Un exemple d'annotation sur corpus écrit
Référence et chaînes de référence

Frédéric Landragin

CNRS

Lattice
Langues, Textes, Traitements Informatiques, Cognition

Ecole thématique « Annotation de données langagières »

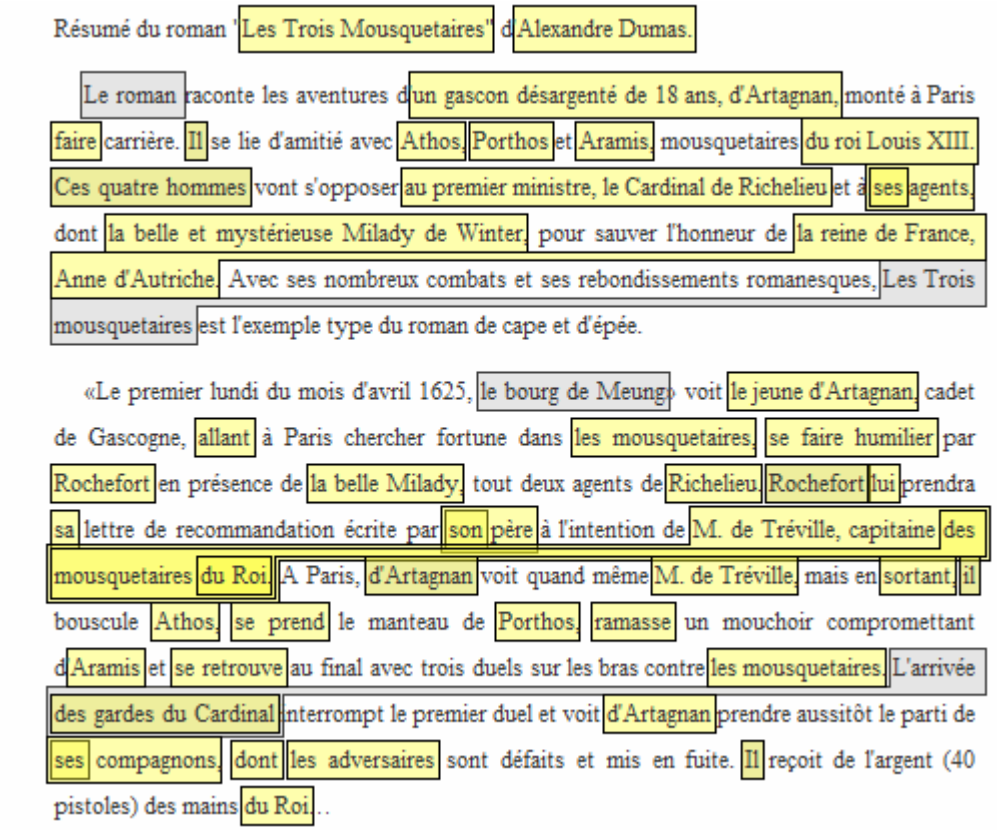
Biarritz, du 10 au 16 septembre 2011

Introduction

- Bilan d'une expérience d'annotation, avec les problèmes rencontrés et les solutions adoptées :
  - délimitation des phénomènes à annoter
  - définition d'une méthodologie
  - choix d'outils d'annotation
- Liens avec les préoccupations de l'école thématique :
  - comment délimiter ce qu'il est utile d'annoter ?
  - que peut-on annoter ? suivant quelle perspective théorique ?
  - qu'apporte l'utilisation de données annotées ?
  - quelle méthode d'annotation utiliser ?
  - comment concilier approches manuelles et automatiques ?
  - quelle exploitation peut-on faire du corpus annoté ?

