Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation

Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

Analyse sémantique

Extraction information : C’est quoi?

- Extraction d'informations
  - Définition : L'extraction d'informations vise à récupérer des données structurées, telles que des événements ou des relations, à partir de textes non structurés.
  - Exemple : Analyse des critiques de produits pour en extraire les caractéristiques et les sentiments correspondants (positifs/négatifs) afin de fournir des informations permettant d'améliorer le produit.

NAMED ENTITY RECOGNITION (NER): LA PIERRE ANGULAIRE DE L'IE

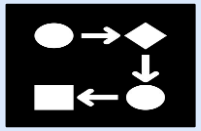
- Traditionnellement, le NER consiste à identifier les noms propres dans les textes et à les classer dans un ensemble de catégories d'intérêt prédéfinies
  - Personne
  - Organisation (entreprises, organisations gouvernementales, comités, etc.)
  - Expression de la date et de l'heure
- D'autres types sont fréquemment ajoutés, en fonction de l'application, par exemple les journaux, les montants monétaires, les pourcentages.

Pourquoi NE est important ?

- Le NER utilise des algorithmes qui fonctionnent sur la base de la grammaire, de modèles statistiques de NLP et de modèles prédictifs.
- Ces algorithmes sont entraînés sur des ensembles de données que les gens étiquettent avec des catégories d'entités nommées prédéfinies, telles que des personnes, des lieux, des organisations, des expressions, des pourcentages et des valeurs monétaires.
- Les catégories sont identifiées par des abréviations ; par exemple, **LOC est utilisé pour les lieux, PER pour les personnes et ORG pour les organisations**

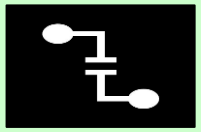
- Le NER fournit aux entreprises des informations essentielles sur leurs clients, leurs produits, leurs concurrents et les tendances du marché. Par exemple, les entreprises l'utilisent pour détecter quand elles sont mentionnées dans des publications. Les prestataires de soins de santé l'utilisent pour extraire des informations médicales clés des dossiers des patients.

Une pipeline NE typique



Prétraitement (tokenisation, sentence splitting, morphological analysis, POS tagging)

Recherche d'entités (gazetteer lookup, NE grammars)



Co-référence (alias finding, orthographic coreference etc.) :
Identification des différentes occurrences d'une entité spécifique dans un texte qui se réfèrent toutes à la même entité

Exportation vers Base de donnes / XML / ontologie

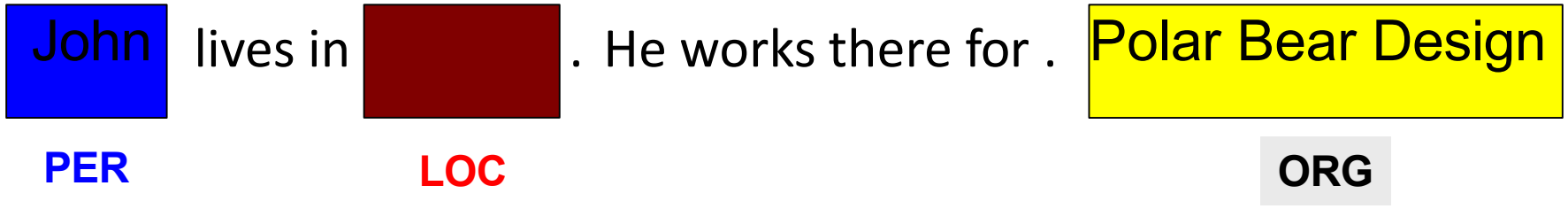
Un exemple de NER

John lives in London. He works there for Polar Bear Design.

Tokenization

John | lives | in | London | . | He | works | there | for | Polar | Bear | Design | .

