10èmes Journées Internationales de la Linguistique de Corpus
Grenoble - 27 novembre 2019

De la coréférence exacte à la coréférence complexe :

problèmes typologiques et complexité de mise en œuvre en corpus.

Marine DELABORDE & Frédéric LANDRAGIN

Lattice

cnrs
dépasser les frontières

ENS
ÉCOLE NORMALE SUPÉRIEURE

UNIVERSITÉ SORBONNE NOUVELLE
PARIS 3

PSL
RESEARCH UNIVERSITY PARIS

USPC
Université Sorbonne Paris Cité

Contenu

1. Projet et corpus

- **Projet ANR DEMOCRAT** 2016 - 2020 (Landragin, 2016) : http://www.lattice.cnrs.fr/democrat/
  - **DE**scription et **MO**délisation des **C**haînes de **R**éférence : outils pour l'**A**nnotation de corpus (en diachronie et en langues comparées) et le **T**raitement automatique.
  - **Lattice** (Paris), **LiLPa** (Strasbourg), **ICAR** et **IHRIM** (Lyon).

⇒ Objectif : Analyse détaillée et contrastive des chaînes de coréférence dans un corpus diachronique de textes écrits.

- **Corpus** annoté en expressions référentielles et en chaînes de coréférence.
  - **58 blocs** de **10 000 mots** soit **200 000 expressions référentielles** annotées.
  - https://www.ortolang.fr/market/corpora/democrat/3

2. Méthodologie

- **Exemples**
  1. Corpus **Democrat** : au moment de l'annotation (4 textes de 10 000 mots) + autres blocs.
  2. Corpus **Frantext** : recherche pour raisons pédagogiques.
- **Difficultés**
  - Phénomène **sans indice sur la forme de surface** : difficulté de trouver des exemples si le corpus n'est pas annoté.
  - Si on repère une forme de surface qui est typique : nécessité de faire un **filtrage** avec une analyse linguistique.

3. Définitions : Qu'est-ce qu'une chaîne de coréférence ?

- **Référence** Désigner un référent mentionné dans le discours au moyen d'une **expression référentielle (ER)** (Karttunen 1976)
- **Coréférence** Relation **symétrique** entre deux ER qui désignent la même entité (Corblin 1985)
- **Anaphore** Relation **asymétrique** : implique de se référer à un élément présent dans le discours pour en interpréter un autre
  - ✓ **Coréférence non anaphorique** : *Paris est la capitale de la France. La Ville Lumière est située en Île-de-France.*
  - ✓ **Anaphore coréférente** : *Jean est allé au restaurant. Il a mangé des lasagnes.*
  - ❌ **Anaphore non coréférente** : *J'ai acheté du pain. La croûte était trop cuite.* (anaphore associative)
  - ❌ **Anaphore sans expression référentielle comme antécédent** :
    Anne a quitté son travail. Cet événement a marqué un tournant dans sa vie. (anaphore résomptive)
- **Chaînes de (co)référence(s)** L'ensemble des **ER** coréférentes à un même référent = **maillons** (**mentions** en TAL)
  - Ex : "Marthe était à moi ; ce n'est pas moi qui l'avais dit, c'était elle." (*Le Diable au corps*)
  - ER non coréférentes = **singletons**

4. La (co)référence non stricte

- **Coréférence** → Le référent est **strictement le même** pour deux expressions référentielles.

- **Coréférence non stricte** Le référent est n'est **pas strictement le même** pour deux expressions référentielles.

  Problèmes d'identification du référent :

  - **Ambiguïté référentielle** : **Choix** entre plusieurs référents candidats potentiels (Fuchs 1996).

  - **Near identity** (Recasens 2011) : Référents proches (métonymie, méronymie…). 3 degrés de coréférence :
    - 3 : total identity (same referent)
    - 2 : strong near identity (almost the same)
    - 1 : weak identity (rôle, representation, gpe, component, location, other)

  ✓ **Flou référentiel** : Référent pas clairement identifiable (désigné de manière imprécise) : **pas de choix** imposé. (Landragin, 2007)

5. Annotation de la coréférence
5.1. Concepts dans les corpus existants

- **ACE** (Doddington et al., 2004) : 5 types de relations (**rôle**, **partie**, **localisation**, **proche** et **sociale**) + coréférence **stricte** ou **métonymie**.
- **OntoNotes** (Pradhan et al., 2011) : coréférence **identique** ≠ coréférence **appositive** .
- **WikiCoref** (Ghaddar & Langlais, 2016) : coréférence **identique** ≠ coréférence **attributive** ≠ coréférence **attributive dans des constructions copulatives**.
- **ARRAU** (Poesio & Artstein, 2008) & **Phrase Detectives** (Chamberlain et al., 2016) : scores sur les **ambiguïtés référentielles** en fonction des avis des annotateurs.
- **Polish Coreference Corpus** (Ogrodniczuk et al., 2015) : coréférence **identique** ≠ coréférence **quasi-identique**.
- **ANCOR** (Muzerelle et al., 2013) : 5 types d'anaphores : **directes**, **indirectes**, **pronominales**, **associatives** et **associatives pronominales.**