La référence

- Expression qui désigne un personnage, un objet concret ou abstrait, un lieu, un événement…
- Exemple avec les individus :
  - Alexandre Dumas
  - D'Artagnan
  - Athos
  - Porthos
  - Aramis
  - le roi Louis XIII
  - Richelieu
  - et les groupes…



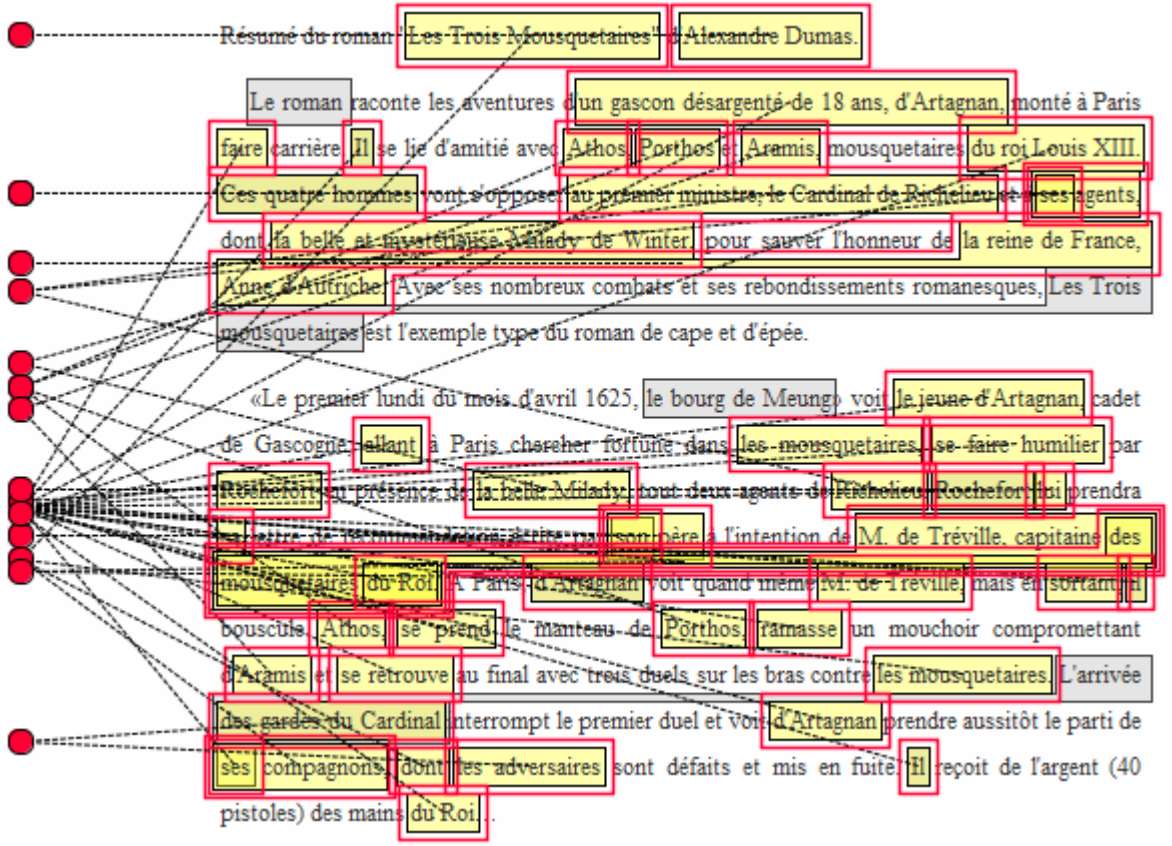
Résumé du roman "Les Trois Mousquetaires" d'Alexandre Dumas.

Le roman raconte les aventures d'un gascon désargenté de 18 ans, d'Artagnan, monté à Paris faire carrière. Il se lie d'amitié avec Athos, Porthos et Aramis, mousquetaires du roi Louis XIII. Ces quatre hommes vont s'opposer au premier ministre, le Cardinal de Richelieu et à ses agents, dont la belle et mystérieuse Milady de Winter, pour sauver l'honneur de la reine de France, Anne d'Autriche. Avec ses nombreux combats et ses rebondissements romanesques, Les Trois mousquetaires est l'exemple type du roman de cape et d'épée.

