Frédérique Bénard

Présentation du Projet Innovant
de l’Ecole Doctorale 268¹ Langage et Langues :

Septembre 2005

¹ http://ed268.univ-paris3.fr/

Responsable du projet : Serge Fleury
Remerciements

21 Septembre 2005

Je tiens à remercier l’ILPGA pour y avoir suivi des cours vraiment passionnants depuis deux ans et rencontré tant de personnes auxquelles je suis très attachée.

Je remercie les membres du Projet Innovant de l’Ecole Doctorale 268 Langage et Langues pour m’avoir intégré au projet et fait passer une très bonne année de travail collectif.

Je voudrais remercier Michel Jacobson pour m’avoir fait passer une très bonne journée à la BNF, et pour le travail au LACITO.

Je remercie également tous mes camarades de maîtrise et doctorants pour la superbe année que j’ai passé avec eux.

Merci à tous,

Frédérique Bénard
Sommaire

PRESENTATION DU PROJET INNOVANT............................................................................. 1
DE L’ECOLE DOCTORALE 268 LANGAGE ET LANGUES : ........................................... 1
SEPTEMBRE 2005............................................................................................................ 1

Remerciements ................................................................................................................ 2

Introduction .................................................................................................................... 6

I Outils pour la normalisation linguistique..................................................................... 9
1. Norme pour la représentation des caractères : UNICODE ....................................... 9
2. Normes génériques d’annotation des textes............................................................... 9
   a) SGML (Standard Generalized Markup Language)............................................... 9
   b) HTML (HyperText Markup Language)................................................................. 9
   c) XML (eXtensible Markup Language)................................................................. 10
3. Normes d’annotation par métadonnées : du générique au linguistique .............. 14
   a) Métadonnées ....................................................................................................... 14
   b) Dublin Core.......................................................................................................... 15
   c) TEI (Text Encoding Initiative)........................................................................... 17
   d) OAI (Open Archives Initiative)......................................................................... 17
   e) OLAC (Open Language Archive Community)............................................... 18

II Présentations de différents projets représentatifs de normalisation...................... 21
1. Projet Archivage du LACITO................................................................................. 21
   a) SoundIndex ...................................................................................................... 21
   b) Moteur de recherche des archives du LACITO.............................................. 22
2. Journée BNF........................................................................................................... 23
3. Bilan ....................................................................................................................... 24

III Faciliter la normalisation pour les linguistes : le projet innovant de l’ED268...... 25
1. Présentation du projet............................................................................................. 25
   a) Collaboration pluridisciplinaire ...................................................................... 25
   b) Utilisation et création d’outils ......................................................................... 25
      i. Praat ............................................................................................................. 27
      ii. Transcriber ............................................................................................... 28
      iii. ELAN ....................................................................................................... 29
      iv. SoundIndex ............................................................................................ 30
2. Vers une aide plus adaptée pour les chercheurs..................................................... 31
   a) MakeMetadata (MKM) .................................................................................. 31
   b) Catalogue du PI-ED268................................................................................... 37
   c) Bilan sur le PI-ED268 .................................................................................... 40

Conclusion .................................................................................................................... 41

BIBLIOGRAPHIE .............................................................................................................. 42

ANNEXE .......................................................................................................................... 43
Table des figures

Figure 1: Exemple d'un codage HTML visualisé sur un navigateur Firefox. ........................................ 10
Figure 2: Rendu d'une description HTML sur un navigateur Firefox. .................................................. 12
Figure 3: Exemple de feuille de style XSLT. ......................................................................................... 13
Figure 4: Rendu de la feuille de style XSLT de la figure 3 en HTML. ..................................................... 13
Figure 5: Exemple de métadonnées au format HTML d'une page Web. ............................................. 14
Figure 6: Métadonnées du Dublin Core d'après Firefox Dublin Core. .................................................. 17
Figure 7: Schéma récapitulatif de l'imbrication des différentes normes et standards. ....................... 20
Figure 8: Exemple d'un codage de métadonnées associant Dublin Core et OLAC avec l'outil MakeMetadata. .......................................................................................................................... Erreur ! Signet non défini.
Figure 9: Copie d'écran SoundIndex. ..................................................................................................... 22
Figure 10: Moteur de recherche des archives du LACITO sous un navigateur Firefox. ..................... 22
Figure 11: Options langues et formats proposées par le moteur de recherche du LACITO. .................. 23
Figure 12: Copie d'écran du logiciel Praat. ............................................................................................ 27
Figure 13: Copie d'écran du logiciel Transcriber. .................................................................................. 28
Figure 14: Copie d'écran du logiciel ELAN. .......................................................................................... 29
Figure 15: Copie d'écran de SoundIndex. ............................................................................................... 30
Figure 16: Copie d'écran de l'onglet "HOME" du MKM. ......................................................................... 32
Figure 17: Copie d'écran d'un extrait de l'onglet "HELP-MKMETADATA". ........................................ 32
Figure 18: Copie d'écran d'un extrait de l'onglet "HELP-DC-OLAC". .................................................. 33
Figure 19: Copie d'écran de l'onglet "MKMETA1" du MKM. ................................................................. 33
Figure 20: Fenêtre pour écrire le texte associé au titre ou à la description. ............................................ 34
Figure 21: Copie d'écran de l'onglet "RESULT" du MKM. ....................................................................... 34
Figure 22: Sortie HTML numéro 1 possible avec l'onglet "EXPORT" du MKM. ................................. 35
Figure 23: Sortie HTML numéro 2 possible avec l'onglet "EXPORT" du MKM. ............................... 35
Figure 24: Moteur de recherche du catalogue du PI-ED268. ............................................................... 37
Figure 25: Exemple des métadonnées disponibles dans le moteur de recherche du catalogue du PI-ED268. ........................................................................................................................... 38
Figure 26: Résultat de la recherche pour dc:description. ................................................................. 38
Figure 27: Exemple de recherche en spécifiant une recherche de l'expression "bise" au sein des métadonnées "dc:description"........................................................................................................ 39
Figure 28: Exemple de recherche en spécifiant uniquement une recherche de l'expression "phonèmes" sur l'ensemble des métadonnées. ................................................................. 39

