L'usage de la morphologie dans les systèmes de Question-Réponse

Delphine Tribout
Delphine Bernhard & Bruno Cartoni

LingLunch
27 octobre 2011

Systèmes de Question Réponse

- Les systèmes de Question-Réponse (QR) ont pour objectif de fournir une réponse précise à une question posée en langue naturelle.

  ex. Q : Quand a été assassiné Henri IV ?
  R : le 14 mai 1610

- Ils reposent généralement sur un composant de recherche d'information (RI) qui apparie les mots de la question avec les mots des documents contenant la réponse potentielle

Problème du fossé lexical

Question

Quand Henri IV a-t-il été assassiné ?

Documents

sdfslkd fhg ksjh dfgkl. Le 14 mai 1610 Ravaillac assassine Henri IV. ghsldfjfh hgsk djfhg rgtjfg bkhs nj

Jh jdhfjgh sd klhg kdjfhg kjh hsd L'assassinat d'Henri IV a eu lieu en 1610. slk dfhgk jsdhf gjhdfjg

hdslkf jhgklsj d hfgklj dhfgj hgkhz lkf Henri IV a été tué en 1610. bg hltgsflk dfhgk jsdhf gjhdfjg kjsbxtf

- Les systèmes de RI et de QR doivent prendre en compte des relations autres que l'identité
  - relations sémantiques : synonymes, hyperonymes...
  - relations morphologiques : flexion, construction

Connaissances morphologiques en RI et QR

- **Méthodes**
  - extension de la requête
    assassiné → assassine assassina assassinat assassin
  - conflation des mots des documents et des requêtes
    assassiné assassinat assassine → assassin

- **Conflation**
  - Raciniseurs [Porter, 1980]

- **Extension de requête**
  - Outils
    - Analyseurs à base de règles [Namer, 2009]
    - Analyseurs non supervisés [Moreau and Claveau, 2006]
  - Ressources
    - Lexiques flexionnels : Morphalou, Le*fff*...
    - Lexiques dérivationnels : Verbaction, Prolexbase, Dubois...

Évaluation des connaissances nécessaires aux systèmes QR

- **Évaluation traditionnelle :** utilisation d'outils et/ou de ressources lexicales, et observation de l'amélioration (ou non) des résultats
  - Pas de distinction entre divers phénomènes morphologiques
  - L'évaluation de la performance du système reste globale
- **Hypothèse :**
  - Toutes les relations morphologiques ne sont pas pertinentes pour la recherche d'information
  - Il faut cibler les relations utilisées pour améliorer un système de QR
- **Une évaluation fine du rôle de la morphologie :**
  - Caractérisation des relations morphologiques dont les systèmes de QR ont besoin
  - Évaluation de la couverture des outils et ressources existants en fonction des relations identifiées
  - Évaluation des résultats d'un système en utilisant des relations ciblées

Plan

Un gold-standard pour les relations morphologiques dans les systèmes de QR

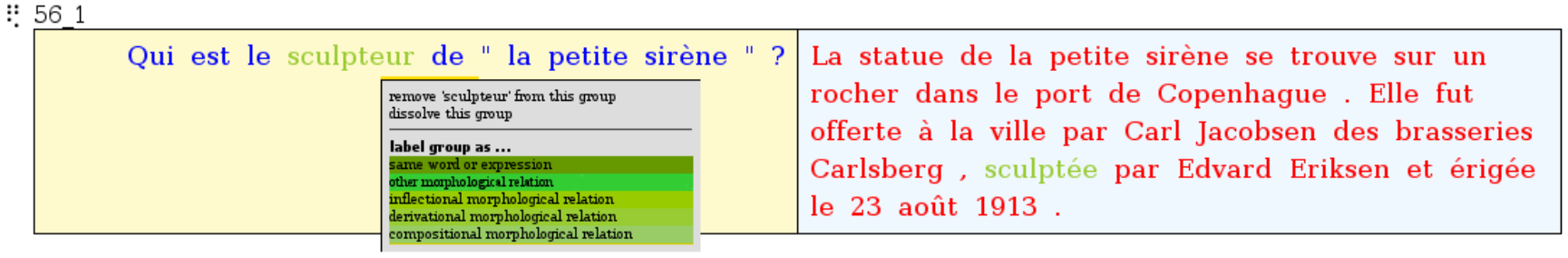
- Annotation de trois corpus de question–passage
- **Corpus Quæro** (Web)
  - réponses retournées par un système QR + validation manuelle (projet Quæro)
  - 566 paires question–passage
- **Corpus EQueR médical** (corpus spécialisé)
  - campagne d'évaluation EQueR–EVALDA, partie QR spécialisé corpus médicaux, QR validée par des spécialistes
  - 394 paires question–passage
- **Corpus Conique** (Wikipedia française)
  - corpus de question–justification (long passage)
  - 664 paires question–passage
- **Total** : 1624 paires