«Le premier lundi du mois d'avril 1625, le bourg de Meung voit le jeune d'Artagnan, cadet de Gascogne, allant à Paris chercher fortune dans les mousquetaires, se faire humilier par Rochefort en présence de la belle Milady, tout deux agents de Richelieu. Rochefort lui prendra sa lettre de recommandation écrite par son père à l'intention de M. de Tréville, capitaine des mousquetaires du Roi. A Paris, d'Artagnan voit quand même M. de Tréville, mais en sortant, il bouscule Athos, se prend le manteau de Porthos, ramasse un mouchoir compromettant d'Aramis et se retrouve au final avec trois duels sur les bras contre les mousquetaires. L'arrivée des gardes du Cardinal interrompt le premier duel et voit d'Artagnan prendre aussitôt le parti de ses compagnons, dont les adversaires sont défaits et mis en fuite. Il reçoit de l'argent (40 pistoles) des mains du Roi..

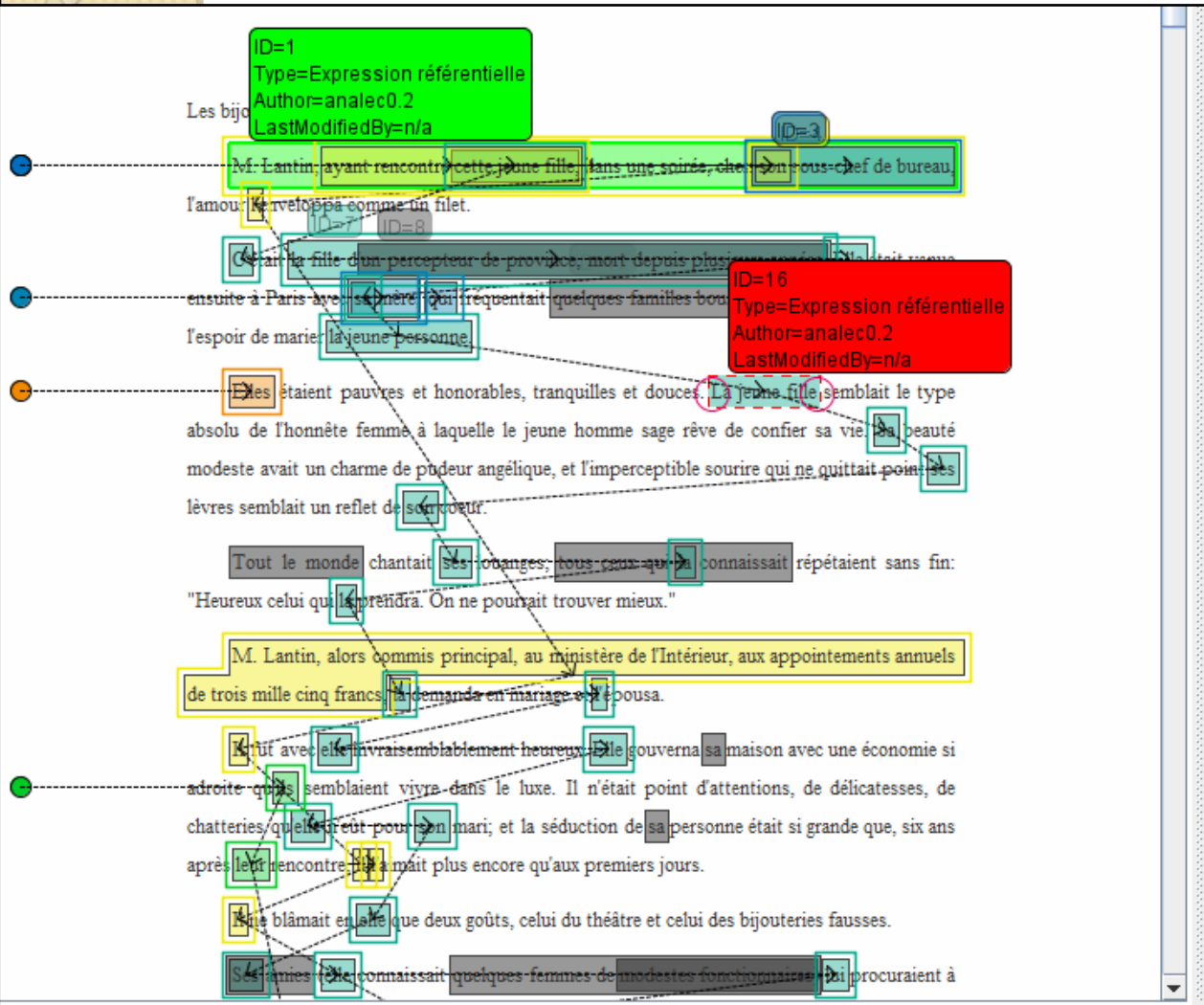
Les chaînes de (co)référence

- Une chaîne regroupe les mentions d'un même référent
- La coréférence est un facteur de cohérence/cohésion
- Exemple :
  - une chaîne par individu
  - une chaîne par groupe d'individus
- Autre étude :
  - une chaîne par lieu
  - une chaîne par date…



Les annotations

- Un ensemble de traits pour chaque élément délimité



Objectifs de l'étude

- Nature des expressions référentielles :
  - qu'est-ce qui réfère dans un texte ?
  - qu'est-ce qui évoque un référent sans pour autant référer ?
  - si on distingue plusieurs degrés de référence, comment en tenir compte dans une méthodologie d'annotation de corpus ?
- Nature des chaînes de coréférence :
  - comment une chaîne commence-t-elle ? se termine-t-elle ?
  - quelles sont les typologies des chaînes ?
  - comment les chaînes se croisent-elles dans un texte ?
  - peut-on prévoir des motifs dans la succession des références ?
  - quelles sont les corrélations entre typologies des chaînes et données syntaxiques, sémantiques, pragmatiques ?
  - comment définir de manière opérationnelle la saillance ?

Objectifs de cette présentation

- Annotation des expressions référentielles :
  - sans entrer dans les théories linguistiques sur la référence, montrer quelques problèmes méthodologiques d'annotation de phénomènes référentiels
  - compte tenu de la nature sémantique des phénomènes concernés, souligner les difficultés auxquelles un annotateur est confronté :
