La base PFC : bilan et évolutions

Giulia Barreca
Julie Peuvergne
Atanas Tchobanov

7 décembre 2013

PLAN

1. Bilan
2. Anonymisation du corpus PFC
3. Annotation morphosyntaxique du corpus PFC
4. Nouveau moteur de recherche (démonstration)

1.Bilan

- 36 points d'enquête en ligne
- En vérification : Maurice, Martigny (Valais – Suisse), Bejaia (Algérie), Béarn, Amiens, Bar-sur-Aube (France), Saguenay, Montréal, Trois-Rivières (Canada)
- En cours : Martinique, Yaoundé (Cameroun), Genève (Suisse), Windsor, Tracadie, Hearst, Québec, La Pocatière, Grande-Rivière , Chelsea, Wickham, Saint-Tite (Canada)
- Prochains points d'enquête : Saint Etienne, Corse…
- Toulouse (LVTI) : 33 en cours, 12 en vérification

2. Anonymisation du corpus PFC

2.1. Les données concernées par l'anonymisation

- Les données à caractère personnel explicites
- directement nominatives :
  - Noms de famille
  - Prénom
  - Sigles

- indirectement nominatives :
  - Lieu et date de naissance
  - Profession, statut, titres…
  - Lieux (toponymes, institutions, services….)
  - Caractéristiques de la personne (physiques, culturelles, médicales…) uniques ou rares dans son milieu identifié.

- Les données à caractère personnel implicite :

  Les données qui permettent indirectement d'identifier le locuteur.

- Les données sensibles :

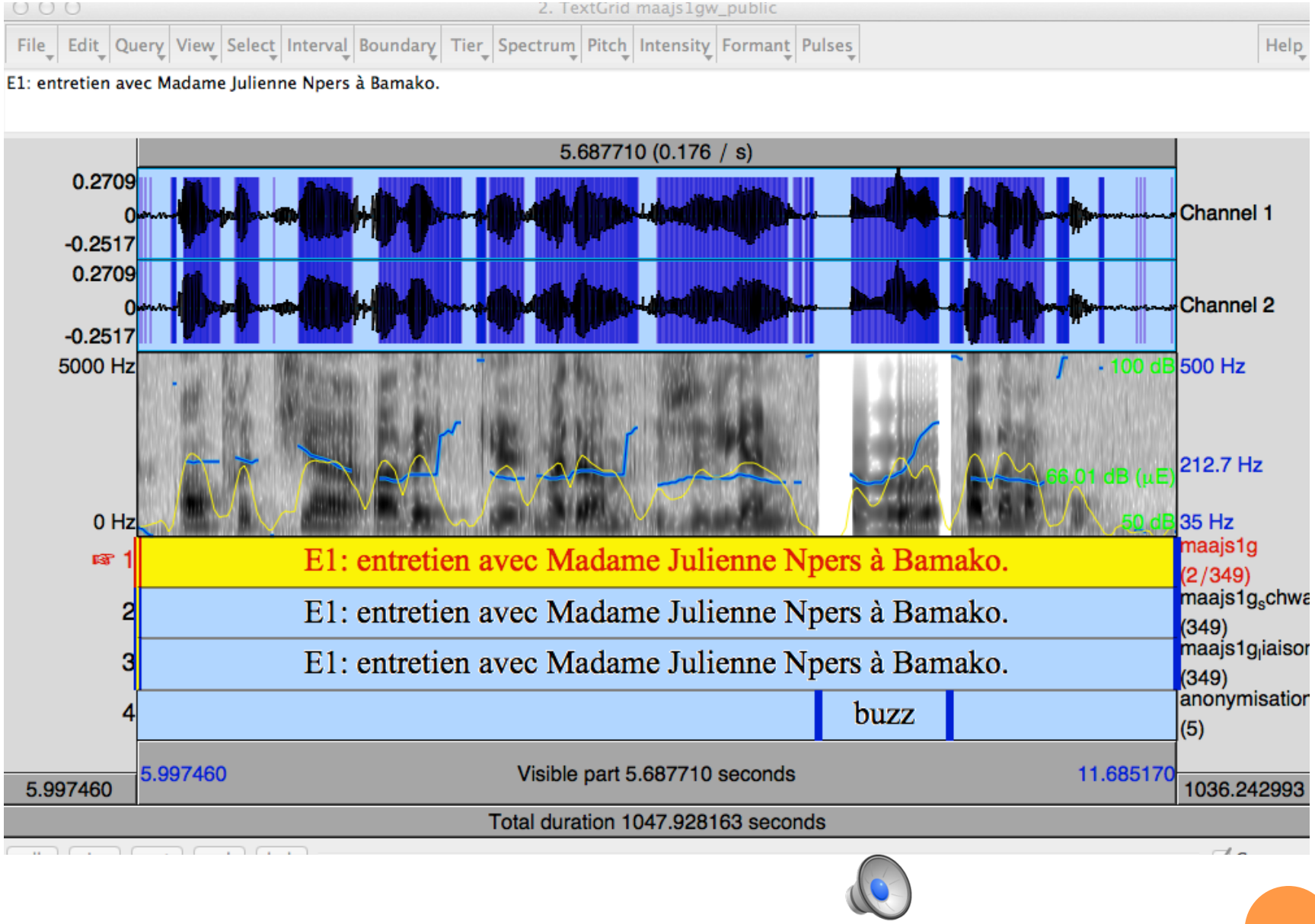
  Les données qui font apparaître les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à leur santé.