Un gold-standard pour les relations morphologiques dans les systèmes de QR

- Annotation de trois corpus de question–passage
- 3 annotateurs indépendants
- Tâches :
  - relier les mots de la question et les mots du passage–réponse
  - assigner une étiquette :
    - flexion
    - dérivation
    - composition
    - autre
- Utilisation de l'outil YAWAT : **Y**et **A**nother **W**ord **A**lignment **T**ool [Germann, 2008]

Annotation avec YAWAT

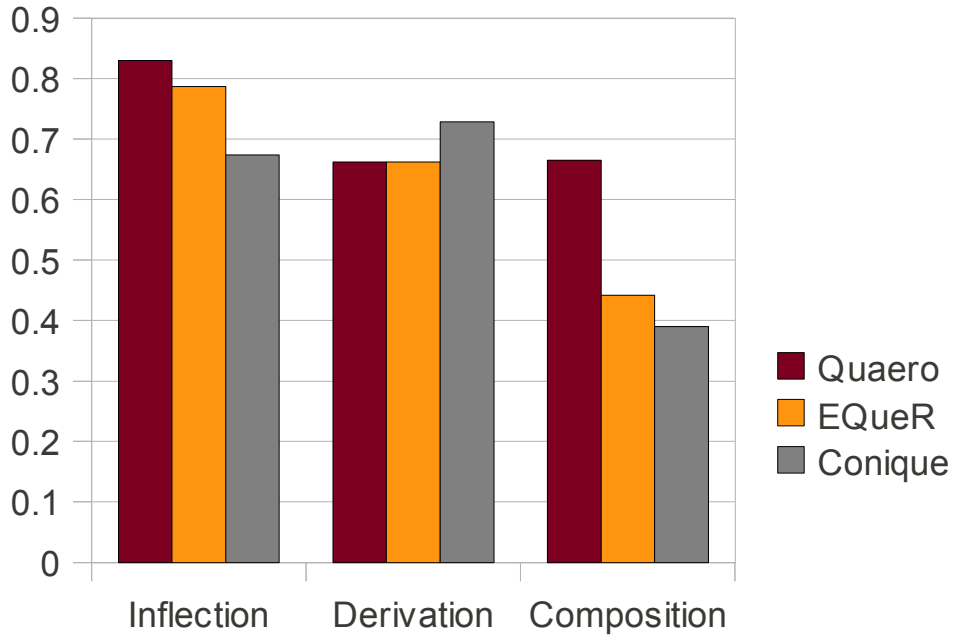
- **Caractéristiques**
  - Créé pour l'alignement de textes parallèles *bilingues*
  - Application Web dynamique
- **Adaptation**
  - Utilisation de notre schéma d'annotation
  - Données parallèles : paires question - passage



Annotation : exemples

Quand est né Philippe d'Orléans ?	Philippe d'Orléans naquit le 2 août 1674.
À combien de milliards de dollars s'élève le déficit budgétaire américain ?	Le PIB des États-Unis s'élève à environ 10 000 milliards de dollars et les déficits atteindraient au moins 300 ou 400 milliards de dollars en 2003
Quels sont les quatre réalisateurs du film "Le jour le plus long" ?	Le Jour le plus long est un film américain réalisé par Ken Annakin, Andrew Marton, Bernhard Wicki et Gerd Oswald sorti en salle en 1962
La pose d'amalgame dentaire peut-elle provoquer des allergies ?	Il est certain que la pose d'amalgames peut entraîner des réactions allergiques plus ou moins graves et prononcées chez les patients.
Où Marcos fut-il dictateur ?	Imelda Romualdez Marcos (née le 2 juillet 1929) fut la femme de Ferdinand Marcos, président-dictateur des Philippines de 1965 à 1986.
Qu'engendre la corticothérapie sur l'os ?	Les fractures de l'ostéoporose cortisonique surviennent au moins en partie en raison d'une perte osseuse induite par la corticothérapie