    - problèmes de décision : part importante de l'interprétation du lecteur
    - problèmes de temps : trop de données à annoter
    - problèmes ergonomiques : nécessité de rationaliser la procédure d'annotation
    - problèmes techniques : conciliation des aspects manuels et automatiques
- Annotation et analyse des chaînes de coréférence :
  - montrer comment séparer ce qui relève de données à annoter et ce qui relève de calculs à effectuer sur ces données
  - exploiter des outils de visualisation, de détection de corrélations…

Listes des problèmes rencontrés au cours de l'étude

- Définition de l'objet d'étude, la référence
- Délimitation des expressions référentielles
- Impossibilité de déterminer le référent
- Impossibilité de figer la référence en une annotation
- Prise en compte de certaines formes linguistiques qui ne rentrent pas dans le schéma prévu
- Impossibilité d'annoter telle quelle une notion linguistique (trop abstraite, trop éloignée du matériau linguistique)
- Besoins croissants quant aux outils d'annotation

Problèmes de définition de l'objet d'étude, la référence

- C'est un problème linguistique, qui a des conséquences sur la procédure d'annotation
- Analyse d'un premier exemple :
  - « Pierre et Paul ont chacun eu un fils cette année.
    Il se trouve qu'ils ont la même nourrice. »
  - personnages : Pierre, le fils de Pierre, Paul, le fils de Paul, la nourrice
  - « Pierre et Paul » : du fait de la coordination, faut-il considérer qu'il y a référence à un groupe d'individus ?
  - « un fils » : est-ce référentiel ?
  - « ils » : réfère apparemment au groupe des deux fils, sauf que ce groupe n'a pas été évoqué précédemment. Est-ce pour autant une première mention (ie. le premier maillon de la chaîne) ?
  - « chacun » ?

Problèmes de délimitation d'une expression référentielle

Quelques exemples de premières mentions :

1. **Le président Jacques Chirac** a dit…
2. **Le président de la République, Jacques Chirac**, a dit…
3. **Jacques Chirac, président de la République,** a dit…
4. **Jacques Chirac** – eh oui ! – **président de la République,** a dit…
5. **Le président de la République, qui s'appelle Jacques Chirac,** a dit…
6. **Cet imbécile de président** a dit…
7. **Jacques Chirac** est **le premier président qui a été maire de Paris**.