Table des tableaux

Tableau 1: Comparaison des corpus disponibles à l'écrit et à l'oral en 2004. ................................. 6
Introduction

Pour les chercheurs en linguistiques, il est beaucoup plus facile d'obtenir des données textuelles que des données orales. Ainsi, Véronis (Véronis 2004) estime à environ 8 milliards le nombre d'occurrences pour le français indexé par Google, alors que les plus grands corpus oraux pour cette langue ne dépassent pas à ce jour les 4 millions d'occurrences. Le tableau 1 ci-dessous donne un ordre de grandeur pour quelques corpus significatifs.

<table>
<thead>
<tr>
<th>Langue</th>
<th>Corpus écrits</th>
<th>Corpus oraux</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglais</td>
<td>100 millions d’occurrences British National Corpus (BNC) (échantillonnage textuels)</td>
<td>10 millions d’occurrences BNC</td>
</tr>
<tr>
<td></td>
<td>100 milliards d’occurrences (Google)</td>
<td>5 millions d’occurrences Longman American Spoken Corpus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 millions d’occurrences CANCODE</td>
</tr>
<tr>
<td>Français</td>
<td>210 millions d’occurrences Frantext (80% d’œuvres littéraires, 20% d’œuvres scientifiques ou techniques) quelques centaines de millions d’occurrences avec le journal Le Monde</td>
<td>4 millions d’occurrences Valibel (français de Belgique) 3,5 millions d’occurrences GARS-DELIC (français hexagonal)</td>
</tr>
<tr>
<td></td>
<td>8 milliards d’occurrences (Google)</td>
<td></td>
</tr>
<tr>
<td>Néerlandais</td>
<td></td>
<td>Corpus Gesproken Nederlands</td>
</tr>
<tr>
<td>Hébreu</td>
<td></td>
<td>Corpus of Spoken Israeli Hebrew</td>
</tr>
</tbody>
</table>


Les données linguistiques orales se composent d’enregistrements oraux en format audio et/ou vidéo, accompagnés de documents écrits tels que transcriptions, annotations, traductions. Le traitement de documents oraux s'avère ainsi beaucoup plus complexe que celui de documents écrits. Dans un premier temps, les données orales, si elles ne sont pas déjà dans un format numérique, doivent être numérisées et stockées. Puis des transcriptions de ces données sont réalisées. Puisque les transcriptions visées dépendent de l'étude linguistique, un même enregistrement oral pourra être transcrit de plusieurs manières : transcription phonétique, transcription orthographique, transcription privilégiant l’intonation, la syntaxe, etc.

De plus, l’oral présente des caractéristiques qui lui sont propres par rapport à l’écrit, avec des « erreurs » de prononciation, des expressions de l’oral qui ne se retrouve pas à l’écrit mais qu’il faut savoir transcrire malgré tout. Tout comme un même enregistrement peut être transcrit de plusieurs façons, une même transcription pourra elle-même être annotée de plusieurs façons dépendant des études linguistiques visées. Il en résulte que les données associées à un enregistrement audio sont amenées à pouvoir être enrichies, et qu’aucune transcription n’est figée. Cela explique la différence de quantités de données de l’oral traités par rapport à l’écrit.

Un des principaux objectifs des linguistes d’aujourd’hui est de pouvoir archiver leurs documents sonores dans un format qui en assure la pérennité et l’accessibilité, afin d’en favoriser la diffusion dans la communauté scientifique. En effet, avec le temps, les données matérielles sur lesquelles les enregistrements ont été réalisés, comme les bandes magnétiques, se détériorent, il faut donc transférer les données qu’ils contiennent sur des supports numériques qui sont d’une part plus fiables et d’autre part plus facile à dupliquer.
Cependant, il est indispensable de créer et de maintenir à jour un répertoire permettant de retrouver les ressources, et de leur attribuer un endroit de stockage limitant leur dégradation, afin d’en faciliter la réexploitation. Une fois les données orales numérisées et répertoriées, il va de l'intérêt scientifique général de pouvoir les mettre à disposition des autres chercheurs, qui pourront compléter leurs études grâce à ces nouvelles données. Il faut donc des mécanismes pour pouvoir faire une promotion efficace de ces données : il faut indiquer que ces données existent, où elles se trouvent, et comment et sous quelles conditions les exploiter.