NE Recognition basique

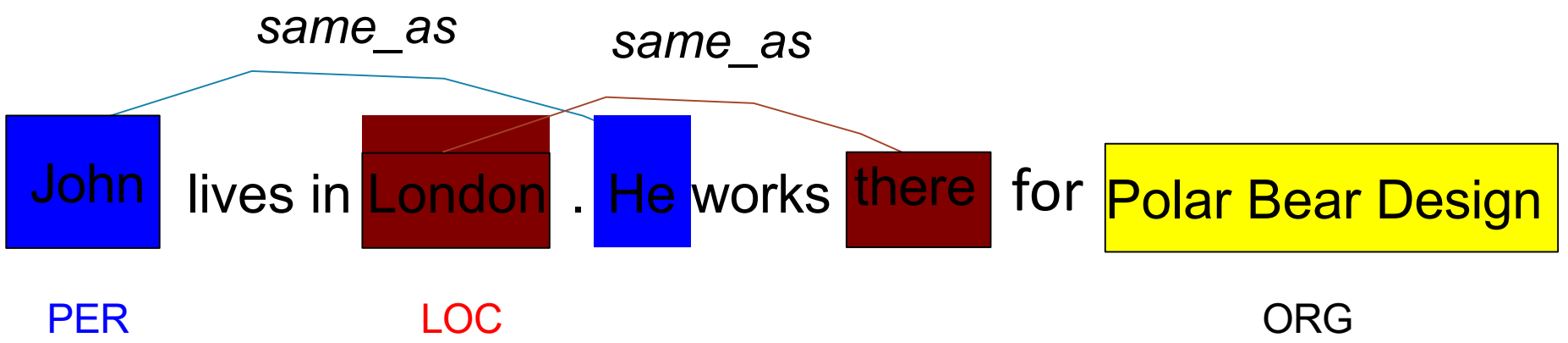


Co-référence

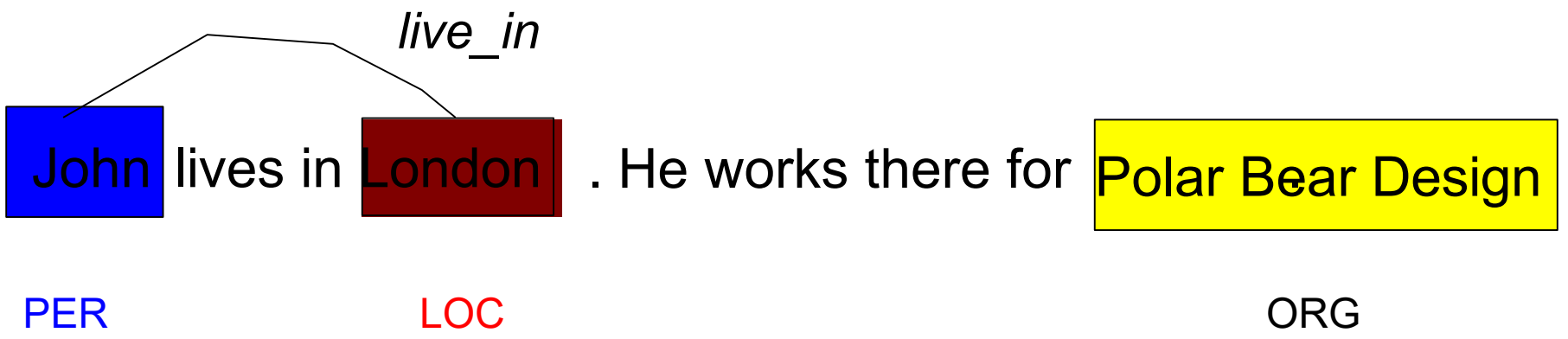
- La résolution des coréférences est un autre problème linguistique que le NLP tente de résoudre.
- Par définition, on parle de coréférence lorsque deux ou plusieurs termes/expressions d'un texte font référence à la même entité.
- On dit alors qu'ils ont le même référent.

- Prenons l'exemple de la phrase suivante : "**Jean vient de me dire qu'il se rend à la salle d'examen"**. Dans cette phrase, le pronom "il" a pour référent "Jean".
- La résolution de ces pronoms fait partie de la résolution de la coréférence et devient un défi dès que nous avons plusieurs référents dans un texte.
- dans un corps de texte. Voici un exemple de texte : **"John vient de parler à Jim. Il m'a dit que nous avions un test surprise demain"**. Dans ce corps de texte, le pronom "il" peut se référer soit à "John", soit à "Jim", ce qui rend difficile l'identification du référent exact