Plusieurs possibilités selon les cas :

- une seule expression référentielle (qui groupe parfois plusieurs syntagmes)
- plusieurs expressions référentielles
- plusieurs expressions, seule la 1ère étant considérée comme référentielle
- plusieurs expressions, la plus directe (nom propre) étant considérée référentielle

Problèmes d'annotation :

- on a parfois du mal à déterminer des limites précises → borne inf et borne sup
- l'exemple avec du texte discontinu pose des problèmes techniques

Ce qui n'est pas réalisé en délimitant peut l'être avec un trait

- Si on repère une seule expression référentielle, alors on peut prévoir un trait pour annoter sa structure, selon les cas :
  - N_{titre} N_{propre}
  - N_{titre} ponctuation N_{propre}
  - N_{propre} ponctuation N_{titre}
  - N_{propre} relative
  - $N_{qualité}$ de N_{propre}
- Si on repère une seule expression et un ou plusieurs ajouts, alors il faut repérer les ajouts selon une typologie : dénomination, attribution…
- Si on repère plusieurs expressions, alors se pose la question de les intégrer toutes dans la chaîne de coréférence du référent en question
- **Dans tous les cas, on met au point des règles pour que les annotateurs puissent résoudre au mieux les cas difficiles**

Attribuer un référent peut s'avérer impossible

- Certains pronoms peuvent rester ambigus, même en tenant compte des connaissances encyclopédiques d'un lecteur averti
- Exemple : résumé du film « *Le cave se rebiffe* »

Eric Masson, un "demi-sel", est devenu l'amant de la belle Solange Mideau, femme d'un graveur raté. Eric veut se servir de Robert Mideau pour monter, à **son** insu, un trafic de fausse monnaie. Il s'associe à Charles Lepicard, tenancier d'une ancienne maison close, et à Lucas Malvoisin, l'homme d'affaires de celui-ci. Charles et Lucas n'ont pas grande confiance en Eric, mais Solange **leur** promet **son** concours. Elle souhaite en effet mener la grande vie. Avec l'accord de **ses complices**, Charles contacte Ferdinand Maréchal, dit le Dabe, vieux truand célèbre qui s'est retiré dans une île des Tropiques. Il le décide à venir à Paris.

- « à son insu » : Robert Mideau ou Solange Mideau ? pas si simple…
- « leur » : Charles (sûr) + Lucas (sûr) + Eric (peu probable, mais possible)
- « son concours » : ambigu entre Solange et Eric
- « ses complices » : Lucas (sûr) + Solange (pas si sûr) + Eric (?)
- question subsidiaire : qui est « le cave » ?

→ rôle du titre ?

Attribuer un référent peut évoluer au fur et à mesure de la lecture

- Il arrive qu'en poursuivant la lecture d'un texte, on remette en question une référence qui semblait pourtant non ambiguë

L'ancien président de la République de Côte d'Ivoire, Henri Konan Bédié et son épouse ont reçu à dîner l'ancien Premier ministre Alassane Dramane Ouattara et son épouse, le 23 septembre. La rencontre très médiatisée avait un objectif, celui de montrer que **les héritiers du premier président de Côte d'Ivoire** peuvent se retrouver pour reconquérir le pouvoir. **Les deux leaders** ont l'habitude de se voir et de s'appeler depuis le déclenchement, le 19 septembre 2002 de la rébellion en Côte d'Ivoire. A Paris, à Abidjan, à Accra, **les deux hommes** se côtoient, mais dans des cadres formels. **Leur** rencontre en soi n'est donc pas un événement, sauf qu'**ils** ont voulu donner à cette entrevue un cachet particulier. Les retrouvailles autour d'un même idéal politique que commande la mémoire du "Vieux" dont **ils** se réclament. […]
Mais après que tout le monde ait perdu le pouvoir, en faveur d'**un autre héritier, le général Robert Guéi**, par un coup d'Etat en décembre 1999, la gestion du pays semble échapper aux "**enfants**".