Les progrès technologiques issus de l'ingénierie du document permettent de proposer des solutions pour répondre aux questions posées concernant l'archivage, la pérennisation, l'accessibilité et la diffusion des données linguistiques orales. Outre les normes d'encodage de données audio (qui ne seront pas traitées dans ce mémoire), de nombreuses normes et de standards existent, et je présenterai dans ce travail ceux qui semblent s'imposer parmi les linguistes. Tout d'abord, le standard Unicode permet de normaliser l’encodage des caractères afin de représenter toutes les graphies pour l'ensemble des langues du monde. Les normes SGML (Standard Generalized Markup Language) et XML (eXtensible Markup Language) se sont imposées pour l'annotation des documents. La norme du Dublin Core permet la description d'un contenu documentaire à l'aide d'un jeu de métadonnées, qui peut être complété par les métadonnées à caractères plus linguistiques proposées par le standard OLAC (Open Language Archive Community). Les normes du Dublin Core et d'OLAC font référence à d’autres normes et concepts, notamment la TEI (Text Encoding Initiative). Par ailleurs, le protocole de l’OAI (Open Archive Initiative) permet de créer, d’alimenter et de tenir à jour, par des procédures automatisées, des réservoirs d’enregistrements qui signalent, décrivent et rendent accessibles des documents, sans les dupliquer ni modifier leur localisation d’origine.

L'objectif d'une communauté de linguistes telle qu'OLAC est de proposer des standards pour la création, la manipulation et l'échange des données linguistiques. Des efforts plus locaux apportent des solutions qui peuvent ensuite se généraliser si les outils proposés répondent aux besoins d'autres chercheurs. Par exemple, le laboratoire LACITO (laboratoire Langues et Civilisations à Traditions Orales) a développé une architecture pour l'archivage et le partage de données orales. Des journées d'études permettent aux différents acteurs de la recherche en langue de débattre des problèmes rencontrés et d'essayer de proposer des solutions. Par exemple, une telle journée s'est tenue récemment pour la communauté francophone (journée d'étude « Constitution, exploitation, diffusion et conservation des corpus oraux » du mardi 17 mai 2005 à la Bibliothèque Nationale de France).

Une des problématiques majeures qui ressort des préoccupations des linguistes est le besoin de faciliter l'utilisation des outils informatiques de description et d'échange pour les linguistes. Ainsi, un projet innovant à l'Ecole Doctorale 268 Langage et Langues a vu le jour pour réconcilier des normes existantes parfois difficiles à manipuler, telles que celles du Dublin Core et d'OLAC, avec leur utilisateurs linguistes. L'objectif de l'outil MakeMetadata, créé dans le cadre de ce projet, est de faciliter l'annotation de données linguistiques à l'aide de telles normes en ayant recours à une interface utilisateur conviviale.

Dans une première partie, je vais introduire les outils pour la normalisation linguistique avec les normes génériques d’annotation des textes, tels SGML et XML, et les normes du générique au linguistique avec le Dublin Core et OLAC, associés à des concepts de TEI et

I Outils pour la normalisation linguistique

1. Norme pour la représentation des caractères : UNICODE

Le standard Unicode\(^3\) attribue un code unique à chaque caractère ou symbole utilisé dans les langues du monde, indépendamment de la plate-forme informatique, des logiciels et des langages de programmation. Les symboles de l'alphabet phonétique international, les idéogrammes chinois, les alphabets indiens et autres peuvent ainsi être codés. Unicode est synchronisé sur la norme ISO/IEC 10646\(^4\) qui normalise l’ensemble des caractères codés de façon universelle, permettant ainsi la saisie et la représentation des caractères des langues du monde sous forme écrite et de symboles complémentaires. La norme XML prend en charge le codage Unicode, et il existe maintenant de nombreuses polices implémentant tout ou partie des caractères représentés par la norme Unicode\(^5\).

2. Normes génériques d’annotation des textes

a) SGML (Standard Generalized Markup Language)

SGML (Standard Generalized Markup Language ou langage normalisé de balisage généralisé - norme ISO 8879)\(^6\) est un langage à balises qui permet de décrire la structure logique d’un document. SGML a été un langage pionnier pour la description intégrée des documents textuelles par le biais de métadonnées, et a été reconnu comme standard international de l'ISO en 1986. Les documents SGML peuvent être décrits par une grammaire appelée DTD (Document Type Définitions, ou définitions de types de documents), qui détermine les types d'éléments autorisés et les liens qu'ils peuvent entretenir entre eux. SGML, très adapté à la normalisation de documents, s'est néanmoins avéré difficile à utiliser en pratique car trop complexe et coûteux pour beaucoup de cas d'utilisation, sans être vraiment adapté au Web.

b) HTML (HyperText Markup Language)

Le langage HTML (HyperText Markup Language)\(^7\) est une simplification de SGML destinée à la présentation de documents pour Internet. Ce langage s'est très largement imposé comme le langage de description de documents sur Internet et est reconnu par de très nombreux outils, principalement les navigateurs web. Contrairement à SGML, la vocation d'HTML est très orientée vers la présentation des données textuelles annotées, puisque chaque élément correspond à une représentation visuelle, qui peut néanmoins varier légèrement en fonction des navigateurs. Un des points forts d'HTML est donc son indépendance vis-à-vis des architectures matérielles et logicielles, qui lui assure donc un usage universel.