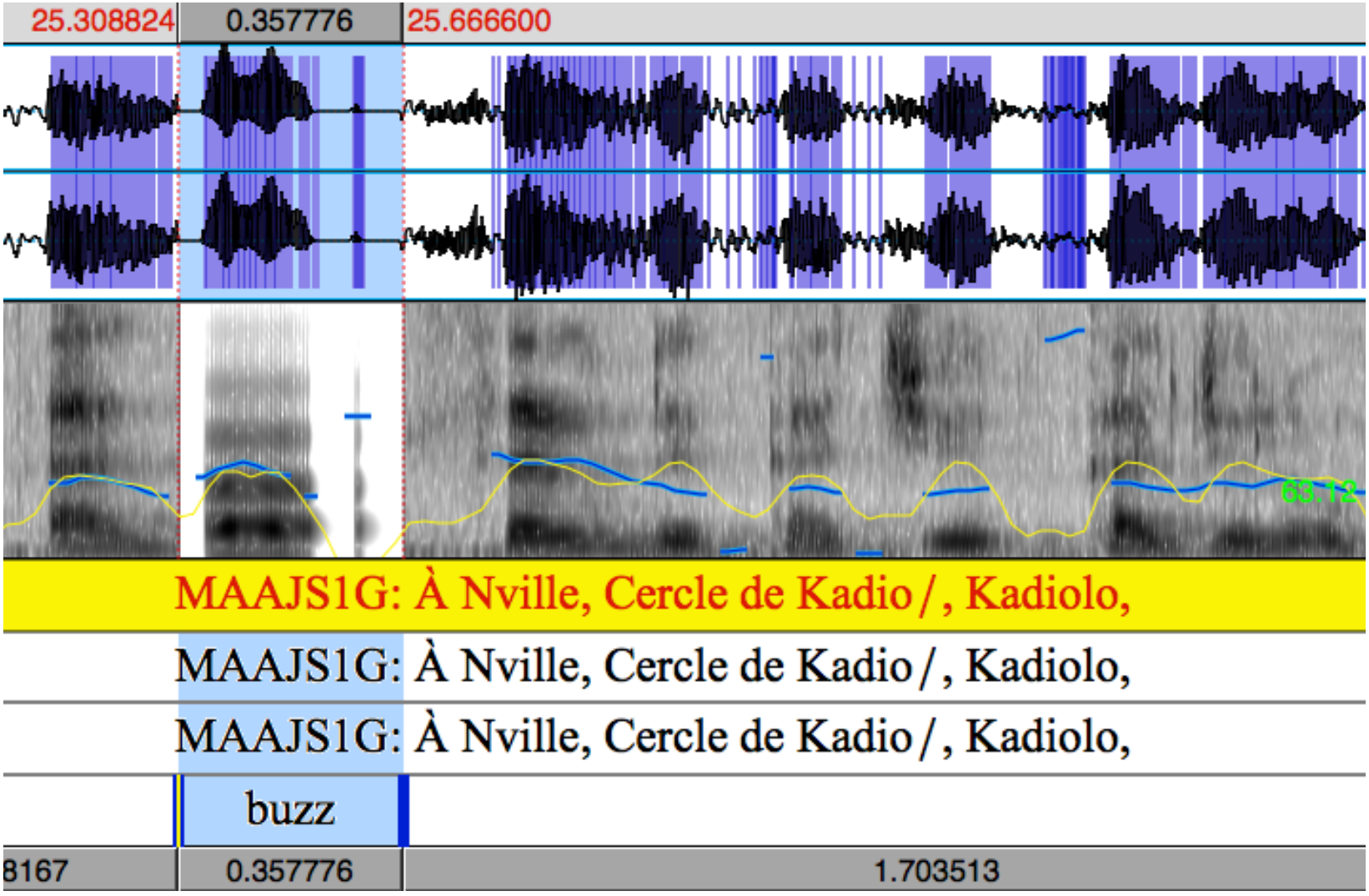
2.2. Anonymisation des textgrids

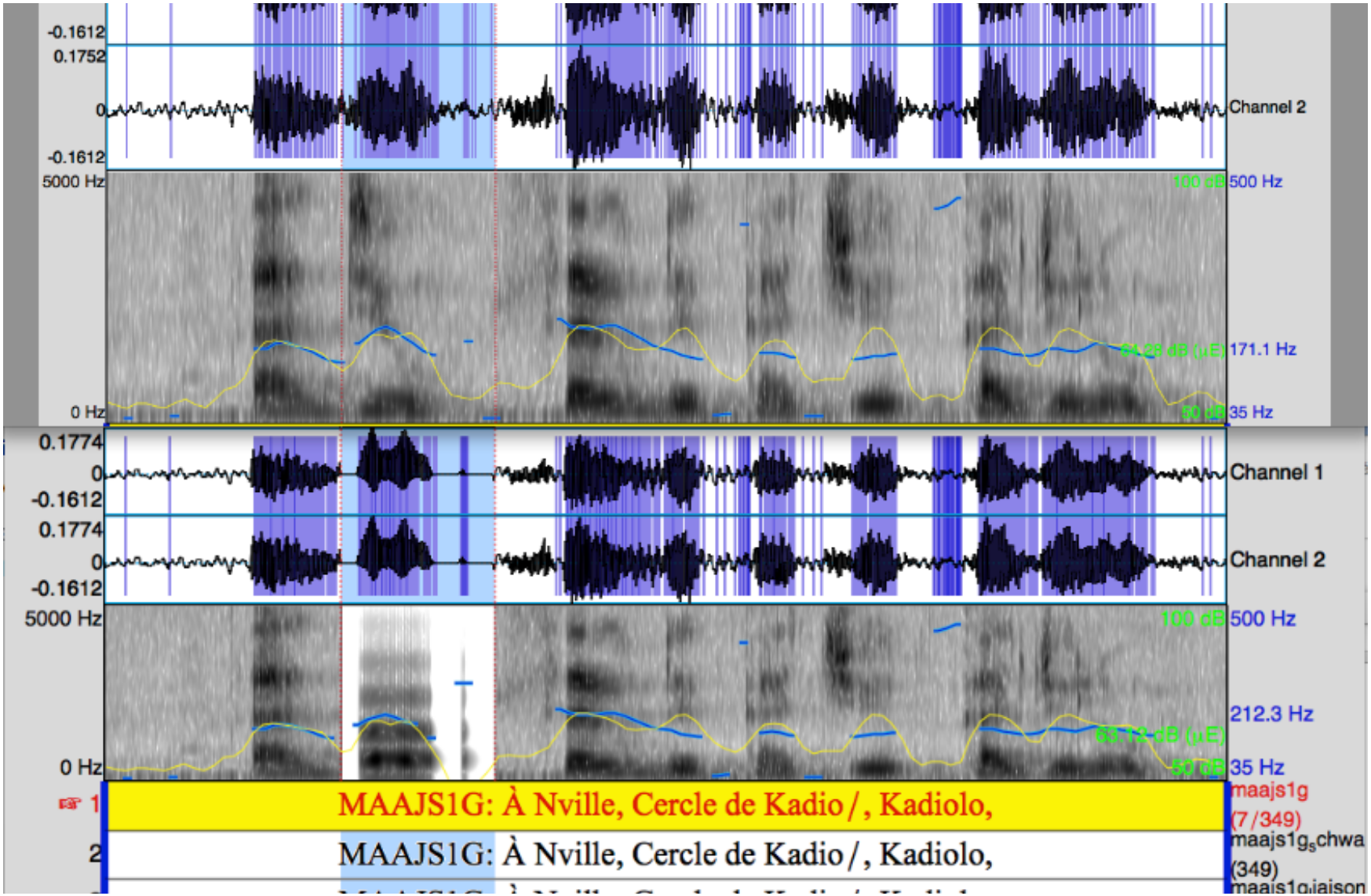
- Convention : Npers, Nville….
- Anonymisation systématique :
  - Noms de famille
  - Lieu de naissance
  - Année de naissance
- Anonymisation au cas par cas :
  - Prénom
  - Profession
  - ….

