Repérage de créations lexicales sur le Web francophone

Franck SAJOUS & Ludovic TANGUY
ERSS / Université de Toulouse 2

Plan

- Créations lexicales : pour quoi faire ?
- Repérage des créations sur corpus classique
- Repérage sur le Web via un moteur de recherche
  - Méthode hypothético-déductive vs inductive
- Un moteur dédié : Trifouillette
  - Filtrage du bruit
  - Résultats

Quelques trouvailles en vrac

- Formes attestées absentes des listes de référence
- Termes techniques
  - *Aquamarquage*, *hémagglutination*, *immunofixation*
- Concepts récents
  - *Pacser* (*se*) / *pacsage*, *surencadrement*, *intermédiation*
- Langue populaire
  - *Baisage*, *poilade*
- Créations diverses
  - *Pêchable*, *japonisation*, *europhobie*, *googler*

Intérêt du repérage de créations lexicales

- Linguistique traditionnelle
  - Morphologie, lexicologie, terminologie
  - Étude des procédés de création lexicale
  - Veille lexicale, évolution de la langue
- Traitement automatique :
  - Analyse (morpho)-syntaxique, traduction, etc.
  - Constitution / extension de lexiques

Exemples d'études à l'ERSS

- Étude de suffixes particuliers
  - *-able*, *-esque*, *-este*, *-ien*, *-ouill-*, etc.
  - Recensement des formes puis analyse
  - Travaux de M. Plénat, M. Roché, N. Hathout, S. Lignon
- Noms déverbaux d'action
  - Famille de suffixes : *-age*, *-ment*, *-tion*, *-erie*, *-ance*, *-ence*, *-ure*
  - Recensement et analyse : repérage de couples nom/verbe
  - Extension du lexique *Verbaction* utilisé par des analyseurs de corpus (*Syntex*) et en recherche d'information
  - Travaux de N. Hathout, L. Tanguy

Approches sur corpus classiques

- À partir d'une liste de référence
  - Dictionnaire de langue (formes fléchies)
  - Repérage de toute forme non référencée
  - Mathieu et alii, 1998
- Sans liste de référence
  - Repérage des formes rares (hapax)
  - Janicijevic & Walker, 1997
- Par accumulation
  - Repérage des apparitions sur corpus évolutif
  - Renouf et alii (projet APRIL 2004-06)
- Problèmes et résultats communs :
  - Noms propres, fautes d'orthographe, mots collés, etc.
  - La dérivation est le procédé le plus productif

La tentation du Web

- Intérêts :
  - Quantité de données nécessaire pour l'étude de phénomènes rares
  - Représentation de nombreux types de textes, domaines, niveaux de langue
  - Créativité et spontanéité débordante
- Méthodes envisageables :
  - Constitution d'un corpus à partir de sites sélectionnés
  - Utilisation d'un moteur de recherche généraliste
  - Parcours systématique du Web

Utilisation des moteurs de recherche

- Deux approches : hypothético-déductive ou inductive
- Hypothético-déductive :
  - Construction d'un mot-candidat en appliquant des processus de création lexicale
  - *Google* → *googlisation* ?
  - Vérification de son existence sur le Web
- Inductive :
  - Utilisation de patrons
  - `*isation`
  - Filtrage par une liste de référence

Méthode hypothético-déductive

- Adaptée aux créations dérivées à partir de bases connues
  - Exemple : *verbe* → *nom* par suffixation
- Système Walim (F. Namer)
  - *gratiner* → ?*gratinage* ?*gratination* ?*gratinement*
  - Après vérification via un moteur : *gratinage*
- Couverture limitée :
  - Bases connues
  - Procédés morphologiques connus

Méthode inductive

- Utilisation de jokers dans les requêtes
- Un programme : *Webaffix* (Tanguy & Hathout)
  - Désormais inutilisable depuis la disparition d'AltaVista en Avril 2004
- Principes : pour un suffixe donné
  - Construction et lancement de sous-requêtes
  - Analyse et filtrage des résultats
    - Vérification de la langue
    - Correction orthographique, détection des contextes bruités
  - Analyse des créations repérées
    - Calcul de la base et vérification