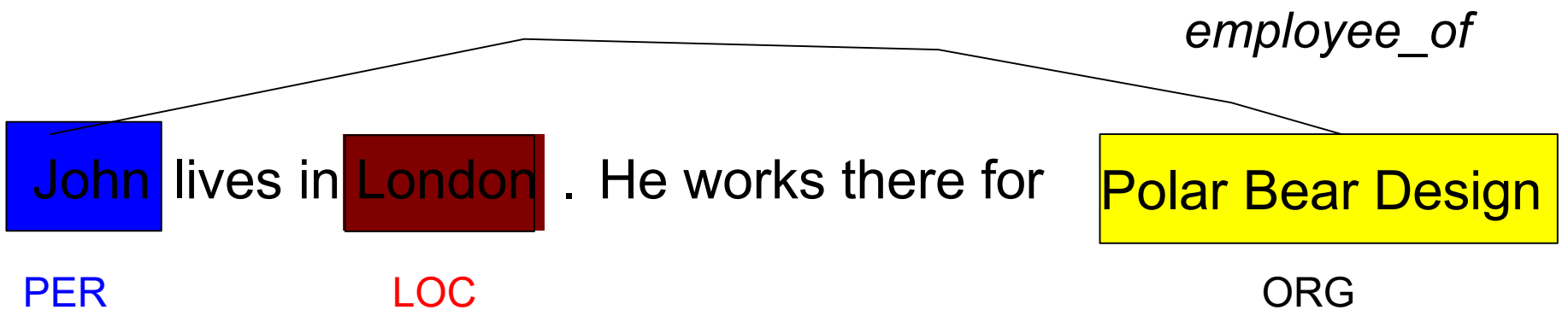
Co-référence



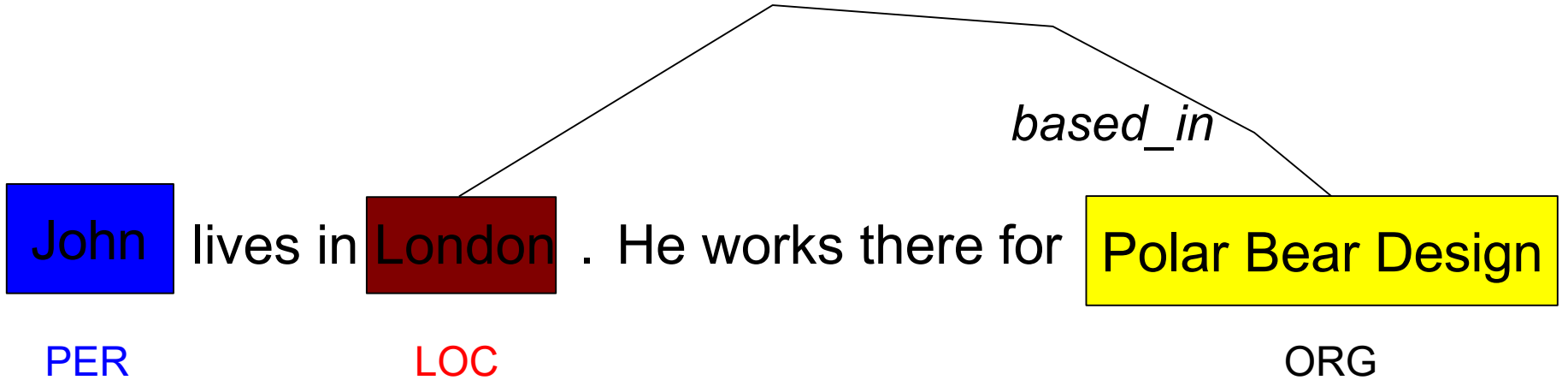
Relations


live_in
John
lives in
London
. He works there for
Polar Bear Design
PER
LOC
ORG

Relations (2)



Relations (3)


based_in
John
lives in
London
. He works there for
Polar Bear Design
PER
LOC
ORG

GAZETTEERS

Les (GAZETTEERS) répertoires géographiques sont des fichiers texte contenant des listes de noms (par exemple, villes, rivières, personnes, ...). Ces listes sont utilisées pour trouver les occurrences de ces noms dans le texte.

