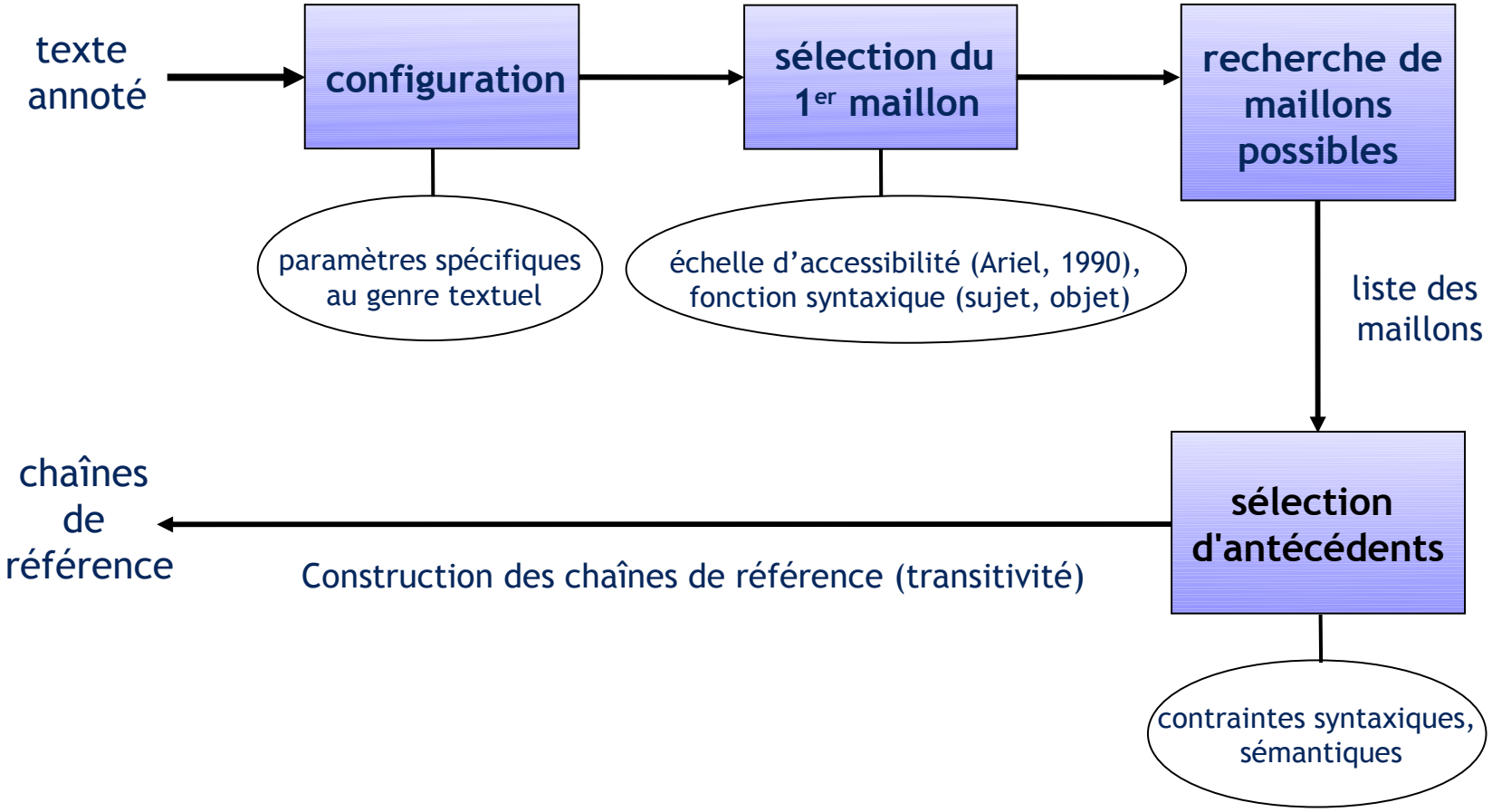
Architecture de *RefGen* (Longo et Todirascu, 2010b)




```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma="il" ana="Pp3ms" feat="imp">il</w>
<w lemma="y" ana="Pp3">y</w>
<w lemma="avoir" ana="Vaip3s">a</w>
<w lemma="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" pana="Af-mp">législatifs</w>
<w lemma="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Calcul de la référence
(Longo et Todirascu, 2010c)

EvalRefGen
Amalia Todirascu



Le corpus d'évaluation

Corpus d'évaluation :

textes libres de droits (15192 *tokens*), plusieurs genres :

- journalistique (fait divers -L'Est Républicain
- littéraire (roman libre de droits – Dumas « Les trois mousquetaires)
- juridique (Acquis Communautaire)
- rapport public (Eurosfaire):
- Résumés de films

annotations manuelles avec *Glozz* (Wildlöcher et Mathet, 2009) (mémoire de M.E.Vallette d'Osia)

1061 expressions référentielles annotées

267 chaînes de co-référence

Schéma d'annotation

Les types de CR

- Chaînes de référence :
  - au moins 3 expressions référentielles;
  - en cas de référent humain, la redénomination implique ouverture d'une chaîne (Schnedecker, 1997)
  - *Ex1 : [Barack Obama…il…il...] [Barack Obama ... le président américain...]*
  - les référents non-humains sont annotés

  Ex2 : Les changements climatiques…ces changements…ils

Méthodologie

- lecture du texte et identification des expressions référentielles et annotation
  - type d'expressions
  - Pas d'annotation morpho-syntaxique ou syntaxique
- établir les liens de coréférence
  - par paires
- relier les paires appartenant à la même chaîne

Expressions référentielles annotées

- noms propres : nom de personne (*Lucien Bertrand*), d'organisation (*La Commission*), nom propre modifié (*L'avocat général Pierre Denier*)
- groupes nominaux indéfini *: une concurrence effective, un Etat membre*
- descriptions définies : *L'article 12, cette décision*
- groupes nominaux complexes *: la notion de concentration*
- pronoms personnels
- pronoms réfléchis
- déterminants possessifs (*son choix*)