De très nombreux outils permettent la création de documents HTML, allant de simples éditeurs de textes jusqu'à des éditeurs spécialisés offrant le plus souvent des modes de création visuels. L'annotation d'un fragment de texte se fait en l'entourant de deux balises, dites balises ouvrante et fermante. Les balises reprennent le nom d'éléments décrits dans une

---

\(^3\) Site de l’UNICODE : www.unicode.org/
\(^4\) Voir le site : www.iso.org/iso/fr/
\(^5\) Consulter le site suivant : http://www.alanwood.net/unicode/index.html
\(^6\) Site officiel du SGML : www.w3.org/MarkUp/SGML/
\(^7\) Site officiel du HTML : www.w3.org/MarkUp/
DTD, qui dans le cas d'HTML correspondent donc à un jeu restreint d'éléments de présentation. Ainsi, par exemple, la mise en gras du texte « XML » à l'aide de l'élément « b » (pour bold) se fait de la manière suivante: `<b>XML</b>`. Un navigateur tel que Mozilla Firefox\(^8\) interprétera cette annotation en faisant le rendu visuel illustré sur la figure 1:

![Affichage du navigateur Firefox](image1.png)

Les origines du **XML**, une combinaison du **SGML** et du **HTML**.

![Codage HTML sur un traitement de texte du type Notepad.](image2.png)

**Figure 1**: Exemple d'un codage HTML visualisé sur un navigateur Firefox.

c) **XML (eXtensible Markup Language)**

XML (eXtensible Markup Language)\(^9\) est né du besoin de pouvoir créer et manipuler des documents annotés plus facilement que ne le permettait SGML. La norme XML est donc une simplification de la norme SGML dont elle garde de nombreuses caractéristiques. Il est donc possible de produire du XML à partir d'outils développés pour SGML (mais bien entendu pas l'inverse, XML étant apparu après SGML). XML a obtenu le statut de recommandation du W3C (World Wide Web Consortium)\(^10\) en 1998.

Tout comme avec SGML et HTML, annoter un texte avec XML se fait en « encadrant » des fragment de texte avec des balises correspondant à des éléments décrits dans une DTD qui indique le type sémantique du fragment. L'extrait de document XML suivant illustre par exemple comment des informations relatives à une entrée bibliographique peuvent être décrites:

\(^8\)Site internet des produits mozilla : http://www.mozilla-europe.org/fr/products/firefox/

\(^9\) Site officiel d’XML : www.w3.org/XML/

\(^10\) Site du World Wide Web Consortium: www.w3.org/
La DTD associée étant la suivante :

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!DOCTYPE entree [
  <!ELEMENT entree (titre, auteurs, edition,)>
  <!ELEMENT titre (#PCDATA)>  
  <!ELEMENT auteurs (auteur+)>  
  <!ELEMENT auteur (#PCDATA)>  
  <!ELEMENT edition (annee, editeur, editeurs+, ouvrage, pages*)>
  <!ELEMENT annee (#PCDATA)>  
  <!ELEMENT editeur (#PCDATA)>  
  <!ELEMENT editeurs (#PCDATA)>  
  <!ELEMENT ouvrage (#PCDATA)>  
  <!ELEMENT pages (#PCDATA)>  
]
```

Ces informations auraient déjà pu être contenues dans une page au format HTML telle que celle de l'exemple suivant:

```html
<html>
  <body>
    Bird Steven and Gary Simons (2004), "Building on Open Language Archives Community on the DC Foundation", in Hillman and Westbrooks (editors), Metadata in Practice: A Work in Progress, ALA Editions, pp 203-222.
  </body>
</html>
```

Néanmoins, cette description ne permet pas de pouvoir extraire automatiquement des informations telles que le titre ou le nombre et le nom des auteurs, puisqu'elle se concentre uniquement sur des choix arbitraires de présentation. Par exemple, ici comme le montre la figure 2, le titre apparaît en gras, le titre de l'ouvrage en italique et les noms des auteurs en style normal, et l'ordre des éléments a été choisis pour répondre à un style de présentation particulier. XML permet donc de s'affranchir des informations de présentation et ainsi d'extraire facilement l'information recherchée en faisant référence à leur description sémantique contenue dans la DTD utilisée dans le document.
Les documents structurés obtenus par annotation en SGML ou XML permettent donc l'échange et la réutilisation de leur contenu, ainsi que la pérennisation de celui-ci, quelles qu'en soient ses utilisations futures. Chaque élément d'un tel document peut effectivement être extrait par filtrage et réutilisé pour créer de nouveaux documents ou alimenter des bases de données structurées. Des langages de transformation de documents XML tels que XSLT\(^{11}\) (eXtensible Stylesheet Language Transformation) existent, et permettent ainsi de produire simplement de nouveaux documents XML pouvant être exploités de différentes façons. Il est en effet possible d'extraire sélectivement des informations dans un document XML et de les disposer dans un nouveau document XML destiné à la présentation, par exemple en respectant la DTD XHTML, sous-partie de HTML respectant la norme XML.

L'exemple suivant illustre la transformation d'un document XML à l'aide d'une transformation XSLT en document HTML\(^{12}\):

---
\(^{11}\) Site des feuilles de styles XSLT : www.w3.org/TR/xslt
**Feuille de style adresses1.xsl**

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="html"/>
  <xsl:template match="/">
    <html>
      <head>
        <title>Carnet d'adresses</title>
      </head>
      <body>
        <h1>Adresses postales</h1>
        <ul>
          <xsl:apply-templates select="adresses/adresse"/>
        </ul>
      </body>
    </html>
  </xsl:template>
  <xsl:template match="#adresse">
    <li>
      <b><xsl:value-of select="prenom"/>
        &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &nbsp; &n...
3. Normes d'annotation par métadonnées : du générique au linguistique

a) Métadonnées

Face à l'augmentation constante de la quantité de documents électroniques disponibles sur le Web et dans les bases de données, il est devenu indispensable de structurer les informations décrivant ces ressources afin de pouvoir réaliser des recherches plus performantes sur ces informations.

En effet, le critère retenu pour la recherche sur le Web est souvent la fréquence d'un mot recherché dans les pages, ce qui n'assure pas toujours une bonne pertinence des résultats. Associer les documents mis en ligne avec un ensemble d'informations permettant de les décrire de façon rigoureuse apparaît ainsi comme une façon promise de d'améliorer la recherche d'informations sur le Web.

Des informations telles que les auteurs d'un document, sa date de création, son type de contenu, permettent de décrire la ressource de façon fine et de pouvoir mieux juger de sa pertinence par rapport à une requête. Les moteurs de recherche du Web s'appuient encore peu sur ce type d'informations. En revanche, elles constituent des informations susceptibles de constituer des bases de données structurées.

Les « métadonnées » représentent l'ensemble des informations techniques et descriptives associées aux ressources électroniques afin de décrire leur contenu et leur format. Une description fine du contenu à l'aide de ces éléments, également appelés « descripteurs » ou « étiquettes » en informatique, permet de faciliter la recherche des informations décrites. Ainsi les métadonnées permettent de cataloguer les ressources électroniques, et donc de pouvoir les retrouver efficacement et les réutiliser de façon pertinente.

Les métadonnées sont décrites à l'aide de modèles décrivant des éléments et les manières de les agencer entre eux, notamment les DTD et les schémas SGML et XML. La figure 5 suivante illustre un exemple de métadonnées dans une page Web au format HTML.

Voici un exemple de métadonnées pour le Web :

Figure 5: Exemple de métadonnées au format HTML d'une page Web.
Les métadonnées d'une page Web sous format HTML sont décrites à l'aide des éléments «META» et de leurs attributs «NAME» et «CONTENT»:

\[<\text{META NAME}="\ldots" \text{CONTENT}="\ldots">\]

L'attribut «NAME» spécifie un type d'information, par exemple le logiciel ayant servi à la création de la page (valeur «GENERATOR»). L'attribut «CONTENT» permet de spécifier la valeur du type d'information précisé (ici «PIED268»).

Les métadonnées peuvent être liées aux ressources qu’elles décrivent de deux manières. Elles peuvent être contenues dans un fichier séparé de la ressource (comme une fiche de catalogue à la bibliothèque), ou intégrées dans la ressource elle-même. Le choix se fera en fonction des besoins de la communauté, et des caractéristiques propres de la ressource. Par exemple, pour le projet innovant ED268, l'utilisation de fichiers séparés de la ressource a été retenue, afin de pouvoir repérer facilement l’ensemble des données disponibles dans la base de données, et ne télécharger effectivement que les ressources souhaitées.

Les métadonnées doivent cependant pouvoir être reconnues de manière uniformes pour permettre une utilisation efficace. Il est donc important que les différents acteurs (éditeurs de logiciels, annotateurs des ressources, utilisateurs) utilisent une définition et une nomenclature commune pour la description et la manipulation de ces métadonnées.

Pour répondre à ce besoin de standardisation, plusieurs organismes de standardisation ont proposé et publié des schémas de métadonnées conçus de sorte à être adoptés par le plus grand nombre. On retrouve notamment les normes du Dublin Core (DC)\(^\text{13}\) et de l’Open Language Archive Community (OLAC)\(^\text{14}\).

\textit{b) Dublin Core}

La norme du Dublin Core est issue d’une journée d’étude ayant eu lieu en 1995, à Dublin (Ohio), sous le nom d’Initiative aux Métadonnées du Dublin Core (\textit{Dublin Core Metadata Initiative} (DCMI)). Dans un premier temps, les objectifs du Dublin Core était d’améliorer les standards de recherche d’information des ressources en uniformisant les données essentielles à fournir pour décrire le contenu d’un document et en déduire sa pertinence par rapport à une requête.

Pour ce faire, un ensemble d’acteurs provenant de différentes communautés (bibliothécaires, informaticiens, producteurs de données, experts en annotations de documents…) ont décrit une liste minimale de descripteurs ayant un intérêt général pour le plus grand nombre de communautés. Cette liste se voulait assez fine sans pour autant saturer le créateur de métadonnées d’éléments spécialisés.

Il en résulte une liste de quinze éléments optionnels et répétables, ayant été normalisée en 2003 par la norme ISO-15836 sous le nom «Information and documentation – The Dublin Core metadata element set »\(^\text{15}\). Cette liste se décompose selon les trois catégories suivantes\(^\text{16}\):

\(^\text{13}\) Site du Dublin Core : http://www.dublincore.org/

\(^\text{14}\) Site d’OLAC : www.language-archives.org/

\(^\text{15}\) http://dublincore.org/documents/dces/
Contenu :
Titre (Title) : nom donné à la ressource
Sujet (Subject) : mots-clés, code de classification, vocabulaire contrôlé
Description (Description) : résumé, table des matières, texte libre
Source (Source) : référence à une ressource à partir de laquelle la ressource actuelle a été dérivée.
Langue (Language) : langue du contenu intellectuel de la ressource
Relation (Relation) : référence à une autre ressource en rapport avec cette ressource.
Couverture (Coverage) : portée spatio-temporelle de la ressource.

Propriété intellectuelle