- Le répertoire ANNIE contient environ **60 000 entrées réparties dans 80 listes**.
- Chaque liste reflète une certaine catégorie, par exemple les aéroports, les villes, les prénoms, etc.
- Les entrées de la liste peuvent être des entités ou des parties d'entités, ou elles peuvent contenir des informations contextuelles (par exemple, les titres d'emploi indiquent souvent des personnes).
- Chaque répertoire possède un fichier d'index qui répertorie toutes les listes, ainsi que les caractéristiques de chaque liste (type principal, type secondaire et langue).
- Les répertoires toponymiques génèrent des annotations de type "Lookup" avec les caractéristiques pertinentes correspondant à la liste correspondante.
- Les annotations Lookup sont principalement utilisées par le transducteur NE.
- Les listes peuvent être modifiées soit en interne à l'aide de l'éditeur du répertoire toponymique, soit en externe dans votre éditeur préféré.

ANNIE GAZETTEER

- **Ajoutez le PR ANNIE Gazetteer à la fin de votre pipeline.**
- **Réexécutez le pipeline**
- **Recherchez les annotations "Lookup" et examinez leurs caractéristiques.**
- Double-cliquez sur le PR ANNIE (sous Ressources de traitement dans le volet de gauche) pour l'ouvrir.
- Assurez-vous que "Gazetteer Editor" est sélectionné dans l'onglet du bas.
- Dans le volet de gauche (définition linéaire), vous voyez le fichier d'index contenant tous les
- listes.
- Dans le volet de droite, vous voyez le contenu de la liste sélectionnée dans le volet de gauche.
- Les entrées sont en lecture seule.
- Pour modifier les ressources ANNIE, nous devons d'abord en faire une copie que nous pourrons modifier.

ANNIE GAZETTEER

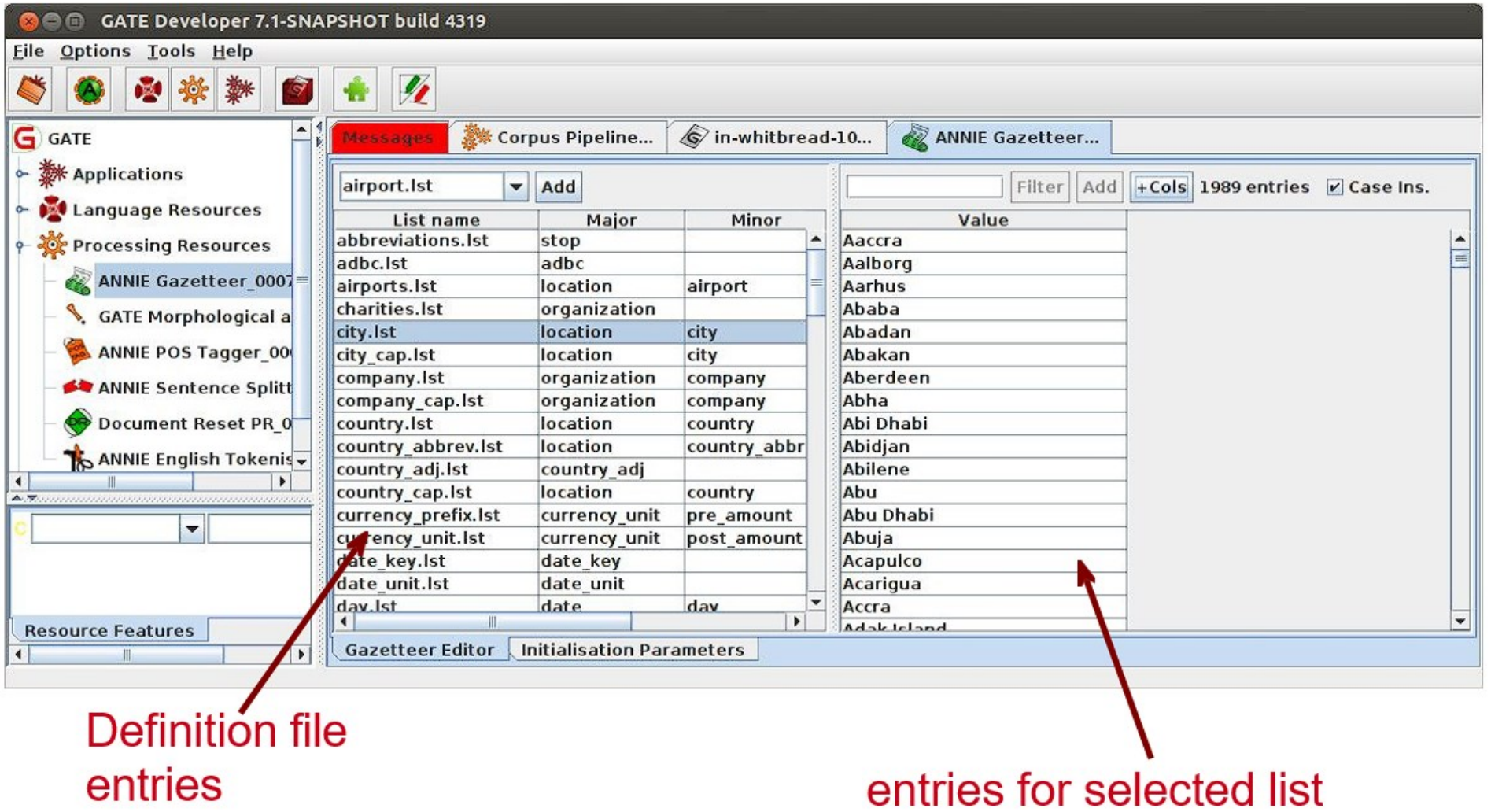
- Ajoutez un autre PR Gazzetter a partir des Processing ressources
- Cliquez sur paramètres d’initilisation
- Cliquez sur le bouton de navigation ListURL
- Allez à l'endroit où vous avez téléchargé le répertoire gazetteer et sélectionnez lists.def et cliquez sur OK.
- Double-cliquez sur ANNIE Gazetteer dans PR.