Annotation : accord inter-annotateur

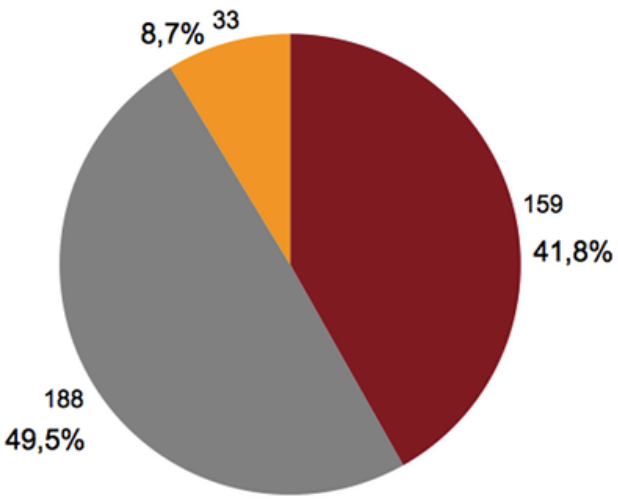


- Kappa de Fleiss
  - Fort à presque parfait pour la flexion
  - Fort pour la dérivation
  - Faible à bon pour la composition
- Les désaccords sont résolus après discussion des annotations

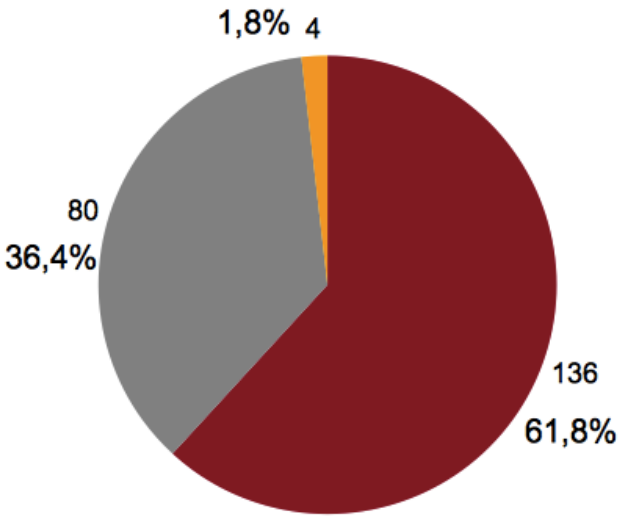
Plan

Types de relations morphologiques

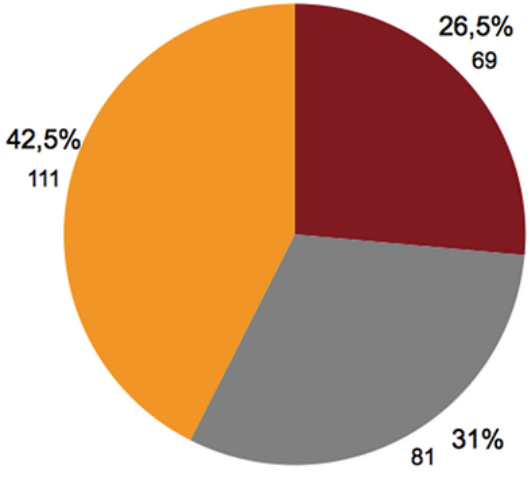
Conique



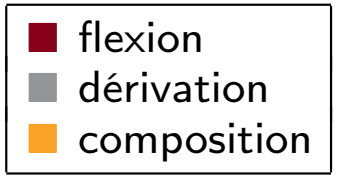
Quæro



EQueR

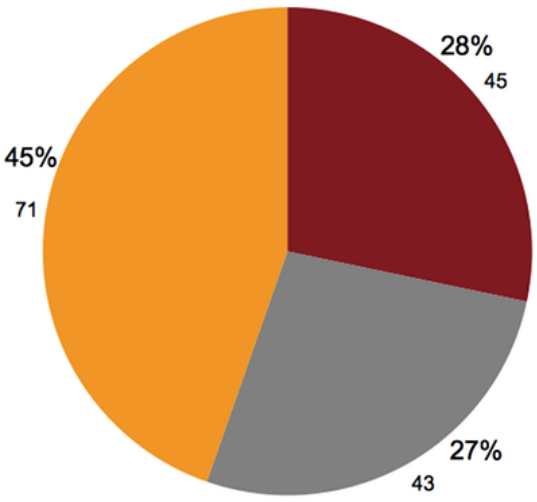


- La dérivation est un phénomène important
- La composition est marginale sauf dans le corpus médical

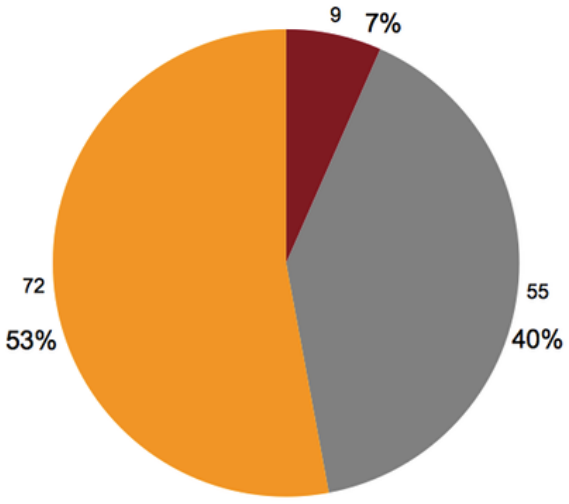


Flexion