5. Annotation de la coréférence
5.1. Concepts retenus

1. **Coréférence identique / exacte**

Le référent est le même pour chaque **expression référentielle** (= **maillon** = **mention**), sans aucun doute :

*« **L'enfant** n'en était pas moins heureuse. **Elle** n'avait pas osé le montrer devant **son** grand-père, mais, après son départ, **elle** s'était livrée à une danse désordonnée, à laquelle s'était mêlé le chien, et qui avait fait rire aux larmes mère Clarisse. »*

AUDOUX Marguerite, *Douce Lumière*, 1937.

5. Annotation de la coréférence
5.1. Concepts retenus

2. (Co)référence floue

- **Inclusive** ou **non**
- **Référent « flou » mais coréférence exacte** :
  - *« **On** racontait que des gens importants étaient venus mais **on** ne donnait pas de noms. »*
- **Coréférence floue** :
  - *« **Tout le monde** paraissait gai, ce soir-là à Némoville. **On** parcourut les rues-couloirs, toutes éclairées à l'électricité, et **on** pénétra dans un sous-marin que le prêtre n'avait pas encore visité. »* BOURGEOIS Adèle, *Némoville*, 1917.
- **Écueils possibles**
  - Annoter dans la même chaîne des ER non coréférentes de manière stricte alors qu'il y a une nuance.
  - Ne pas relier des maillons qui pourraient très probablement être coréférents.

5. Annotation de la coréférence
5.2. Outils

❌ **CADIXE** (Bessieres et al., 2001) : annotation des entités mais pas des relations anaphoriques.

❌ **MMAX 2** (Müller & Strube, 2006) : pas de fonctionnalité de représentation et d'analyse des chaînes.

❌ **GLOZZ** (Widlöcher & Mathet, 2009) : annotation des entités et relations. Représentation visuelle des chaînes mais nécessité de passer par GLOZZQL. Fonctionne bien sur des textes courts.

- **SACR** (Oberlé, 2018) : annotation simple de la coréférence (drag-and-drop). Pas de visualisation pour l'instant (autre outil en développement)

✅ **ANALEC** (Landragin et al., 2012) : annotation des entités et des relations + analyse et représentation des chaînes mais peu utilisé.

✅ **TXM** (Heiden et al., 2010) : très grosse plateforme de manipulation de corpus largement diffusée. Comporte l'extension URS qui implémente une partie d'analec.

5. Annotation de la coréférence
5.3. Modèle

- **Unités-Relations-Schémas (URS)** : Développé pour GLOZZ puis implémenté dans ANALEC et TXM par extension

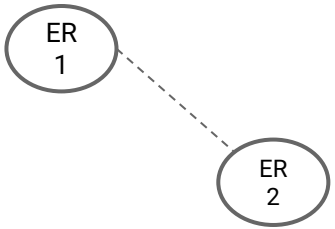
  - **Unités** : Séquence contiguë de mots → **expressions référentielles**
    - ✔ Trait rempli **manuellement** (propriétés de l'ER) : **référent**
      - Traits remplis **automatiquement** : **catégorie grammaticale**, (**code SEM**)

  - **Relations** : Relation de type 1 to 1 (binaire) entre deux éléments (U, R ou S) du modèle
    - Pas dans Democrat mais création automatique possible

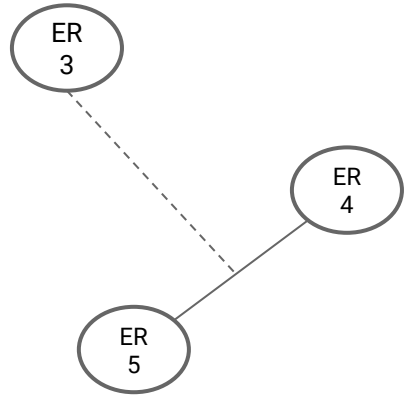
  - **Schémas** : Ensemble d'éléments URS → **chaînes de coréférence**
    - ✔ Traits remplis **automatiquement** lors de la création de la chaîne : **référent**, **nombre de maillons**
      - Traits remplis **manuellement** (propriétés du référent) : **genre**, **nombre**, **type de référent**