cliquez sur n'importe quelle liste pour voir les entrées.
Essayez d'ajouter, de supprimer et de modifier les listes existantes ou le fichier de définition des listes.
Pour enregistrer un répertoire gazette modifié, utilisez le raccourci Ctrl-S ou faites un clic droit sur le nom du répertoire gazette dans les onglets en haut ou dans le panneau des ressources à droite, et sélectionnez "Enregistrer et réinitialiser" avant de relancer le répertoire gazette.
Essayez d'ajouter un nouveau mot d'un document que vous avez chargé (qui n'est pas actuellement reconnu comme une recherche) dans le GAZETTER, exécutez à nouveau le GAZETTER et vérifiez les résultats.

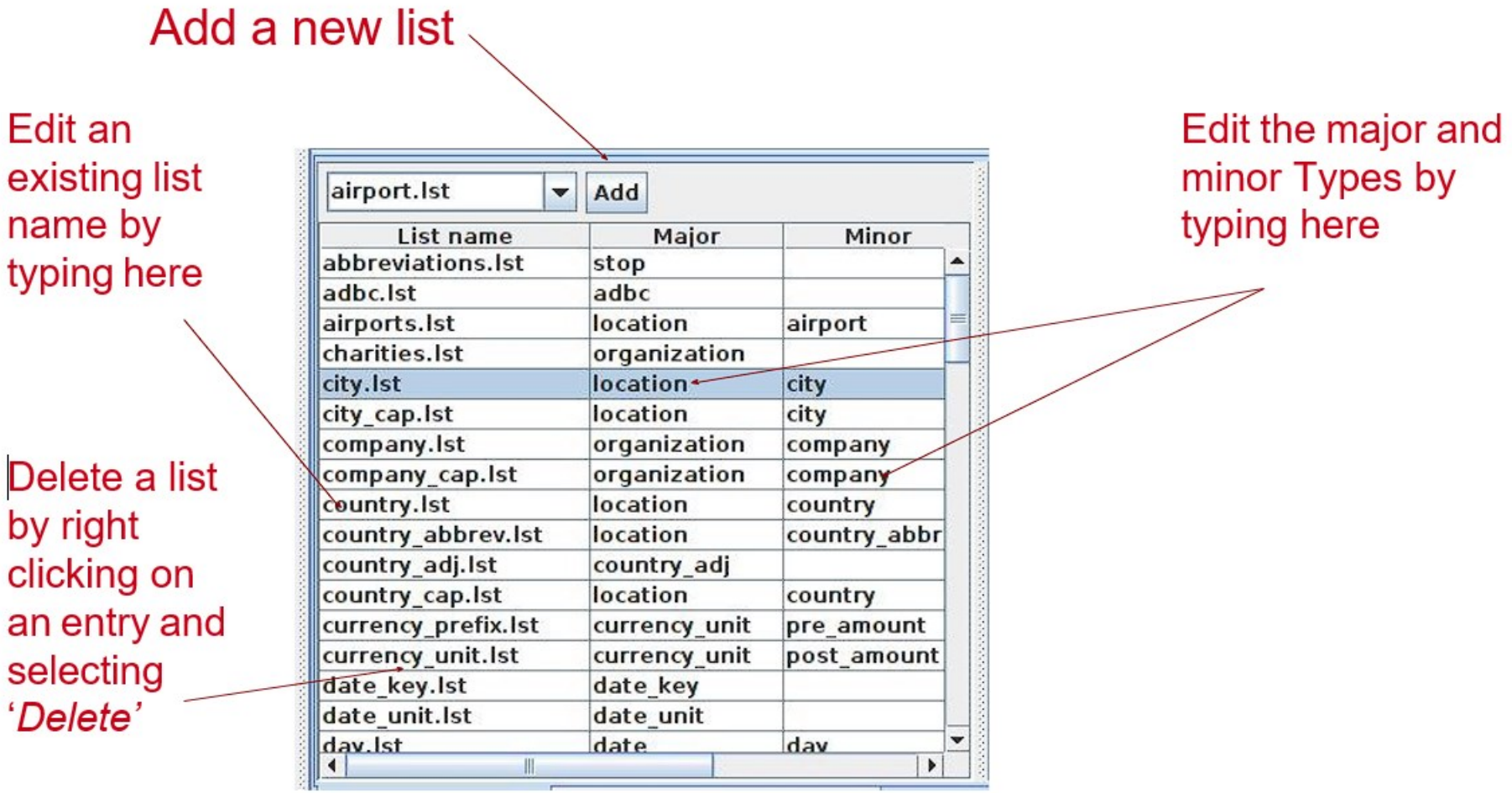
Les attributs de la liste

- Lorsqu'un élément du texte correspond à une entrée du Gazzetter, une annotation Lookup est créée, avec diverses caractéristiques et valeurs.
- Le Gazzetter ANNIE possède les types de caractéristiques par défaut suivants : majorType, minorType, langue.
- Par exemple, la liste "city" a un majorType "location" et un minorType "city", tandis que la liste "country" a pour types "location" et "country".
- Plus tard, dans les grammaires JAPE, nous pourrons nous référer à toutes les listes de recherche de type "lieu", ou nous pourrons être plus spécifiques et nous référer uniquement à celles de type "ville" ou de type "pays".