2.3. Anonymisation du son

- Script pour Praat “Anonymise Sound Files” de Daniel Hirst (Hirst 2010)
- ‘Brouillage’ du son
- Préservation partielle des caractéristiques prosodiques
- Problème en cas de chevauchement
- Réécriture des fichiers


2. TextGrid maajs1gw_public
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
E1: entretien avec Madame Julienne Npers à Bamako.
5.687710 (0.176 / s)
0.2709
0
-0.2517
Channel 1
Channel 2
5000 Hz
0 Hz
100 dB
500 Hz
212.7 Hz
66.01 dB (µE)
50 dB
35 Hz
1
2
3
4
E1: entretien avec Madame Julienne Npers à Bamako.
buzz
maajs1g
(2/349)
maajs1g_schwa
(349)
maajs1g_liaison
(349)
anonymisation
(5)
5.997460
Visible part 5.687710 seconds
11.685170
1036.242993
Total duration 1047.928163 seconds


25.308824
0.357776
25.666600
63.12
MAAJS1G: À Nville, Cercle de Kadio / , Kadiolo,
MAAJS1G: À Nville, Cercle de Kadio / , Kadiolo,
MAAJS1G: À Nville, Cercle de Kadio / , Kadiolo,
buzz
8167
0.357776
1.703513


-0.1612
0.1752
0
-0.1612
Channel 2
5000 Hz
100 dB
500 Hz
171.1 Hz
0 Hz
35 Hz
0.1774
0
-0.1612
Channel 1
0.1774
0
-0.1612
Channel 2
5000 Hz
100 dB
500 Hz
212.3 Hz
0 Hz
35 Hz
maajs1g
(7/349)
maajs1g_schwa
(349)
maajs1g_liaison
MAAJS1G: À Nville, Cercle de Kadio / , Kadiolo,
MAAJS1G: À Nville, Cercle de Kadio / , Kadiolo,

2.4. VISIBILITÉ DES DONNÉES NATIVES ET ANONYMISÉES

- Les données anonymisées sont accessibles aux conditions habituelles
- Seules les portions transcrites des fichiers sonores sont disponibles
- Les données natives sont conservées mais non accessibles sauf convention spéciale signée avec la direction de PFC

3. Annotation morphosyntaxique du corpus PFC

3.1. Objectifs

Rendre **PFC** un **corpus de français oral :**

- Annoté en morphosyntaxe (POS et lemmes)
- Aligné tokens-POS-son
- Librement accessible en ligne

  Très peu de corpus de français parlé annoté et open source :
  - corpus de français parlé *PERCEO* (Benzitoun et al., 2012)
  - *Rhapsodie* : corpus de référence en français parlé (Lacheret, Kahane, Pietrandrea 2014)
- Exploitable pour tout traitement syntaxique et lexical

3.2.MÉTHODOLOGIE

1. Fractionnement des intervalles du Textgrid:

   Fractionnement de longs intervalles du Textgrid à l'aide du logiciel EasyAlign (Goldman 2011) assure la conservation de l'alignement token-POS-son.

2. Tokenization:
   - Identification des mots composés grammaticaux et lexicaux à l'aide du guide morpho-syntaxique (Abeillé et Clément 2006).
   - Tokenization à l'aide du tokanizer spécifique au français SxPipe (Boullier & Sagot 2008) basé sur Lefff (Lexique de Formes Fléchies du Français). (Sagot 2010).

3. « Dépliage » de la transcription Textgrids.