5. Annotation de la coréférence
5.4. Annotation du flou : à quel niveau ?

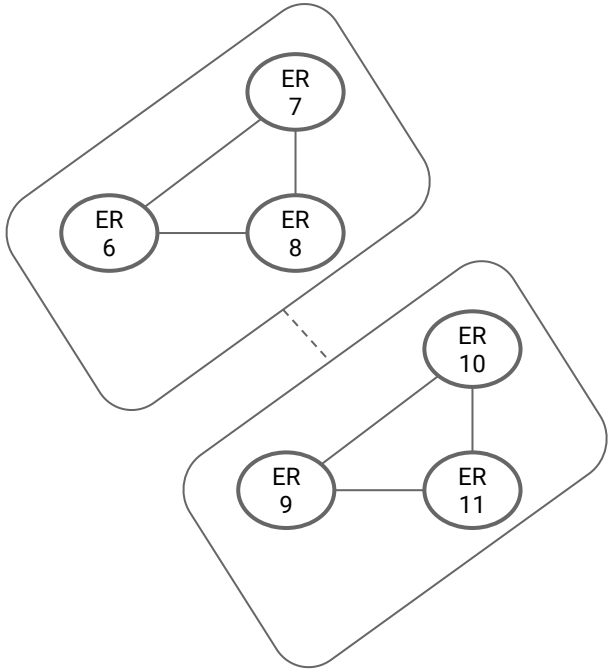
Relation unité-unité



Relation unité-relation :



Relation schéma-schéma :



6. Difficultés rencontrées

- **Délimitation des phénomènes à annoter :**
  - Sous-catégories de coréférence floue ?
  - Prise en compte de la coréférence proche ? De l'ambiguïté ?
- **Difficulté de trouver des exemples non annotés**
  - Pas/peu d'indices sur la forme de surface
  - Beaucoup de ON flous, mais pas tous !

7. Discussion

- **Conclusion**
  - **Prise en compte des difficultés d'annotation du flou** : permettre la notion de doute sur la relation de coréférence entre plusieurs maillons potentiels.
  - **Groupes flous typiques** : pronoms ON et NOUS, groupes de personnes (soldats, foule, famille, enfants…), instances/organisation (entreprise, pays…), etc.
  - **Flou sur du singulier** : pas impossible mais peu rencontré pour le moment.
- **Perspectives**
  - Petit corpus analysé de manière fine à visée linguistique mais pas pour un grand corpus de TAL.

Bibliographie

- Bessières, P., Nazarenko, A., & Nédellec, C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. *Actes du 4e Colloque International sur le Document Électronique*, 1-11.
- Chamberlain, J., Poesio, M., & Kruschwitz, U. (2016). Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2039-2046. Portorož, Slovenia.
- Corblin, F. (1985). Remarques sur la notion d'anaphore. *Revue québécoise de linguistique*, *15*(1), 173–195.
- Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 837-840. Lisbon, Portugal.
- Fuchs, C. (1996). *Les ambiguïtés du français*. Ophrys.
- Ghaddar, A., & Langlais, P. (2016). *WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles*. Présenté à Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.
- Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *Proc. of 10th International Conference on the Statistical Analysis of Textual Data*, *2*, 1021-1032. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.
- Karttunen, L. (1976). Discourse Referents. In J. D. McCawley (Éd.), *Syntax and Semantics Vol. 7* (p. 363–386). Academic Press.
- Landragin, F. (2007). L'anaphore à antécédent flou: une caractérisation et ses conséquences sur l'annotation des relations anaphoriques. *Journée d' étude de l'Association pour le Traitement Automatique des LAngues (ATALA) sur la résolution des anaphores*, 3. Paris, France.
- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, (92), 11-15.

Bibliographie

- Landragin, F., Poibeau, T., & Victorri, B. (2012). ANALEC: a New Tool for the Dynamic Annotation of Textual Data. European Language Resources Association (ELRA). *International Conference on Language Resources and Evaluation*, 357-362. Istanbul, Turkey.
- Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*.
- Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2011). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (Éd.), *20e conférence sur le Traitement Automatique des Langues Naturelles* (p. 555-563). Les Sable d'Olonne, France: ATALA.
- Oberle, B. (2018). SACR: A Drag-and-Drop Based Tool for Coreference Annotation. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*. Présenté à LREC, Miyazaki, Japan.
- Ogrodniczuk, M., Glowinska, K., Kopec, M., Savary, A., & Zawislawska, M. (2014). *Coreference: Annotation, Resolution and Evaluation in Polish*. Walter de Gruyter GmbH & Co KG.
- Poesio, M. & Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In Proc. of LREC, Marrakesh.
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, 1-27. Portland, Oregon, USA.
- Recasens, M., Hovy, E., & Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, *121*(6), 1138–1152.
- Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz: environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelle*, *session posters*, 10. Senlis, France.