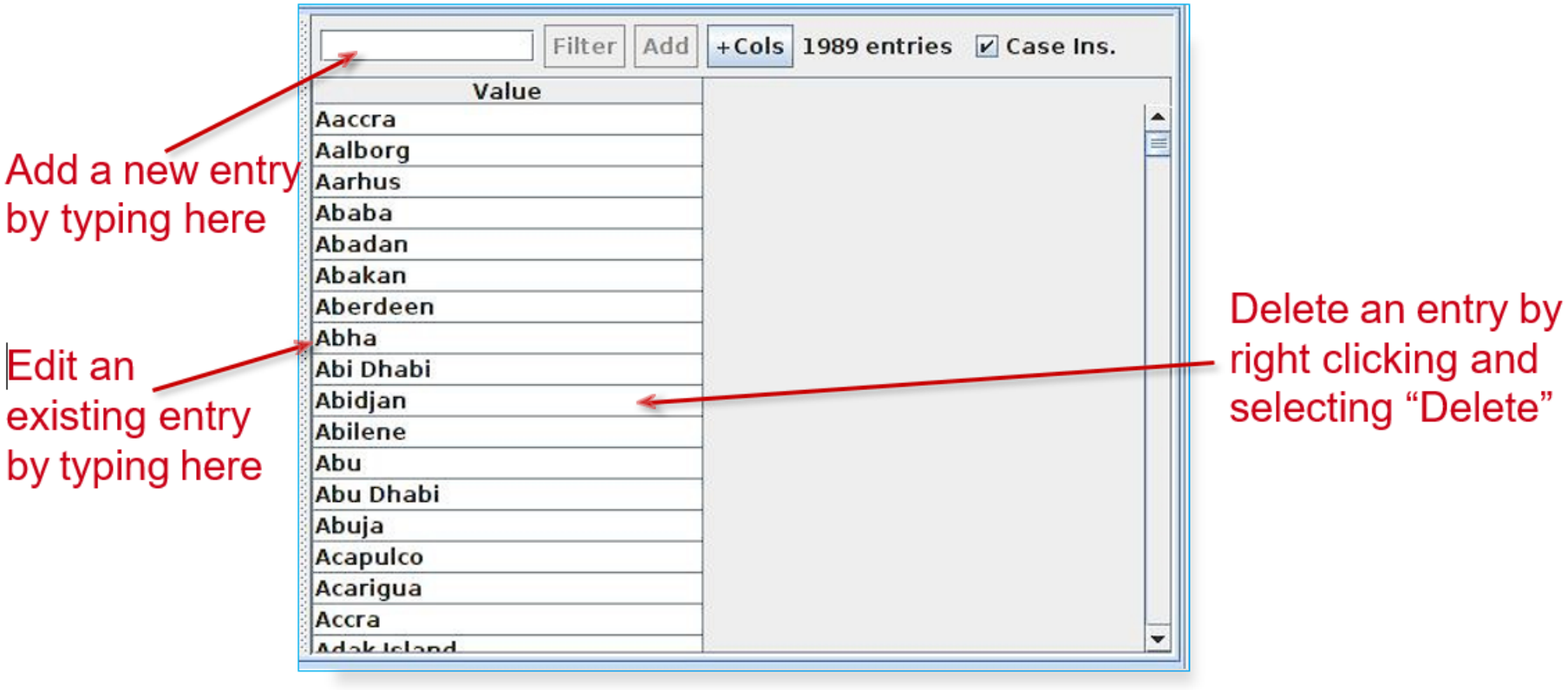
Editeur PR GAZETTEER



Modifier le Fichier des Definitions



Modifier une Liste



NE TRANSDUCER

Les Gazzetters peuvent être utilisés pour trouver des termes qui suggèrent des entités.

- Toutefois, les entrées peuvent souvent être ambiguës.
- "May Jones" vs "May 2010" vs "May I be excused ?" (puis-je être excusé ?)
- "M. Parkinson" vs "Maladie de Parkinson".
- General Motors" vs "General Smith".
- Des grammaires élaborées à la main peuvent être utilisées pour définir des modèles à partir des listes de référence et d'autres annotations.
- Ces modèles peuvent contribuer à la désambiguïsation et combiner différentes annotations, par exemple : les dates peuvent être composées de {jour} + {numéro} + {mois}. {jour} + {nombre} + {mois} Annotations de recherche.
- Le transducteur NE consiste en un certain nombre de grammaires écrites dans le langage JAPE. Le module 2 sera consacré à JAPE.

ANNIE NE TRANSDUCER

- Charger un PR transducteur ANNIE NE L'ajouter à la fin de l'application
- Exécuter l'application
- Regardez les annotations
- Vous devriez voir de nouvelles annotations telles que Personne, Lieu, Date, etc.
- Ces annotations comporteront des caractéristiques indiquant des informations plus spécifiques (par exemple, de quel type d'endroit il s'agit) et les règles qui ont été déclenchées (pour faciliter le débogage).

