# Outils Linguistiques

**Jean-Yves Antoine**
LI - Université Rabelais de Tours
Jean-Yves.Antoine@univ-tours.fr

Mastere L&R — U. Tours



# Outils Linguistiques

**TRANSCRIPTION DE CORPUS ORAUX**

Mastere L&R — U. Tours



## Corpus oraux et transcription

- **Utilité des corpus transcrits**
  - Etudes linguistiques sur la langue parlée : évitent un retour au signal audio pour des études lexicales, syntaxiques, sémantiques ou pragmatiques.
  - TALP : apprentissage de modèles de langage

- **Deux types de transcription**
  - **Phonétique ou prosodique** : prononciation phonétique exacte et/ou le rythme adopté (longueur des pauses etc…)
  - **Orthographique** : transcription moins fine rendant compte sous forme écrite de ce qui a été prononcé: restitution **littérale** ce qui a été prononcé



## Corpus oraux et transcription

- **Le dilemme de la transcription orthographique**
  - Transcription littérale sans interprétation : comment assurer ce passage objectif, sans variabilité, de l'oral à sa transcription écrite ?
  - Séparation phonétique (prononciation) et lange (syntaxe de la langue parlée)

**Exemples**

*douskipudonctan* [Queneau, Zazie dans le métro]
transcription des toponymes dans la langue du colonisateur
fautes d'accord : * *des grands hommes* ou *des grands-t-hommes* ?
ambiguïté : * *il marche dans la rue* ou *ils marchent dans la rue* ?

**Règle d'or de transcription orthographique**

*Transcrire littéralement ce qu'on entend et non pas ce qu'on croit entendre*

- *ne jamais corriger ce qu'on entend pour le rendre plus « acceptable »*
- *ne pas inventer de nouvelles formes écrites pour simuler la prononciation orale*



## Corpus oraux et normalisation

- **Réutilisabilité des ressources existantes**
  - Transcription: coût très important (20 à 40h d'écoute pour 1h de signal)
  - Normaliser pour optimiser l'utilisation de ces ressources
- **Normalisation linguistique : conventions de transcription**
  - Objectiver la transcription
  - Eviter les biais méthodologiques : une réalité linguistique doit être partout transcrite de la même manière.
- **Normalisation du codage**
  - Normalisation technologique : permet la réutilisation des mêmes outils informatique pour la gestion des corpus
  - Besoins et contenus différents d'un corpus à un autre
  - Codage sous format structuré : langage de balisage XML
  - TEI (*Text Encoding Initiative*)



## Conventions de transcription

- **Objectifs**
  - Définir des règles systématiques de comportement face à des observations problématiques : limiter la variabilité entre transcripteurs et entre corpus
- **Limites**
  - Limite de l'objectivité : toute convention répond à des a priori théoriques.
  - Toujours joindre les conventions à un corpus afin d'expliquer les choix de transcription qui ont été adoptés
  - Recours à l'audio reste toujours possible
- **Exemples de conventions (français)**
  - GARS-DELIC [Blanche-Benveniste, 1991]
  - Parole Publique [Antoine, 2002]
  - Transcriber www.etca.fr/CTA/gip/Projets/Transcriber/ [Barras *et al*, 1998]

## Transcription : ne pas céder à l'écrit

- **Phrase et énoncé**
  - la notion de phrase n'a pas de sens à l'oral. On distingue simplement des énoncés qui correspondent à une prise de parole ininterrompue du locuteur

    *« et donc il y avait également sans doute et qui va avec tout ce que je viens de dire avant il y avait aussi un homme mythique complètement mythique et c'est la / première et là en plus c'est intéressant parce que c'est la première image que j'ai eue de la Tchécoslovaquie lorsque je suis arrivé à la frontière »*
  - La transcription ne comporte donc pas de points, virgules ou points virgules. Les énoncés ne débutent pas par une majuscule, ceci pour faciliter l'analyse syntaxique ou lexicale des corpus (majuscule réservée aux noms propres)
- **Segmentation en mots**

  pour permette l'analyse lexicale, il est préférable de n'utiliser le tiret que pour les unités polylexicales insécables, indépendamment des normes de l'écrit

  *peux tu* et non pas *peux-tu*
  *porte-feuille* et non pas *porte feuille*



## Transcription : éviter la caricature

- **Ne pas céder à une vision caricaturale de l'oral**

  **Exemple : contractions phonétiques**

  Eviter une transcription orthographique qui colle trop à la prononciation

  *il y a* et non pas *y'a*
  *il part* et non pas *i'part*
  *je vais* et non pas *j'vais*

  MAIS…

  Rendre compte de toute élision complète d'une unité lexicale :

  *il pleut pas* et non pas *il ne pleut pas* si élision du discordanciel *ne*



## Conventions de transcription : exemples

**Majuscules / minuscules**

- Majuscules réservées aux noms propres, épellations, sigles
- Quelles règles pour les associations nom communs / noms propres ?