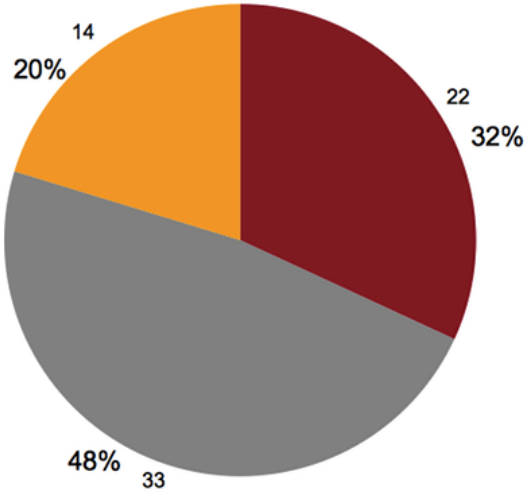
Conique



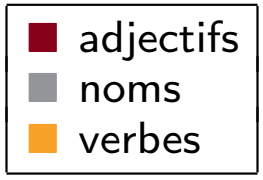
Quæro



EQueR



- La majorité des variations flexionnelles concernent les verbes
- Le corpus médical diffère des deux autres de ce point de vue



Dérivation

- quels sont les types de procédés les plus fréquents ?

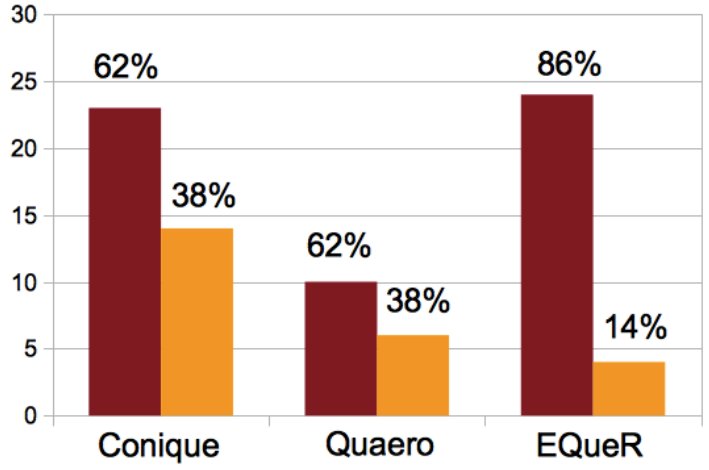
	Exemple	**Conique** (174) %	**Quæro** (70) %	**EQueR** (70) %
N > A	commerce > commercial	21	23	40
N propre > A	Afrique > africain	26	11,5	1
N > N	président > présidente	17	7	3
N propre > N	Arménie > Arménien	3	11,5	3
A > N	national > nationalité	2	0	13
V > N	traiter > traitement	24	43	36
Autres	complet > complètement	7	4	4