 Créateur (Creator) 17: entité principalement responsable de la création de la ressource
 Editeur (Publisher) : entité responsable de la diffusion de la ressource.
 Contributeur (Contributor) : entité qui a contribué à la création du contenu de la ressource.
 Droits (Rights) : propriété intellectuelle, copyright

Administratif (Instanciation)
 Date (Date) : date de création ou de publication de la ressource.
 Type (Type) : nature ou genre du contenu de la ressource
 Format (Format) : matérialisation physique ou digitale de la ressource
 Identifiant (Identifier) : référence non ambiguë de type ISBN, URL


Les métadonnées du Dublin Core représentent un grand intérêt pour décrire des ressources. Elles doivent être facile à créer et à comprendre, pour permettre à des communautés diverses et variées de se les réapproprier. Pour cela, elles ont même « été traduites dans une vingtaine de langues et approuvé par une dizaine de gouvernement »21. Le Dublin Core n’a pas pour intention de remplacer les autres standards de métadonnées, mais de coexister, parfois au sein

17 Nous noterons qu’à présent le Dublin Core recommande de ne plus utiliser l’étiquette « créateur », mais « contributeur » à la place, car il est rare qu’une personne crée entièrement une ressource toute seule, elle participe plus ou moins, et contributeur répond à ce critère. Ce qui nous fait un total de 14 étiquettes au lieu de 15.
18 Consortium World Wide Web http://www.w3.org/
d’une même description de ressource en proposant une sémantique simplifiée, comme c’est le cas avec OLAC.

La figure 6 donne un aperçu des métadonnées du Dublin Core issues du site du projet innovant de l’Ecole Doctorale 268 Langage et Langues. L’obtention de cette fenêtre se fait grâce à une extension du navigateur Firefox qui permet de visualiser les métadonnées du Dublin Core associées aux documents électroniques du Web.

![Dublin Core Metadata](image)

Figure 6: Métadonnées du Dublin Core d’après Firefox Dublin Core.

c) TEI (Text Encoding Initiative)


Bien que la TEI mette l’accent sur les données textuelles, elle permet également de décrire des fichiers représentant du son ou de l’image. Les recommandations de la TEI permettent d’expliciter certaines caractéristiques d’un document afin d’en faciliter le traitement par des programmes informatiques pouvant s’exécuter sur des plates-formes différentes. En effet, la TEI s’appuie sur la norme SGML, et donc tous les outils SGML généralistes peuvent être utilisés pour manipuler des textes conformes à la TEI.

d) OAI (Open Archives Initiative)

En 1999, lors de la convention de Santa Fé, l’OAI (Open Archives Initiative) a mis en place le protocole OAI-PMH (Open Archives Initiative’s Protocol for Metadata Harvesting) permettant de faciliter l’échange de données. Ce protocole simplifie la description et la

---

diffusion des métadonnées de documents scientifiques ou culturels disponibles en accès ouvert sur Internet. Le protocole a atteint une certaine stabilité avec sa version 2 publiée en 2002.

Ce protocole d'échange permet de créer, d'alimenter et de tenir à jour, par des procédures automatisées, des réservoirs d'enregistrements qui signalent, décrivent et rendent accessibles des documents. Cette accessibilité est assurée sans que les documents ne soit dupliqués ni déplacés de leur emplacement d'origine.

Le protocole de l'OAI offre donc une visibilité des ressources disponibles et de leur emplacement sur le Web, en collectant les métadonnées des documents accessibles\(^\text{27}\). Il permet le signalement de ressources non accessibles aux moteurs de recherche, de faciliter l'interopérabilité des ressources documentaires, et de mettre à jour facilement et automatiquement des métadonnées collectées et des liens. Cette mise à jour est effectuée en récupérant les dernières modifications apportées aux réservoirs sources, sans que l'intégralité des données n'ait à être copiée, ce qui permet d'éviter l'encombrement des serveurs qui les hébergent.

e) OLAC (Open Language Archive Community)

La communauté OLAC\(^\text{28}\) (Open Language Archive Community) est l'émanation d'une collaboration entre trois organismes linguistiques internationaux : le LDC (Linguistic Data Consortium)\(^\text{29}\), le SIL (Summer Institute of Linguistics) International\(^\text{30}\) et la liste de diffusion LINGUIST \(^\text{31}\). Le LDC crée et partage des ressources linguistiques, appuie le développement de formations de recherche et technologie concernant le langage, et publie des bases de données linguistiques, ouvertes aux institutions mondiales de recherche (Bird et Simons (2004)). Le SIL International facilite le développement de la linguistique sur plus de 1000 langues dans le monde, et il est utilisé par des individus du monde entier dans les domaines de la recherche, de la traduction et de la littérature. La liste LINGUIST est une liste de diffusion sur le thème de la linguistique, qui accueille plus de 2 000 pages de contenu, 100 listes de diffusion et 20 000 abonnés dans le monde. OLAC est donc une communauté regroupant de nombreux linguistes, et qui peut donc être à même de répondre de façon pertinente aux problèmes qu'ils rencontrent dans leur travail.

Les normes d'OLAC permettent de créer et de réutiliser des collections de ressources linguistiques comprenant des données, des outils et des recommandations. Les métadonnées proposées permettent de décrire les ressources que les linguistes ont créées et de les aider à trouver les ressources dont ils ont besoin si elles existent. Ces métadonnées sont construites à partir de la norme du Dublin Core.

Afin de respecter la norme du Dublin Core, OLAC ne remplace pas les métadonnées du Dublin Core, mais les spécifie par rapport aux attentes de la communauté linguistique. Cinq

\(^{27}\) Site de François NAWROCKI du Ministère de la culture et de la communication http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm

\(^{28}\) Site OLAC : www.language-archives.org/

\(^{29}\) Site du LDC : www.ldc.upenn.edu/

\(^{30}\) Site du SIL : www.sil.org/

\(^{31}\) Site LINGUIST : www.linguistlist.org/
extensions au Dublin Core concernant la description de ressources linguistiques ont été proposées:

- **Type de discours (Discourse Type)** : comprend des types du genre « dramatique », « narration », « jeux de mots », etc.
- **Identification de la langue (Language Identification)** : fait référence aux codes ISO (« fr » pour français, « en » pour anglais, etc)
- **Champ linguistique (Linguistic Field)** : propose des champs tels que sociolinguistique, phonétique, etc.
- **Type de données linguistiques (Linguistic Data Types)** : on retrouve ici trois type de données, « lexique », « texte-primaire » et « description de langue ».
- **Rôles des participants (Participant Roles)** : liste des différents rôles que peuvent avoir les participants, tels qu’annotateur, auteur, locuteur, etc.

Voici une illustration de métadonnées utilisant à la fois Dublin Core et OLAC :

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<olac:olac
 xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
 xmlns:dc="http://purl.org/dc/elements/1.1/
 xmlns:dcterms="http://purl.org/dc/terms/
 xmlns:xsl="http://www.w3.org/2001/XMLSchema-instance">
<dc:title xml:lang="fr">Exemple d'un codage associant Dublin Core et OLAC </dc:title>
<dc:subject xsl:type="olac:discourse-type" olac:code="language_play"/>
<dc:subject xsl:type="olac:linguistic-fields" olac:code="phonetics"/>
<dc:description xml:lang="fr">Représente un exemple de génération de métadonnées provenant à la fois du Dublin Core et d’OLAC. </dc:description>
<dc:contributor xsl:type="olac:role" olac:code="annotator"> Fred </dc:contributor>
<dc:contributor xsl:type="olac:role" olac:code="author"> Zia Zoupsie Louve </dc:contributor>
<dc:contributor xsl:type="olac:role" olac:code="researcher"> Jodard </dc:contributor>
<dc:contributor xsl:type="olac:role" olac:code="translator"> Tchoumf </dc:contributor>
<dc:created xml:lang="fr"> 2005-09-14 </dc:created>
<dc:valid xml:lang="fr">2005-09-21 </dc:valid>
<dc:type xsl:type="olac:linguistic-type" olac:code="primary_text"/>
<dc:language xsl:type="olac:language" olac:code="fr"/>
</olac:olac>
```
Figure 7: Exemple d’un codage de métadonnées associant Dublin Core et OLAC avec l’outil MakeMetadata.