Expressions référentielles annotées (II)

Choix de ne pas annoter:

- sujets zéro des verbes infinitifs ou participes passés
- anaphores plurielles

  Paul et Marie…le couple…ils

- anaphores associatives (relation méronymique)

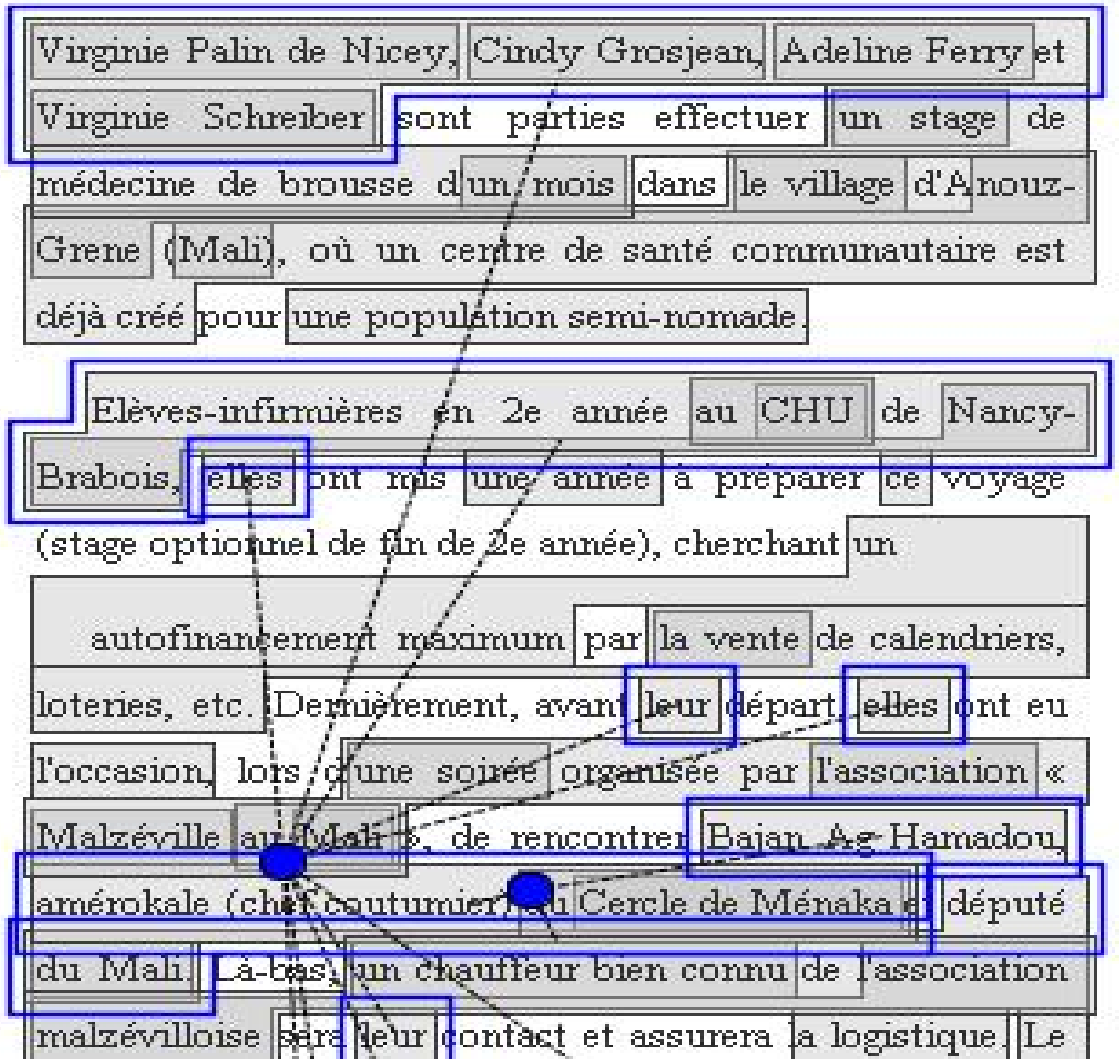
Comparaison d'outils

- Critères de comparaison:
  - objets annotés (simples ou imbriqués)
  - niveau intra-phrastique ou interphrastique;
  - définition d'un modèle d'annotation adaptée à ses objectifs;
  - format;
  - ergonomie;
  - facilité de prise en main;
  - fonction de visualisation

Comparaison d'outils (II)

- Plusieurs outils :
  - MMAX (Muller et Strube, 2006);
  - PALinkA (Orasan, 2003);
  - Annie (Gate) (Cunningham et al, 2000);
  - **Glozz (Wildocher et Mathet, 2009);**

Annotation avec *Glozz*



Glozz (II)

Plusieurs étapes

- Identification des unités et de leur propriétés
- Identification de schémas
- Identification de relations

Glozz (III)

Unités

- nom propre
- noms de fonction (grade, titre, métier)
- groupe nominal indéfini
- groupe nominal défini
- Pronoms
- Démonstratifs
- possessifs
- une propriété « type » pour la catégorie

Glozz (IV)

Schémas

- Plusieurs catégories
  - Nom propre (nom propre seul; rédenomination)
  - Chaînes anaphoriques
  - Chaînes de référence (si au moins trois unités sont reliées)

Glozz (V)

Relations

- Coréférence
  - entre anaphore et son antécédent
  - Entre unités
- entre schémas
  - Entre plusieurs chaînes de référence qui commencent par le même nom propre
  - Entre noms propres seuls

Glozz (VI)

Avantages

- Possibilité d'annoter des expressions imbriquées
  - [la notion de [concentration]]
- Facilité pour délimiter les unités et les relations
- Adapté pour les chaînes : les schémas
- Annotations exportées (séparées des données)
- Visualisation des chaînes en couleur

Glozz (VII)

Difficultés

- Outil en cours de développement
  - Fonctions ajoutées au fur et à mesure
  - Perte de données
- Contraintes : créer un schéma à partir d'un exemple (avant la version 1.0)
- Limite en taille
- Les annotations multiples sont peu lisibles