- adjectifs dénominaux : 47% dans Conique, 41% dans EQueR
- nominalisations : 61,5% dans Quæro, 54% dans EQueR, 46% dans Conique

Dérivation : cas particuliers

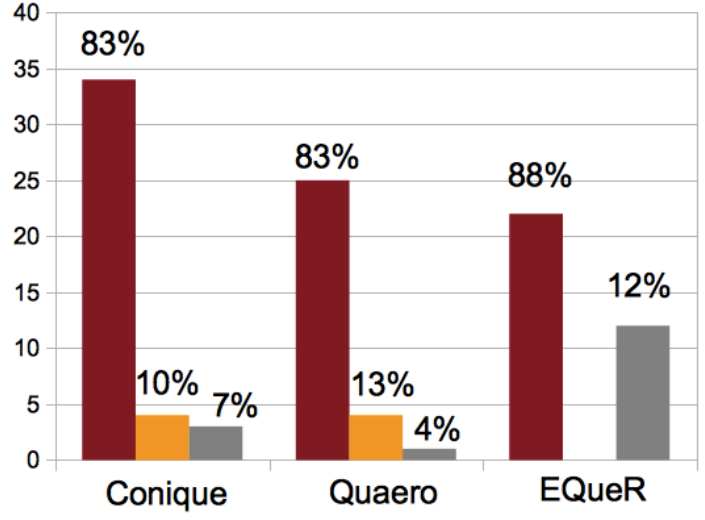
- Adjectifs dénominaux

adj. relationnels	région>régional
adj. qualificatifs	âge>âgé



- Nominalisations

N d'évnt	inaugurer>inauguration
N d'agent	réaliser>réalisateur
N de résultat	produire>produit



Dérivation : cas marginaux et absences

- Cas marginaux
  - A>Adv : *complet>complètement*
  - préfixation : *boucher>déboucher*
- Absences
  - verbes désadjectivaux (*régional > régionaliser*)
  - verbes dénominaux (*hôpital > hospitaliser*)
  - ☞ sémantique moins prédictible
- Semble confirmer l'hypothèse de départ : toutes les relations morphologiques ne sont pas pertinentes pour les systèmes de RI et de QA

Plan

Évaluation : présentation des ressources

Ressources de morphologie dérivationnelle pour le français :

- Prolex [Bouchou and Maurel, 2008], [Tran and Maurel, 2006]
  - dictionnaire multilingue de noms propres, contenant également les adjectifs associés (20 614 relations dérivationnelles)
- VerbAction [Hathout and Tanguy, 2002]
  - ressource contenant les noms d'action dérivés de verbe (9 393 paires nom-verbe)
- Dubois [Dubois and Dubois-Charlier, 1997]
  - description des verbes du français, contenant une information partielle sur la dérivation (33 955 dérivés)

Évaluation de la couverture de ces ressources sur la base du gold standard

Évaluation : présentation des ressources

- Les ressources morphologiques sont consacrées à un phénomène particulier

	Exemple	**Prolex**	**VerbAction**	**Dubois**
N > **A**	commerce > commercial			
N propre > **A**	Afrique > africain	✓		
N > **N**	président > présidente			
N propre > **N**	Arménie > Arméniens	✓		
A > **N**	national > nationalité			
V > **N**	traiter > traitement		✓	✓
Autres	complet > complètement			

- Leur évaluation se fait sur la partie concernée du gold standard

Évaluation : résultats

- Adjectifs dénominaux
  - Prolexbase

Corpus	**Relation morphologique (nbr.)**		**Trouvé dans Prolexbase**	
			nbr.	%
Conique	Gentilé - Adj. Rel	(1)	1	100
	LocOrg - Adj. rel	(45)	43	96
Quæro	LocOrg - Adj. rel.	(8)	8	100
EQueR	LocOrg - Adj. rel.	(1)	1	100
Total		**55**	**53**	**96,36**