- au début : « les héritiers » = H.K.B. + A.D.O.
- il y a du flou : « les héritiers » = H.K.B. + A.D.O. + leurs femmes
- puis : « un autre héritier » = R.G., d'où :
  - nécessairement, « les héritiers » = H.K.B. + A.D.O. + R.G. + ?
  - et, finalement : « les héritiers » = groupe de personnes aux limites floues, qui comprend au moins les trois hommes cités

Conséquences sur l'annotation : plusieurs stratégies sont possibles

1. **On se focalise sur les formes linguistiques**, sans tenir compte des éventuelles ré-interprétations ultérieures (stratégie linéaire)
   - avantage : théoriquement, on réduit les biais interprétatifs et on rend compte des étapes de l'interprétation
   - inconvénients : vouloir attribuer un référent sur la seule forme linguistique est illusoire, car nos connaissances encyclopédiques interviennent constamment ; annoter quelque chose de faux n'est pas très pertinent…
2. **On se focalise sur les concepts**, et on n'annote qu'après avoir compris tout le texte et calculé toutes les références
   - avantage : on se rapproche de la réalité modélisée derrière le texte
   - inconvénient : on s'éloigne des effets stylistiques voulus par l'auteur
3. **On part des concepts et on élargit aux interprétations possibles**, via un attribut dédié (interprétation immédiate vs. différée)
   - avantage : on modélise de manière complète
   - inconvénient : rédiger un manuel d'annotation peut s'avérer compliqué

Conséquences sur l'annotation : c'est une opération floue…

- On ne tente pas donc d'attribuer un référent à tout prix, mais on prend en compte les possibilités d'ambiguïté, d'imprécision, de flou
- On prend en compte la notion de flou, d'une part pour la détermination des groupes (groupe strict versus groupe flou), d'autre part pour la relation d'appartenance à un groupe (appartenance stricte versus floue)
- On modélise ces aspects avec la Théorie des Ensembles Flous (Zadeh)
  - « Solange Mideau » (référence individuelle sans problème) : $\mathbf{A_{strict}}$
  - « son concours » (ambigu entre Solange et Eric) : $\mathbf{A_{strict}}$ **ou** $\mathbf{B_{strict}}$
  - « le cave » : $\mathbf{A_{flou}}$
  - « Charles et Lucas » (groupe construit) : $\mathbf{groupe_{strict}\ \{\ A_{strict}\ ;\ B_{strict}\ \}}$
  - « ses complices » : $\mathbf{groupe_{strict}\ \{\ A_{strict}\ ;\ B_{strict}\ ;\ C_{flou}\ \}}$
  - « les héritiers » : $\mathbf{groupe_{flou}\ \{\ A_{strict}\ ;\ B_{strict}\ ;\ C_{flou}\ ;\ D_{flou}\ \}}$

Annotation de formes atténuées

- En plus des expressions référentielles, certains mots ou morphèmes participent aux chaînes de coréférences
  - les marques d'accord en genre et/ou en nombre (qui, sans référer, rappellent le référent et participent ainsi aux coréférences) :
    - dans « ils dorment », la terminaison « -ent » rappelle le pluriel
    - à annoter ? via une catégorie spécifique ?
  - les sujets zéro (infinitifs, participiales…) :
    - l'intérêt de les annoter est qu'ils peuvent être saillants et contribuer ainsi fortement aux coréférences
    - on peut alors confronter des exemples tels que « il entra, il prit son chapeau » et « il entra, prit son chapeau »
    - à annoter ? comme on n'annote pas du vide (ni un signe de ponctuation), il faut recourir à une solution telle qu'annoter le verbe en tant que support d'une expression référentielle élidée
  - les constructions pronominales, etc.

Annotation de notion linguistique

- Peut-on annoter la saillance des référents ?
- Première façon de procéder : on ajoute un attribut « statut cognitif » ou « saillance » aux expressions référentielles

  → Quelles valeurs donner à cet attribut ?
  - on peut considérer la saillance comme purement cognitive et commencer par utiliser deux valeurs subjectives (faible et forte)
  - pour que l'annotateur n'hésite pas trop, on peut prévoir un test linguistique simple (entité privilégiée pour une reprise = annotée comme saillante)
- Deuxième façon de procéder :
  - on configure un outil pour qu'il calcule automatiquement la saillance de chaque entité du discours, à partir des valeurs affectées aux traits « fonction grammaticale », « position dans la phrase », « rôle actanciel »…
  - c'est outil qui calcule et qui permet de visualiser les scores obtenus
  - il reste ensuite à vérifier que les scores sont cohérents par rapport aux théories et/ou à des expérimentations psycholinguistiques à définir

A quoi sert un outil d'annotation ?