  **Exemples**

| | | |
|---|---|---|
| *mairie de Paris* | ou ? | ~~*Mairie de Paris*~~ |
| *amicale bouliste de Lamastre* | ou ? | ~~*Amicale Bouliste de Lamastre*~~ |
| ~~*crédit lyonnais*~~ | ou ? | *Crédit Lyonnais* |
| ~~*parti socialiste*~~ | ou ? | *Parti Socialiste* |
| *Toulouse football club* | ou ? | ~~*Toulouse Football Club*~~ |
| *médecins du monde* | ou ? | ~~*Médecins du Monde*~~ |
| *l'île de la tentation* | ou ? | ~~*L'île de la Tentation*~~ |

**Choix de convention**

- Majuscules pour marques identifiées ou dénominations correspond à un acronyme
- Guillemets + minuscules si minuscules seules trompeuses

  **Exemple** **«** *l'île de la tentation »*



## Conventions de transcription : exemples

- **Epellation**

  Rendre compte de la prononciation : majuscules séparées par des blancs

  **Exemple** *My name's Cholomondely Featherstonaugh C H O L O M O N D E L Y ….*
- **Acronymes**

  Si ambiguïté, distinction entre épellation ou non (voire mixte)

| | | | | |
|---|---|---|---|---|
| ~~*S N C F*~~ | ou ? | *SNCF* | | |
| *SMIC* | ou ? | *Smic* | ou ? | ~~*S M I C*~~ |
| *O N U* | ou ? | *ONU* | | |
| ~~*CDROM*~~…. | ou ? | ~~*CD Rom*~~… | ou ? | ~~*C D Rom*~~ cédérom |

- **Chiffres / Nombres / Dates**

  Pour les nombres (i.e.>10) privilégier l'écriture en chiffre

  *1200* se transcrit *mille deux cents* ou *douze cents*
  *73,5* se transcrit *soixante treize virgule cinq* ou *septante trois virgule cinq*

  Certaines conventions limitent la règle aux nombres à prononciation ambiguë

  Exemple (Transcriber) 60 mille 200



## Conventions de transcription : exemples

- **Troncatures**

  Marquer la troncature tout en essayant de rendre compte de l'intention du locuteur

  **Exemple (Transcriber)**

  *On en recause après-de(main) euh lundi* aucune ambiguïté
  *Alors je vous()* ambiguïté *vous, voue, voulais, voudrais*

  Eviter toute interprétation non certaine dans la complétion
- **Prononciations inattendues, lapsus**
  - Ne pas représenter orthographiquement la liaison
  - Transcription phonétique (ou glose de prononciation) à part

  **Exemple (Transcriber)** *moi aussi* [pron=moi-z-aussi]
- **Ambiguïté**

  Transcriptions alternatives et non pas interprétation personnelle…

  **Exemple (DELIC)** *ils/il marchent/marche dans la rue*



## Transcription et dialogue : tours de parole

Du fait du chevauchement des interlocuteurs, il est difficile de donner une définition claire de la notion d'énoncé,appelée dans ce cas **tour de parole** (*speech turn*). Deux solutions envisageables :

- centrer la définition sur chaque locuteur (période limitée par une prise et une fin de parole) et coder les chevauchements temporels

  **Exemple** (alignement de type GARS/DELIC)

  U1 : *on a maintenant un camping-car à la place de notre caravane*
  S1 : *ah ouais*
  U2 : *ouais c'est bien plus pratique*
- centrer la définition sur le dialogue (tour de parole = période d'interlocution)

  **Exemple** (conventions Transcriber, PAROLE PUBLIQUE)

  T1 : <*U = on a maintenant un camping-car*>
  T2 : <*U = à la place de*>
  <*S = ah ouais*>
  T3 : <*U = notre caravane ouais c'est bien plus pratique*>