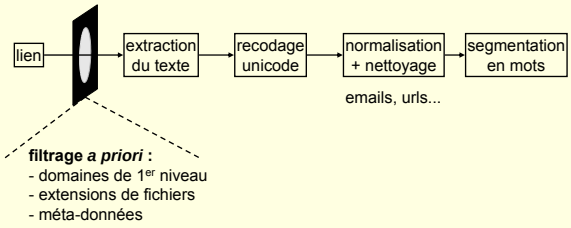
Exemples de campagnes passées

- Adjectifs en *–este*
  - 1 attestation avant 1997 (silvio-pelliqueste), 14 en 2004 (*algueste, bangueste, big-bangueste, blagueste, blogueste, cirqueste, dukeste, gagueste, gangueste, jack-langueste, loqueste, orqueste, swingueste, turqueste*)
- Adjectifs en *–able*
  - 1145 nouveaux adjectifs (1641 dans les dictionnaires généraux)
- Extension du lexique Verbaction
  - Au total, 9400 couples nom/verbe dont 2000 trouvés par Webaffix
- Étude des noms déverbaux concurrents
  - 1150 couples X*age*/X*ment* dont une des formes n'est pas recensée dans les dictionnaires

La recherche continue face à l'adversité : Trifouillette

- Objectif : détection automatique de "formes rares"
  - indépendante des moteurs de recherche
    → mise en œuvre d'un crawler
  - recherche non ciblée
    → l'utilisateur définit ses requêtes *a posteriori*
  - stockage des pages *pertinentes*
    pertinente ≈ contient au moins une forme rare
    rare ≈ nombre d'occurrences rencontrées faible

Filtrage et traitements pré-analyse



Détection de la langue

- Recours à :
  - lexique de référence + mots gramm. FR
  - mots gramm. autres langues
- Critère d'acceptabilité :
  - %mots connus > $Seuil_{MC}$
  - & %mots grammaticaux FR > $Seuil_{GrFR}$
  - & %mots grammaticaux Autres < $Seuil_{GrA}$
  - & %mots différents > $Seuil_{Diff}$
- Au niveau global + en contexte autour des mots inconnus (∉ lexique)

Filtrage du bruit

- Mots collés/Mots séparés
  - *"des technologies d'accès sans fil **tellesque** IEEE802"*
    tellesque ← telles que
  - *"des **technol ogies** d'accès sans fil telles que IEEE802"*
    technol ogies ← technologies
- MAIS
  - *"Bises **pâquestes**."*
    pâquestes ↚ pâques tes
  - *"Syd Barret, spécialiste **es planade**, fut interné en 1968"*
    es planade ↚ esplanade

Filtrage du bruit (2)

- Présence de majuscules :
  - noms propres
  - sigles
  - MAIS
    *"les **ViWistes** vont me haïr, mais je passe aux japonaises, surtout Honda"*
- Suite de lettres "impossibles"
  - "*il travaille au **cnrs***"
  - MAIS : "*l'**xmlisation** récente des corpus*"

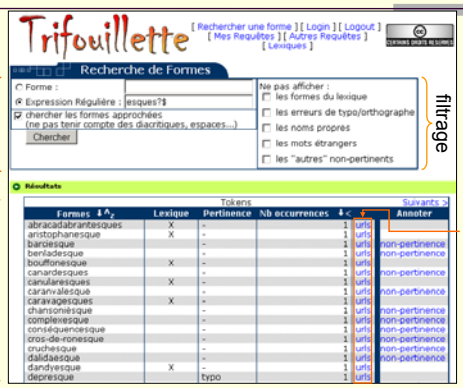
Filtrage du bruit (3)

- Lettres répétées n fois (n > 2)
  - *"Co**uuuu**co**uuuuu** !", "**Aaaaaaaa**rg !"*
  - MAIS : *"la société des xv**iii**èmistes se réunira demain"*
- Lettres parasites
  - *acce**r**ptable ← acceptable*
  - MAIS : *hivernable ↚ hivernale*
- Lettres manquantes
  - *accepable ← acce**p**table*
  - MAIS : *entarter ↚ entart**r**er*

Premiers résultats

- Chaque jour :
  - 100 000 à 700 000 pages, 2 à 35 millions de formes
  - 2000 à 70000 nouvelles entrées, dont :
    - sans filtrage : 80 à 90% de bruit
    - avec filtrage :
      - 19 à 28% de bruit non signalé
      - 7 à 37% de surcorrection
        (parmi les formes réellement pertinentes)
- Trouvailles difficilement "prévisibles" :
  rhône-alpettes, downesque, ViWiste (Volkswagen) warriorisme, repositoires (repository)

Coté utilisateur...



requête

filtrage

accès au contextes

formes trouvées

Perspectives

- Mise à disposition des résultats
- Système d'alerte pour des couples {*requête*, *utilisateur*}
- Étude des contextes d'apparition des créations lexicales
- Collaborations bienvenues :
  - Traitements spécifiques additionnels
  - Autres langues