atilf
ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

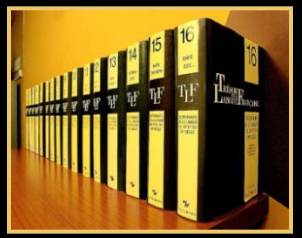
Premiers tests en vue de l'étiquetage d'un corpus oral par apprentissage automatique exclusivement endogène

Christophe Benzitoun & Lolita Bérard

09/09/2011

Christophe.Benzitoun@univ-nancy2.fr, lolita.berard@atilf.fr






www.atilf.fr

cnrs dépasser les frontières

Nancy-Université

OBJECTIFS

- Objectifs principaux
  - Annoter une banque de données orales de + de deux millions de mots (mot, lemme, POS)
  - Elaborer un système d'étiquetage automatique pour l'oral
  - Diffuser les ressources créées
- Objectifs induits
  - Elaborer/diffuser un jeu d'étiquettes et un manuel d'annotation
  - Elaborer/diffuser un corpus de référence annoté semi-manuellement
  - Elaborer/diffuser un lexique à partir du corpus de référence
  - Entraîner un étiqueteur automatique
  - Etiquetage automatique par type de texte

DES CORPUS ETIQUETES : POUR QUOI FAIRE ?

- **Listes de fréquence des formes utilisées**
- **Recherche de phénomènes grammaticaux complexes**
  - **Exemple : étude de la position adjectif-nom en français parlé**
- Reconnaissance automatique de parole
- Etc.

- En fait, traitement initial de bons nombres d'applications

SOLUTIONS ENVISAGEABLES

- Annotation entièrement manuelle écartée
- On s'oriente donc vers les systèmes automatiques
- Mais à l'heure actuelle, outils entraînés sur/pour l'écrit (Cordial, TreeTagger, etc.)
- Solutions généralement envisagées
  - Adapter les corpus aux outils (cacher les disfluences) : Valli & Véronis (1999) ; Dister (2007)
  - Adapter les outils aux corpus : Eshkol et al. (2010)
- Solution retenue
  - Apprentissage automatique à partir d'un corpus de référence oral non modifié pour ne pas dénaturer les données → entraînement avec le module TrainTreetagger

POURQUOI TREETAGGER ?