Traitement des phénomènes perturbateurs :

- Chevauchements :
  JP: vers le trois peu/ ***<E: Ah oui?>*** *peut-être;*

- Commentaires entre parenthèses :
  (rires);

- Sigles des locuteurs :
  JP*: euh, nous aussi*;

- Coordonnées de synchronisation :
  Intervals: size = 183 intervals [1]
  xmin = 0 xmax = 3.193555237451039

4. Annotation morphosyntaxique (étiquettes POS et lemme)

4.1. **Étiquetage (POS)** semi-automatique en trois étapes :

- Pré annotation automatique : quatre étiqueteurs différents :
  - MElt (Denis & Sagot 2010)
  - MElt (réentrainé sur le corpus oral PERCEO) (Benzitoun, Fort, Sagot 2011)
  - Stanford Postagger (entrainé sur le French Treebank) (Toutanova & Manning. 2000)
  - Treetagger (réentrainé sur le corpus oral PERCEO) (Benzitoun, Fort, Sagot 2011
- Comparaison semi-automatique des sorties (Loftsson et al. 2011)
- Correction manuelle des erreurs restantes

4.2. **Lemmatisation** semi-automatique en trois étapes:

- Emploi de deux techniques:
  - Lemmatisation en contexte à l'aide de Treetagger (Schmid 1997)
  - Lemmatisation hors contexte sur la base de Lefff (Sagot 2010)
- Comparaison automatique des sorties
- Correction manuelle des erreurs

3.3.PRÉTEST ÉTIQUETAGE (POS) COMPARAISON PRÉCISION ÉTIQUETEURS

Échantillon testé: 2640 tokens.

Treetagger (Schimd 1997)	**Tretagger (PERCEO) (Benzitoun et al. 2012)**	**MElt (Denis & Sagot 2010)**	**MElt (PERCEO) (Benzitoun et al. 2012)**	**Stanford (Toutanova & Manning 2000)**
91%	93%	**94%**	**98%**	93%

3.4. Apport de la comparaison

MElt

(Denis & Sagot 2010)

Interjections

(Beh, oh)

75% erreurs

MElt

(PERCEO)

(Benzitoun et al. 2012)

100% correction erreurs interjections

+

45 nouvelles erreurs

3.5. OUTPUT

- **Corpus PFC :**
  - Annoté
  - Lemmatisé
  - Aligné
  - Accessible en ligne

3.6. Exploitations possibles du corpus PFC annoté

- **Corpus d'apprentissage spécifique** au français parlé

- **Corpus exploitable** pour des recherches syntaxiques, lexicales, sociolinguistiques, TAL…

- Création d'un **concordancier lemmatisé** pour l'enseignement/apprentissage du lexique et de la syntaxe du FLE (projet PFC-EF)

4. Nouvelle interface

- http://www.projet-pfc.net/liaisonsn.html

RÉFÉRENCES BIBLIOGRAPHIQUES

- Abeillé, A. & Clément, L. (2006). *Annotation morpho-syntaxique. Les mots simples - Les mots composés Corpus Le Monde*.
- Belião, J. (2011). *Formalisation, implémentation et exploitation d'une hiérarchie objet intono-syntaxique. Étude sur un treebank de français oral spontané*. Mémoire de Master
- Benzitoun, C., Fort, K. & Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de TALN 2011*, Grenoble, France.
- Christodoulides, G. & Grosman, I., (2012). DisMo:A morphosyntactic annotator for spoken French. *Journée d'étude CONSCILA (ENS Paris) Annotation syntaxique de corpus oraux*
- Denis, P. & Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong-Kong, Chine.
- Goldman, J-Ph (2011). EasyAlig : an automatic phonetic alignment tool under Praat. In *Proceedings of InterSpeech*. Firenze, Italy.
- Lacheret A., Kahane S., Pietrandrea P. (2014 ed.), *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, col *Studies in Corpus Linguistics,* Amsterdam:Benjamins.

- Hirst, D. (2010) Anonymisation de fichiers sonores. Outil. Laboratoire parole et langage –UMR 7309 (LPL, Aix-en-Provence FR). Création 2010-07-28. Speech and Language Data Repository. Identifiant hdl:11041/sldr000526 - Archived ark:/87895/1.4-126693.
- Sagot, B. (2010). The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC 2010*, La Valette, Malte.
- Sagot, B. & Bouillier, P. (2008). SxPIPE 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.),* 49(2):155–188.
- Schmid, H. (1997). Probabilistic Part-of-Speech tagging using decision trees. *In* D.Jones & H.Somers (éd.), *NewMethods in Language Processing*,154-164. London: UCL Press.
- Toutanova, K & Manning, C., D. (2000)._Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger.In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 63-70.

MERCI!