- très bonne couverture pour les adjectifs relationnels gentilés

Évaluation : résultats

- Noms déverbaux
  - noms d'événement : VerbAction et Dubois

Corpus (nbr.)	VerbAction nbr.	VerbAction %	Dubois nbr.	Dubois %
Conique (34)	33	97	19	56
Quæro (25)	25	100	9	36
EQueR (22)	22	100	19	86
Total (81)	**80**	**99**	**47**	**58**

    - VerbAction a une très bonne couverture

  - noms d'agent : Dubois
    - 100 % des noms d'agent de Conique, 75 % de ceux de Quæro

Plan

Utilité de la morphologie dérivationnelle : ressources utilisées

Ressources de morphologie dérivationnelle correspondant aux principaux procédés identifiés :

- Prolex [Bouchou and Maurel, 2008], [Tran and Maurel, 2006]
  - dictionnaire multilingue de noms propres, contenant également les adjectifs associés (20 614 relations dérivationnelles)
- VerbAction [Hathout and Tanguy, 2002]
  - ressource contenant les noms d'action dérivés de verbe (9 393 paires nom-verbe)
- VerbAgent
  - ressource développée au LIMSI et en cours de validation, contenant 4 067 paires verbe-nom d'agent dérivé

Utilité de la morphologie dérivationnelle : expériences

- 2 systèmes de QR développés au LIMSI
  - Ritel
  - QAVAL
- Données d'évaluation
  - Quæro (web)
  - QAST (oral)
  - CLEF (écrit journalistique)
- Tests effectués
  - lemmatisation uniquement
  - lemmatisation + chacune des 3 ressources dérivationnelles indépendamment
  - lemmatisation + 3 ressources ensemble

Utilité de la morphologie dérivationnelle : résultats

- Résultats :
  - l'intégration de ressources dérivationnelles n'améliore pas les résultats de façon significative (entre 1% et 2 %)
  - c'est essentiellement VerbAction qui a un effet sur les résultats
  - effets sur le rang des réponses, mais pas nécessairement sur le 1er rang
  - peut améliorer la sélection des documents et des passages, mais pas l'extraction de la réponse
  - l'effet de la morphologie dérivationnelle est + important sur les questions courtes (2 à 3 éléments d'information)

- Expériences à venir
  - comparaison avec le stemming

Conclusion

- Caractérisation des relations morphologiques les plus fréquentes entre question et passage
- Évaluation des ressources existantes en fonction des relations morphologiques identifiées, et développement de ressources manquantes
- Evaluation de l'apport d'une morphologie dérivationnelle ciblée dans des systèmes de QR
  - ☞ les résultats sont mitigés

Conclusion

Merci de votre attention !

Bibliographie

- Bouchou, B. and Maurel, D. (2008).
  Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres.
  Traitement Automatique des Langues, 49(1) :61–88.

- Dubois, J. and Dubois-Charlier, F. (1997).
  Les verbes français.
  Larousse-Bordas.

- Germann, U. (2008).
  Yawat : yet another word alignment tool.
  In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT '08)*, pages 20–23.

- Hathout, N. and Tanguy, L. (2002).
  Webaffix : Discovering Morphological Links on the WWW.
  In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1799–1804, Las Palmas de Gran Canaria, Espagne. ELRA.

- Moreau, F. and Claveau, V. (2006).
  Extension de requêtes par relations morphologiques acquises automatiquement.
  In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, pages 181–192.

- Namer, F. (2009).
  Morphologie, Lexique et TAL : l'analyseur DériF.
  TIC et Sciences cognitives. London : Hermes Sciences Publishing.

- Porter, M. F. (1980).
  An algorithm for suffix stripping.
  Program, 14(3) :130–137.

- Sagot, B. (2010).
  The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French.
  In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

- Tran, M. and Maurel, D. (2006).
  Prolexbase : un dictionnaire relationnel multilingue de noms propres.
  Traitement Automatique des Langues, 47(1) :115–139.