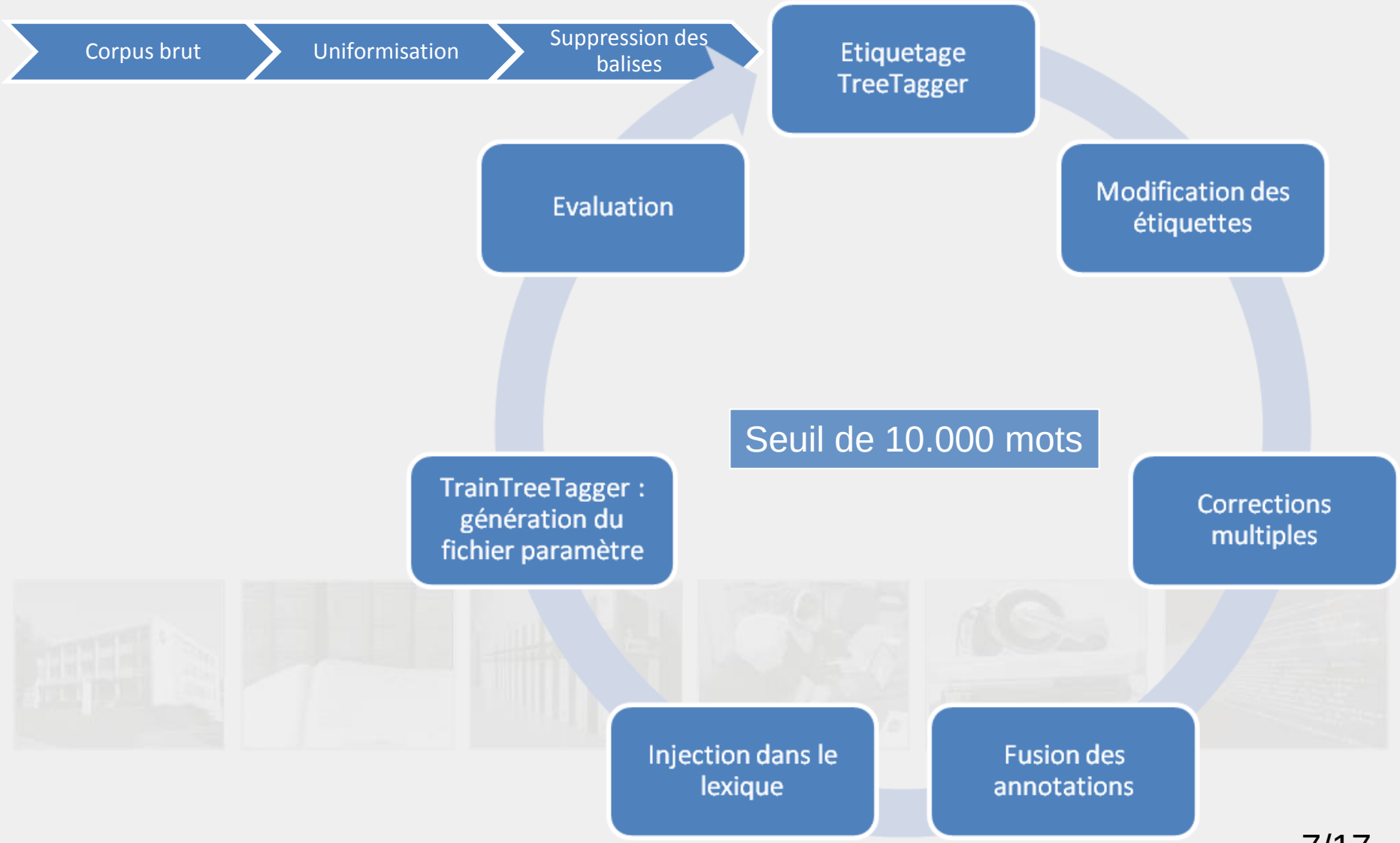
- Jeu d'étiquettes basique et restreint
- Module d'entraînement → possibilité d'obtenir un fichier paramètre spécifique à un type de données
- Étiqueteur probabiliste (basé sur des n-grammes)
- Encodage ISO + UTF8
- Module de segmentation fourni et en partie paramétrable (abréviation + mots multiples)
- Donne de bons résultats à partir de corpus d'entraînement de taille réduite
- Facile à utiliser (!= MElt (Denis & Sagot, 2010))
- Outil libre → résultats diffusables

CORPUS DE TEST

- TCOF (http://www.cnrtl.fr/corpus/tcof/)
  - Partie adulte : 50.000 mots
  - Transcription orthographique
- Conversion automatique de format (Transcriber → Txt) : suppression/ transformation des balises

ETAPES DU TRAVAIL

Travail préliminaire :



RESSOURCES NECESSAIRES POUR L'APPRENTISSAGE

1. Corpus étiqueté
2. Lexique Forme/POS

   création d'un lexique du français parlé à partir du corpus
3. Liste des tokens à ne pas segmenter
4. Liste des étiquettes pour les mots inconnus

TOKENISATION

- Conservation du tokenizer de TT
- Par défaut : segmentation par caractères séparateurs de mots + fichier french-mwls (vide) + fichier abréviations
- Compléter liste des mw et abr.
  - Limiter les ambiguïtés [ex : *de la*]
  - Toujours segmenter si on peut insérer / remplacer un élément [ex : *un peu*]
  - Limiter aux mw effectivement rencontrés → pas d'ambition d'exhaustivité

LEXIQUE

- Extrait à partir du corpus
  - Couverture moins grande que la plupart des lexiques disponibles

  MAIS
  - Étiquettes forcément identiques à celles du corpus – difficile d'adapter un lexique existant
  - Plus facile à maintenir et corriger
  - Écarte les nombreuses formes inusitées, réduit les ambiguïtés (*boulot* : ADJ ; *lourde* : NC in LEFFF)
  - Permet de voir quelles sont les formes réellement utilisées : important pour le FLE et description linguistique en général (possible *vs* attesté)

JEU D'ETIQUETTES UTILISE

- Base
  - TreeTagger : VER:tps_simple, PRP:DET, PRP, PRO, NUM, NOM, NAM, KON, INT, DET, ADV, ADJ
  - Abeillé & Clément (2006)
- Ajouts pour spécificités orales : amorces, multi-transcriptions, locuteur…
- Résulte d'un compromis
  - Faire des distinctions utiles pour des descriptions ultérieures et pour dénombrer les formes/tokens différents
  - Ne pas complexifier la tâche de correction
  - Limiter au maximum les erreurs lors de l'étiquetage automatique
  - Suffisant à partir du moment où la distinction est faite

