

# Combinaison d'approches du traitement automatique des langues et de la recherche d'information pour une recherche de livres orientée contenu efficace

EARIA 2014  
OLLAGNIER Anaïs

Thèse encadrée par :  
Pr. Bellot Patrice  
Dr Fournier Sébastien



Laboratoire des Sciences  
de l'Information et  
des Systèmes



OpenEdition



Aix-Marseille  
université



15 octobre 2014

BERKELE, G. (1983). *Die Entwicklung des Ausdrucks von Objektreferenz am Beispiel der Determinanten. Eine empirische Untersuchung zum Spracherwerb bilingualer Kinder (Französisch/ Deutsch)*. Thèse, Université de Hambourg.

BICKERTON, D. (1981). *Roots of language*. Ann Arbor : Karoma. BOTTARI, P., P. CIPRIANI & A. M. CHILOSI (1993). « Protosyntactic devices in the acquisition of Italian free morphology », *Language Acquisition*, 3 : 327- 369.

### Références dans une section spécifique

3 Cette notation se lit : x ans ;y mois,z jours.

4 Les abréviations f., m., n. désignent respectivement le genre féminin, le genre masculin et le neutre.

5 Ces calculs ont été faits à partir du *Petit Larousse* (1969).

6 Le nombre entre parenthèses indique le nombre absolu des noms dont la règle prédit le genre correct

### Références dans les notes

On peut se demander pourquoi les formes de l'article indéfini ne sont intégrées que tardivement dans le système grammatical de l'enfant. Il ressort de notre analyse des articles indéfinis que les syntagmes qui réfèrent à des objets indéterminés ([- défini]), sont marqués par les enfants d'une façon déviante par rapport aux langues cibles. Berkele (1983) montre que les formes de l'article défini sont utilisées pour exprimer la référence spécifique. L'enfant pourrait faire l'hypothèse que [- défini] (ou [- spécifique]) est marqué par  $\emptyset$  et que [+ défini] (ou [+ spécifique]) est indiqué par un déterminant ou un élément modificateur quelconque.<sup>24</sup> Bickerton (1981) suppose l'existence d'un tel système dans

### Références dans le corps du texte

Cross-linking



# Mise en oeuvre

## Reconnaissance des entités nommées

- Évaluation des capacités de généralisation des systèmes de reconnaissance d'entités nommées
- Étude des processus d'adaptation d'un système existant à un nouveau cadre d'utilisation tout en minimisant l'effort d'adaptation (*Ollagnier & al., 2014*)
  - Exploitation des corpus Quaero et des données d'OpenEdition

## Bilbo (*Kim, Bellot et al. 2012*)

- Démonstration en ligne : <http://bilbo.openeditionlab.org>
- Disponible en téléchargement : <https://github.com/OpenEdition/bilbo/>

*Ollagnier, A.; Fournier, S.; Bellot, P. & Béchet, F. Impact de la nature et de la taille des corpus d'apprentissage sur les performances dans la détection automatique des entités nommées Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles, Association pour le Traitement Automatique des Langues, 2014, 511-516*

*Kim, Bellot et al. 2012*

# Mise en oeuvre

## Optimisation de Bilbo

- Paramétrage et sélection de features
- Ajout de features linguistiques
- (*En cours : ajout de features contextuels*)

## Modèle de Recherche d'informations et de recommandation adapté à la recherche de livres par le contenu

- Modèle basé sur les liens (en plus du contenu comme sur les modèles classiques)
- Analyse de requêtes complexes (type Inex Social Book Search – Participation à INEX 2014 Social Book Search Track)

## Conclusion

- Modèles pour la détection de références bibliographiques dans le corps du texte (références incomplètes et peu explicites)
- Analyse des requêtes complexes en langage naturel (analyse en dépendances, recherche de contraintes exprimées dans les requêtes etc.)
- Modèle de RI combinant sac de mots, concepts et contraintes ainsi que l'exploitation des liens

# Merci de votre attention !

Pour nous suivre : <http://lab.hypotheses.org>