- outil pour la visualisation, l'impression, le regroupement de corpus, etc.
- outil pour l'annotation des expressions référentielles, des formes atténuées, et des syntagmes qui seront utiles pour rechercher d'éventuelles corrélations : verbes, cadratifs, participiales, subordonnées antéposées, etc.
- outil pour l'identification et l'annotation des chaînes de coréférence
- outil pour l'analyse des chaînes de coréférence, par exemple : trouver toutes les chaînes ayant une expression démonstrative en 3e position ; repérage automatique des passages où le référent considéré a été oublié pendant longtemps avant d'être réactivé ; affichage graphique des imbrications…
- outil pour l'analyse des transitions référentielles, par calcul et/ou annotation : détermination et visualisation des patrons de type « R1 R1 R1 R2 R2 »…
- outil pour l'analyse de la saillance, par calcul à partir des traits annotés
- outil pour les calculs statistiques, la recherche de corrélations, la génération de schémas
- outil pour la gestion de plusieurs annotateurs (calcul de l'accord), etc.

A quoi ressemble une procédure d'annotation ?

1. Import du corpus de travail en mode texte brut, avec éventuellement passage par un analyseur morphosyntaxique
2. Annotation manuelle avec GLOZZ des expressions référentielles et des formes atténuées, en saisissant à chaque fois un certain nombre de caractéristiques (traits)
3. Annotation manuelle avec GLOZZ des chaînes de coréférence
4. Toujours dans GLOZZ, ceux qui le souhaitent peuvent annoter également les verbes, les cadratifs, etc.
5. Import du corpus annoté dans ANALEC, qui procède à certains calculs (repérage automatique des paragraphes et des phrases ; affectation des relations d'appartenance d'un syntagme vers une phrase), et permet de visualiser des vues, des calculs statistiques
6. Dans ANALEC, visualisation des chaînes de coréférence pour analyser les transitions référentielles

A quoi servent les annotations réalisées ?

- Constituer un corpus de référence sur la référence…
- Exemples d'interrogations qui sont autant d'études linguistiques rendues possibles par l'annotation :
  - comment les chaînes de coréférence sont-elles constituées ?
  - à quel rang dans une chaîne un démonstratif apparaît-il ?
  - quelle est la fréquence d'apparition des noms propres ?
  - quelles sont les expansions des expressions référentielles ?
  - quels sont les rôles thématiques privilégiés en début de chaîne ?
  - peut-on prévoir des motifs dans la manière dont les chaînes se croisent ?
  - y a-t-il corrélation entre un genre textuel et un type de chaîne ?
  - etc.

Annotation et sémantique : bilan sur les difficultés de l'annotateur

- **problèmes de décision :** part importante de l'interprétation
  → détailler le manuel d'annotation pour prévoir le maximum de cas et de procédures face aux problèmes les plus fréquents

- **problèmes de cas particuliers :** même si schéma d'annotation a prévu le maximum, on se retrouve parfois face à des cas particuliers…
  → un ensemble de règles doit permettre à l'annotateur de ne pas bloquer

- **problèmes de temps :** trop d'éléments à annoter, trop de traits
  → pas de solution miracle, mais on peut séparer le corpus en plusieurs parties pour plusieurs niveaux de détail

- **problèmes ergonomiques**
  → séparer les tâches : faire toutes les délimitations, puis toutes les chaînes, puis tous les remplissages de traits ; pour chaque phase, choisir l'outil qui permet d'effectuer l'opération avec un minimum de clics souris

- **problèmes techniques**
  → concilier annotations automatiques et annotations manuelles ; entraîner des systèmes d'apprentissage pour des pré-annotations ; multiplier les outils, quitte à passer d'un outil à l'outil lors de l'annotation