Récupération de flux de corpus RSS et traitements linguistiques sur le texte

Emmanuel Cartier – LIPN - RCLN

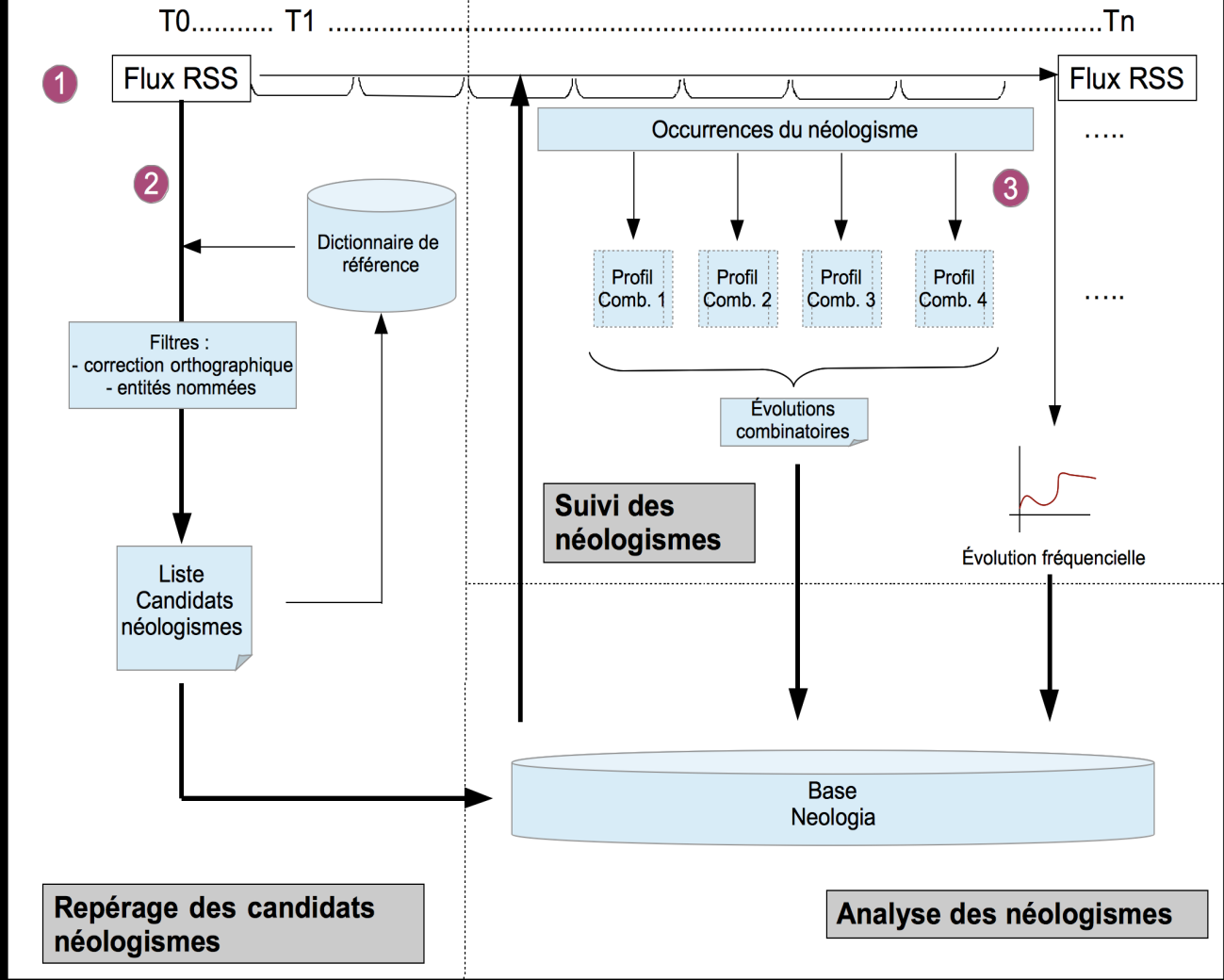
Projet M1 Informatique – UP13 - 2014-2015

Sommaire

- Contexte : projet de veille néologique sur grand corpus
- Objet de la mission
- Contraintes et calendrier