Pour conclure cette première sous-partie, la figure 8 illustre les différentes relations entre les normes et les standards détaillés ci-dessus.

Figure 8: Schéma récapitulatif de l’imbrication des différentes normes et standards.
II Présentations de différents projets représentatifs de normalisation

1. Projet Archivage du LACITO

Des chercheurs du Laboratoire de Langues et Civilisations à Traditions Orales (LACITO) collectent des enregistrements de parole spontanée, principalement dans les langues dites « rares », afin de contribuer à la documentation des langues du monde et à l’étude du patrimoine humain. Ils ont donc à traiter des documents sonores avec leurs transcriptions synchronisées et des traductions associées. La structuration de ces données est réalisée à l’aide de normes informatiques récentes dans un format ouvert, et une application informatique d'archivage à architecture ouverte a été développée au sein du LACITO. Dans cette partie, je décrirai cette application et soulignerai les aspects qui ont été jugés importants pour le développement de notre propre application d'archivage dans le cadre de notre projet innovant à l’Ecole Doctorale 268 Langage et Langues.

Une telle application informatique a émergé des besoins des linguistes de pérenniser les données collectées, ainsi que de les rendre accessibles et disponibles. Elle implique des problèmes de codage de caractères et d'alignements pour lesquelles les normes de l'Unicode et XML du W3C ont été retenues.