## Normalisation du codage : SGML et XML

**SGML** (*Standard Generalized Markup Language*) ISO-8879

- Langage de description pour les documents structurés
- Séparation structure / contenu / mise en forme du document
- **Langage à balises** : **structure** du document décrite par un ensemble de balises et leurs imbrications possibles
  - balise ouvrante (<B>) et balise fermante (</B>) délimitent une partie du document, dans laquelle peuvent s'imbriquer d'autres balises
  - pas de balises ni de type de structure prédéfinie : rôle d'une **DTD** (*Document Type Definition*)
- **Contenu** : format libre entre les balises

  Exemple <livre> <chapitre> ceci est le chapitre 1 </chapitre>
  <chapitre> ceci est le chapitre 2 </chapitre>
  </livre>
- **Mise en forme** : non décrite, laissée à d'autre outils



## Normalisation du codage : SGML et XML

**Normalisation des corpus et langages à balises**

La séparation contenu / structure / mise en forme facilite l'interopérabilité entre les différentes ressources linguistiques :

- parseurs permettant la traduction d'une DTD à une autre
- une norme d'encodage peut se limiter à la définition d'une DTD

**TEI (*Text Encoding Initiative*)**

- Norme d'encodage initialement développée en SGML
- Historiquement : textes littéraires
- Actuellement, pour tout document langagier (dont parole transcrite)

**Transcriber**

- Outil de transcription de corpus oraux
- DTD spécifique



## Normalisation du codage : SGML et XML

**SGML : bilan**

- **Souplesse** de représentation : DTD
- **Puissance d'expression** : syntaxe SGML et DTD complexes
- **Complexité** : analyse (parsage) coûteuse en temps de calcul

**HTML** (*HyperText Markup Language*)

- Langage de balises simplifié pour un usage spécifique : WWW
- Balises et structure prédéfinie : norme HTML peu évolutive
- Mélange structure / présentation dans le système de balisage

**XML** (*eXtensible Markup Language*) www.w3c.org/XML/

- SGML simplifié (syntaxe plus rigoureuse et DTDs plus simples)
- Souplesse, adaptabilité et évolutivité conservées : DTD
- Présentation: feuilles de style CSS (*Cascading Style Sheet*)
  XSL (*eXtensible StyleSheet Language*)



## XML et transcription : exemple

```
<Trans version="1" version_date="981211" audio_filename="fi0428" scribe="YM" xml:lang="fr">
<Speakers>      <Speaker id="sp1" name="Simon Tivolle" type="male"/>
                <Speaker id="sp2" name="Patricia Martin" type="female"/>
</Speakers>
<Episode program="France Inter" air_date="980428:0700">
<Section type="filler" startTime="0.000" endTime="4.736">
<Turn speaker="sp1 sp2" startTime="0.000" endTime="0.387"> <Sync time="0.000"/>
        <Who nb="1"/> ouais
        <Who nb="2"/> sûr ?
</Turn>
<Turn speaker="sp1" startTime="0.387" endTime="4.736"> <Sync time="0.387"/>
                ah bon ?
        <Event desc="rire"/><Sync time="1.008"/>
                non blague blague de Patricia France-Inter il est 7 heures
</Turn>
</Section>
<Section type="nontrans" startTime="4.736" endTime="9.609">
<Turn startTime="4.736" endTime="9.609"> <Sync time="4.736"/>
        <Background time="4.736" type="music" level="high"/>
</Turn>
</Section> </Episode> </Trans>
```

DTD Transcriber



## Transcription et déontologie

**Identité des locuteurs**

- Anonymer toute information permettant d'identifier une personne
- Droit à l'image (corpus multimédias)
- Déclaration CNIL si nécessaire

**Consentement éclairé**

- Expliquer les objectifs de l'enregistrement
- Faire signer une acceptation d'enregistrement après explications
- Problème : parole spontanée ?



## Bibliographie

- **Articles cités ou ouvrages de référence**


  - **Antoine J.-Y.** (2002) Corpus OTG : présentation générale. Rapport technique VALORIA-CORAIL 2002-2, Université de Bretagne Sud, Vannes, France. Disponible à l'URL suivante: http://www.sir.blois.univ-tours.fr/~antoine/parole_publique/OTG/Pres_OTG.pdf
  - **Barras C. et al. (1998)** Transcriber, a free tool for segmenting, labeling and transcribing speech, Actes *Language Resources and Evaluation Conference*, *LREC'1998*, Grenade, Espagne, 1373-1376
  - **Blanche-Benveniste C., Bilger M., Rouget C., van den Eynde K.** (1991) Le français parlé : études grammaticales. CNRS Editions, Paris, France.