Contexte

Projet de veille néologique sur grand corpus : veille des mots nouveaux dans la langue et de l'évolution du sens des mots



Contexte

- Architecture séquentielle :
- Récupération de corpus sur le web en flux continu (fils RSS)
- Stockage sur un serveur en format txt (ou xml) UTF-8
- Traitements linguistiques 1 : recherche des mots nouveaux (formes nouvelles) – par comparaison avec un lexique de référence
- Traitements linguistiques 2 : analyse statistique des séquences de caractères, de mots, de séquences répétées **en vue d'établir des profils combinatoires des usages des lexies**
- Comparaison des profils combinatoires en diachronie => différences signalent une évolution de sens
-

Objet de la mission

Intervention principale sur le crawler de corpus RSS

Mise en place d'un crawler de fils RSS fonctionnant en continu et effectuant les traitements suivants :

- récupération des flux de quotidiens nationaux et régionaux français

- récupération des fichiers xml RSS et des méta-informations => stockage

- récupération des articles liés (url) : nettoyage, conversion et stockage dans une base de corpus.

Nettoyage fichiers (boiterplate cleaning)

Titre,
Méta-informations,
Texte à l'exclusion de tout le reste de la page

Conversion, stockage

- Conversion : format texte, UTF-8
- Stockage – décision en cours entre:
  - Stockage des fichiers texte sur disque
  - Base de données (relationnelle ou XML)
  - Outil dédié à la gestion et l'interrogation de corpus

Traitements linguistiques 1

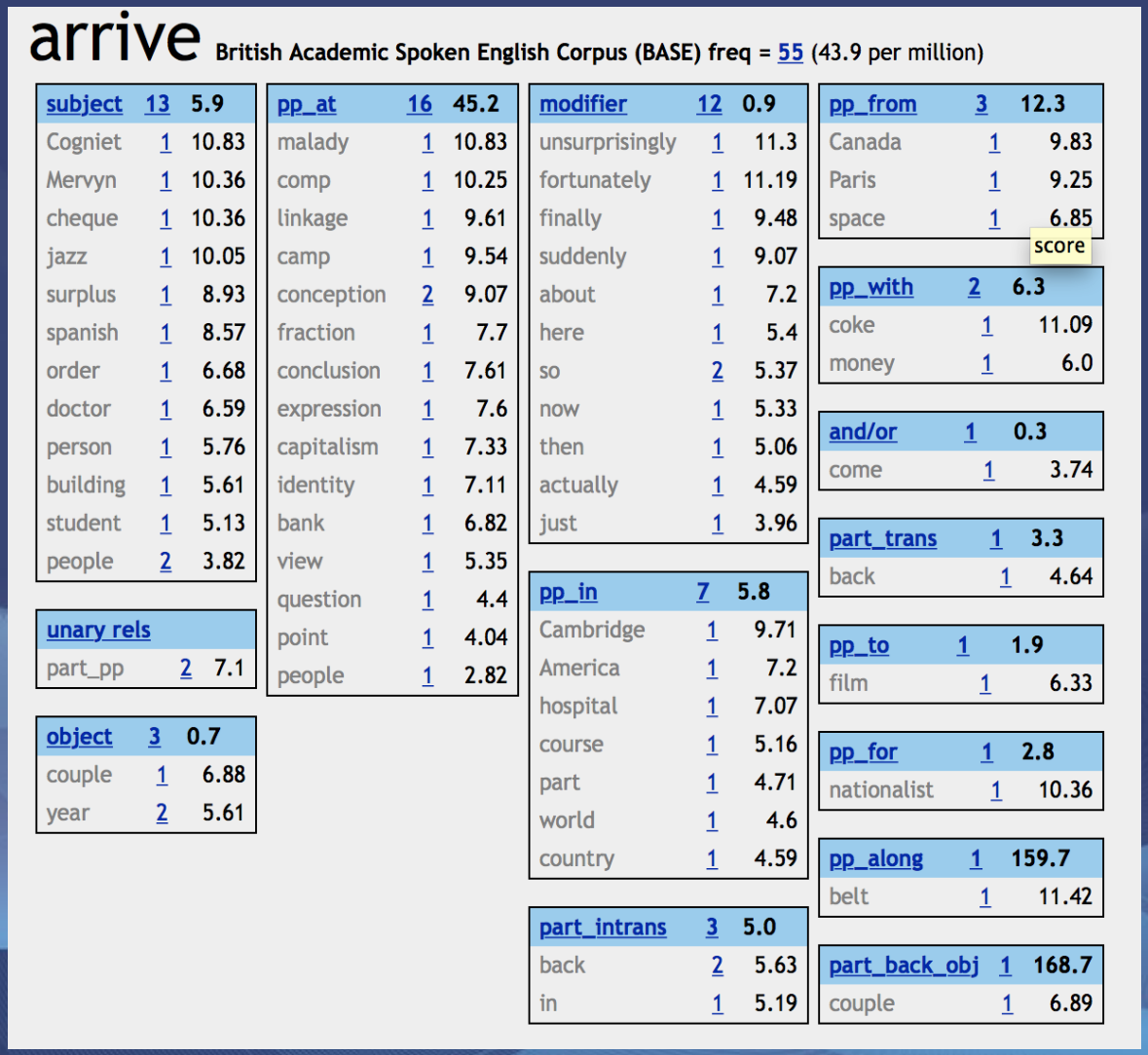
- Objectif : repérage des mots-formes inconnus
- Segmentation en “mots” du texte
- Comparaison avec dictionnaire de référence