Retour d'expérience

- corpus multi-genre à construire
- pertinence de l'évaluation automatique
- difficulté de la tâche :
  - les catégories d'expressions (groupes nominaux complexes, structures imbriquées, groupes indéfinis ayant un emploi non-référentiel)
  - annoter même les phénomènes qui ne sont pas actuellement traités par l'outil
  - les ruptures thématiques
    - répétitions

Perspectives

- Le corpus est complété avec d'autres textes (faits divers, articles de loi)
  - en cours
- Evaluation de l'outil sur d'autres corpus
  - ANNODIS
  - mais ... choix méthodologiques différents

Références

- Corblin, F. (1995) Les formes de reprise dans le discours. Anaphores et chaînes de référence, Presses Universitaires de Rennes
- Cornish F. (1998) Les “chaînes topicales” : leur rôle dans la gestion et la structuration du discours, *Cahiers de Grammaire* 23, décembre 1998 : 19-40.
- Goutsos D. (1997). “Modeling Discourse Topic : sequential relations and strategies in expository text,” Advances in Discourse Processes, vol. LIX, Norwood : Ablex Publishing Corporation.
- Longo, Laurence, Todirascu, Amalia (2010), RefGen: Identifying Reference Chains to Detect Topics, Series Ed.: Kacprzyk, Janusz “Advances in Intelligent and Soft Computing”, Springer Verlag, ISSN: 1860-949X, ISSN: 1860-949X, volume 361, chapitre 3, pp. 27-40.
- Longo, Laurence, Todirascu, Amalia (2010): RefGen: a Tool for Reference Chains Identification. IMCSIT 2010: pp. 447-454 (indexé par DBLP-Trier)
- LONGO L, TODIRASCU, A. (2010). RefGen : un module d’identification des chaînes de référence dépendant du genre textuel, Actes du congrès TALN 2010 , Montreal 19-22 juillet,
- Schnedecker C. (1997). Nom propre et chaînes de référence. Recherches Linguistiques 21, Paris : Klincksieck.
- Schnedecker C. (2005). « Les chaînes de référence dans les portraits journalistiques : éléments de description », Travaux de linguistique 2/2005 (no 51), Duculot, 85-133.
- Vonk W., Hustinx L.G., Simons W.H. (1992). “The use of referential expressions in structuring discourse,” Language and Cognitive Processes, 7, 301-333.