EXEMPLES DE CHOIX

- Transferts d'étiquettes :
  - DET démonstratifs (ce), indéfinis (chaque, quelque), interrogatifs (quel) plutôt que PRO

    Se comportent comme des déterminants

  - Forme noyau : oui, non, ok, bonjour, bien sûr, merci, d'accord, etc. plutôt que ADV ou INT

    Généralement non attendus comme réponse à la requête ADV

  - Est-ce que : particule interrogative

→ **Correction globalement automatique, n'ajoute pas d'ambiguïtés**

CAS PARTICULIEREMENT DIFFICILES

- Voilà : FNO (et pas INT), VER, PRP
- Lemme de *allez, disons*
- Analyse de *s'en aller*
- Lemme des amorces et mots tronqués
- *Celle-**ci*** vs *cet homme-**ci***
- Seul le *il* est considéré comme CLSi

→ **Annotation manuelle**

SYNTHESE DES ETIQUETTES ACTUELLES

ADJ	adjectif
ADV	adverbe
AUX:tps	auxiliaire
DET:def	déterminant défini
DET:dem	déterminant démonstratif (ce, cette, ces)
DET:ind	déterminant indéfini (chaque, quelque, un, des)
DET:par	déterminant partitif (du)
DET:pos	déterminant possessif (ma, ta, etc.)
DET:pre	pré-déterminant (tout (le), toute (la), toutes (les))
EPE	épenthétique
FNO	forme noyau (oui, non, d'accord, ...)
INT	interjection & particules discursives
KON	conjonction
LOC	locuteur
MLT	multi-transcription (/x,y/, (n'))
NAM	nom propre
NOM	nom
NUM	numéral
PRO:clo	clitique objet
PRO:cls	clitique sujet
PRO:clsi	clitique sujet impersonnel
PRO:dem	pronom démonstratif
PRO:ind	pronom indéfini
PRO:int	pronom interrogatif (comment, où, quand, quoi, etc.)
PRO:rel	pronom relatif
PRO:ton	pronom tonique
PRP	préposition
PRP:det	préposition+déterminant (au, du, aux, des)
PRT:int	particule interrogative (est-ce que)
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impératif
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitif
VER:pper	verbe au participe passé
VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passé simple
VER:subi	verbe au subjonctif imparfait
VER:subp	verbe au subjonctif présent
étiq:TRC	mots tronqués

EXEMPLE CORRECTION MANUELLE (>4000 mots)

- Nb de formes regroupées : 39 [84occ]
- Nb de formes segmentées : 1 [2occ] (puisqu'ils)

- Nb d'étiquettes modifiées [hors seg.] : 740 – 18%
  - corrigeables automatiquement : 497 – 67%
  - à corriger manuellement : 243 – 33%

- Nb de lemmes modifiés : 43 occ – 1%

- Trentaine de formes ambiguës

PREMIERS RESULTATS

- **Corpus entraînement : 10857 tokens**
- **Lexique : 1703 formes**
- **Corpus test interne (ApprendreAuLycee)**
  - Taille : 2979 tokens
  - 34 erreurs d'étiquetage : *avoir, autre, certains, des, être, il, même, où, première, que, quoi, tout, vous*
  - Taux d'erreurs : 1,14 %
- **Corpus test externe (Assemblee_DIM_08)**
  - Taille : 895 tokens
  - 124 occurrences inconnues mais bonne étiquette
  - Taux d'erreurs mots connus : 5,4 % (38/696)
  - Taux d'erreurs mots inconnus : 28 % (55/196)
  - Taux d'erreurs global : 10,4 %

PERSPECTIVES

- Eventuellement, rationaliser les étiquettes tel que le proposent Eshkol et al. (2010)
  - Cela **PRO:dem** | PRO:ton => PRO:ton:dem
- Continuer le travail jusqu'au 50.000 occurrences dans un premier temps
- Systématiser le travail de double-annotation
- Faire des évaluations inter-annotateurs
- En fonction des résultats, lancer l'étiqueteur sur la totalité de la base
- Etendre la méthodologie à d'autres types de texte (littérature et journaux)