  => mots “inconnus”, avec tri des noms propres
- **Attention : tâche bien plus complexe qu'il n'y paraît**
  - Segmentation fautive
  - Erreurs typographiques
  - Noms propres…

Traitements linguistiques 2

- Objectif : repérage des nouveaux emplois par une étude quantitative des unités lexicales visant à établir sur corpus la combinatoire des mots
  - Segmentation en “mots” du texte, normalisation
  - Calcul du vocabulaire et de la fréquence de chaque élément
  - Calcul des n-grams répétés (fenêtre variable)
  - Calcul des n-grams sur informations morpho-syntaxiques
  - Calcul n-grams sur informations mixtes (formes et informations morpho-syntaxiques)

Traitements linguistiques 2



arrive British Academic Spoken English Corpus (BASE) freq = 55 (43.9 per million)

subject	13	5.9
Cogniet	1	10.83
Mervyn	1	10.36
cheque	1	10.36
jazz	1	10.05
surplus	1	8.93
spanish	1	8.57
order	1	6.68
doctor	1	6.59
person	1	5.76
building	1	5.61
student	1	5.13
people	2	3.82

unary rels		
part_pp	2	7.1

object	3	0.7
couple	1	6.88
year	2	5.61

pp_at	16	45.2
malady	1	10.83
comp	1	10.25
linkage	1	9.61
camp	1	9.54
conception	2	9.07
fraction	1	7.7
conclusion	1	7.61
expression	1	7.6
capitalism	1	7.33
identity	1	7.11
bank	1	6.82
view	1	5.35
question	1	4.4
point	1	4.04
people	1	2.82

modifier	12	0.9
unsurprisingly	1	11.3
fortunately	1	11.19
finally	1	9.48
suddenly	1	9.07
about	1	7.2
here	1	5.4
so	2	5.37
now	1	5.33
then	1	5.06
actually	1	4.59
just	1	3.96

pp_in	7	5.8
Cambridge	1	9.71
America	1	7.2
hospital	1	7.07
course	1	5.16
part	1	4.71
world	1	4.6
country	1	4.59

part_intrans	3	5.0
back	2	5.63
in	1	5.19

pp_from	3	12.3
Canada	1	9.83
Paris	1	9.25
space	1	6.85

score

pp_with	2	6.3
coke	1	11.09
money	1	6.0

and/or	1	0.3
come	1	3.74

part_trans	1	3.3
back	1	4.64

pp_to	1	1.9
film	1	6.33

pp_for	1	2.8
nationalist	1	10.36

pp_along	1	159.7
belt	1	11.42

part_back_obj	1	168.7
couple	1	6.89

Contraintes et calendrier

- Langage de programmation : (de préférence) Python, Perl ou Java
- Livraison d'une API + batch (interface graphique facultative)
- accompagnement par un expert pour les problématiques TAL
- Possibilité d'utiliser outils préexistants opensource (analyseur morphosyntaxique notamment)
- Livraison obligatoire en mai-juin 2015!