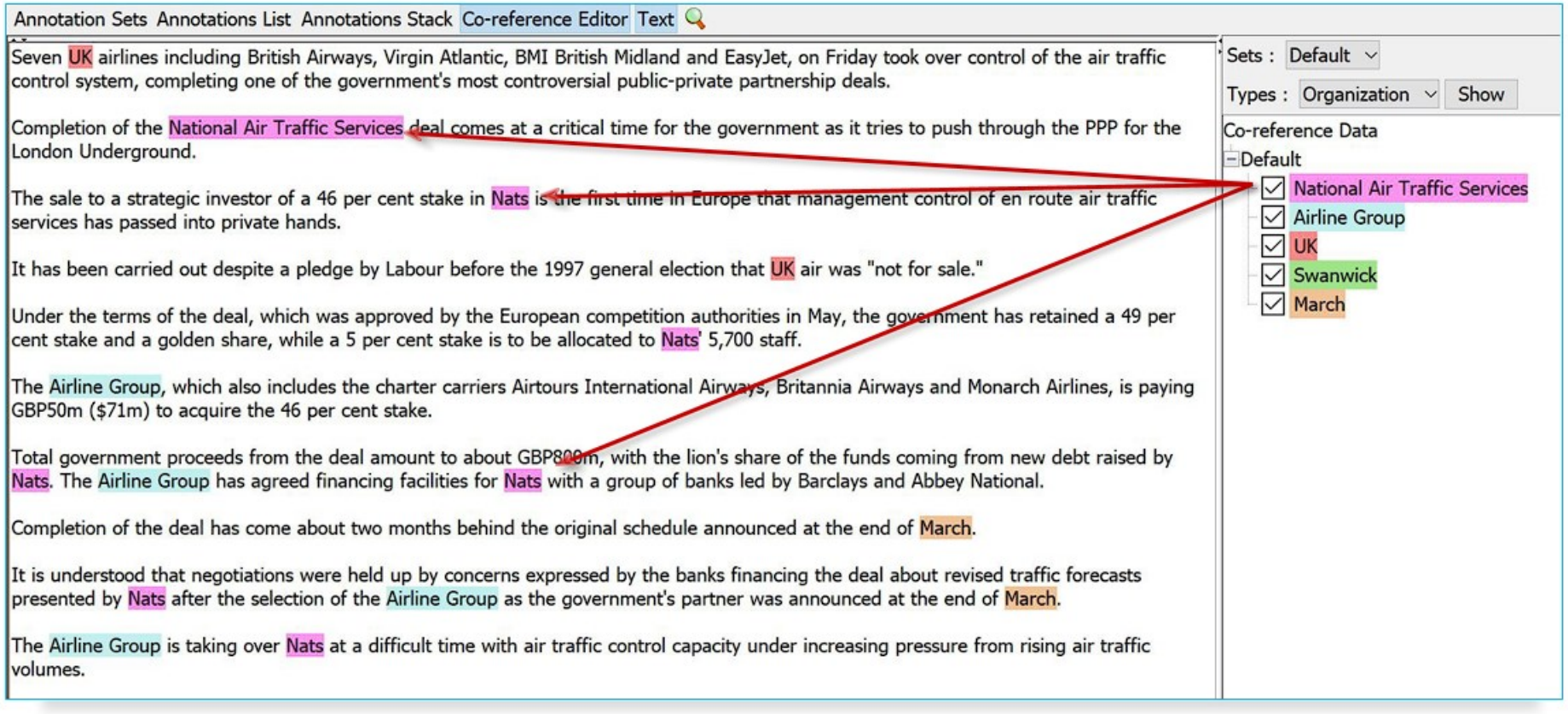
ORTHOMATCHER PR

- Le module de coréférence orthographique (orthomatcher) compare les noms propres et leurs variantes dans un document.
- Effectue la résolution des coréférences sur la base des informations orthographiques des entités
- Produit une liste d'ID d'annotations qui forment une "chaîne" de coréférence
- La liste de ces listes est stockée en tant qu'élément du document appelé "MatchesAnnots".
- Améliore les résultats en attribuant un type d'entité à des noms précédemment non classés, sur la base des relations avec des entités classées.
- La classification des entités inconnues est très utile pour les noms de famille qui correspondent à un nom complet ou à des abréviations,
  par exemple, "Bonfield" <inconnu> correspondra à "Sir Peter Bonfield" <personne>.

EDITEUR DE CO-REFERENCE

- Ajouter un nouveau PR : ANNIE OrthoMatcher.
- Ajoutez-le à la fin de l'application.
- Lancez l'application.
- Dans une vue de document, ouvrez l'éditeur de coréférence en cliquant sur le bouton situé au-dessus du texte.
- Tous les documents du corpus devraient avoir des coréférences, mais certains peuvent en avoir plus que d'autres.

EDITEUR DE CO-REFERENCE



UTILISER EDITEUR DE CO-REFERENCE

- Sélectionnez l'ensemble d'annotations que vous souhaitez visualiser (sélectionnez l'ensemble par défaut pour l'instant).
- Une liste de toutes les chaînes de coréférence basées sur les annotations de l'ensemble sélectionné s'affiche.
- Sélectionnez un élément de la liste pour mettre en évidence toutes les annotations membres de cette chaîne dans le texte (vous pouvez en sélectionner plusieurs à la fois).
- Le survol d'une annotation en surbrillance dans le texte vous permet de supprimer un élément de la chaîne de coréférence.