L'application informatique développée au LACITO est principalement dédiée à la production de documents associant son numérisé et texte. L’aspect texte comprend au minimum la transcription phonologique et la traduction libre en français ou en anglais, accompagnées, selon le cas, de transcriptions orthographiques pour les langues n’utilisant pas l’alphabet latin, de traductions en d’autres langues, de gloses morphologiques, de notes, etc. L’alignement de la transcription avec le son se fait généralement au niveau de « la phrase » ou du groupe intonatif, mais peut se faire également au niveau du mot ou du morphème.

Je vais décrire les trois types de développements logiciels réalisés au LACITO: SoundIndex, un outil pour la création de documents contenant des alignements entre de la parole orale et sa transcription; une interface permettant de lire les documents enrichis produits à l'aide de SoundIndex; une architecture pour diffuser et interroger ces documents sur Internet. La présentation de l'architecture de diffusion et d'interrogation sera l'occasion d'illustrer le concept de « métadonnées », et les normes du Dublin Core Metadata Initiative (DCMI) et de l'Open Language Archives Community (OLAC) utilisées dans le projet du LACITO, ainsi que le protocole de l'Open Archives Initiative (OAI) pour l'accès à ces données.

a) SoundIndex

L’une des attentes des chercheurs consiste à pouvoir accéder à la transcription d’un texte au fur et à mesure de la lecture de l’enregistrement sonore. Dans cette optique, le LACITO a développé le logiciel SoundIndex. Écrit en Tcl/Tk et XML il permet de travailler sur de nombreux systèmes d’exploitation. Le son et la transcription sont lues à partir d’un même fichier. Le logiciel insère dans chaque segment du document une balise XML <AUDIO> pour lié le texte au son.

32 LACITO http://lacito.vjf.cnrs.fr/archivage/index_fr.html
La figure 9 illustre l’utilisation de SoundIndex. La partie du haut de la fenêtre montre la transcription et son annotation correspondant au signal sonore encadré dans la partie du bas.

Figure 9: Copie d'écran SoundIndex.

b) Moteur de recherche des archives du LACITO

Pour accéder aux archives du LACITO, une page Web consultable depuis n'importe quel navigateur, a été créée pour interroger la base grâce aux métadonnées. Ces métadonnées concernent ici principalement la langue, la culture, les participants et la situation d’enregistrement. Ces métadonnées sont à la fois encodées d’après la norme du Dublin Core, et complétées par le standard de métadonnées proposé par OLAC. Ces informations sont accessibles grâce au protocole de l’OAI. L’utilisateur doit donc définir les caractéristiques pertinentes pour sa recherche en fonction des informations contenues dans les métadonnées.

La figure 10 illustre un exemple de recherche dans la base de données du LACITO sous le navigateur Firefox :

Figure 10: Moteur de recherche des archives du LACITO sous un navigateur Firefox.
Nous pouvons noter ici que ce moteur de recherche permet uniquement de retrouver des ressources déjà existantes, et ne permet pas d’en créer de nouvelles. C’est ce que nous montre le choix particulier des langues et des formats qui sont prédéfinis dans la figure 11.

### Figure 11: Options langues et formats proposées par le moteur de recherche du LACITO.

2. Journée BNF

Il n'est pas toujours évident d'exploiter des corpus oraux, car cela pose les problèmes de leur stockage, de leur diffusion et de leur exploitation proprement dite. Un groupe de travail constitué de linguistes, de juristes et de conservateurs, a donc essayé de mettre au point un « Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux » lors d'une journée d'étude ayant eu lieu le 17 mai 2005 à la Bibliothèque Nationale de France. Leur but était d’ouvrir un débat et de proposer un guide dont la vocation à terme sera d'être diffusé le plus largement possible. Ce guide contient de précieuses informations sur les aspects juridiques liés aux corpus oraux, ainsi que des conseils sur les aspects matériels concernant aussi bien la structuration et la mise en forme des données, que la conservation, l’exploitation et la diffusion de ces données. Cela permet d’aider les chercheurs à constituer ou enrichir leurs corpus oraux.

Un des point fort de ce guide est sa partie concernant les aspects juridiques de la création de corpus oraux. En effet, en cas de litige avec un locuteur par exemple, il est bien difficile à un chercheur de décrire le problème à un juriste. Un chercheur n’est pas nécessairement familier avec le jargon juridique, et le juriste ne connaît pas ses préoccupations spécifiques. Le guide

[34] http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm
permet donc de donner aux linguistes des éléments pour se faire comprendre et mieux se faire aider par des juristes en donnant des éléments réponses pour savoir où chercher l'information et l’aide nécessaire. Par exemple, le guide contient des informations relatives aux questions « Quel peut être le statut juridique de chacun des corpus oraux constitués par les chercheurs ? » (p.21 du guide provisoire), « Quand est-il question de droit d’auteur à propos des corpus ? » (p.23), « Le respect de la vie privée dans la constitution, l’exploitation, la diffusion et la conservation des corpus » (p.25), ou « Quelles sont les responsabilités des personnes en charge de la diffusion des corpus sur Internet ? » (p. 26).

Le guide donne également des conseils sur les démarches à suivre pour constituer, exploiter, conserver et diffuser des corpus oraux. Une fois encore, il n’est pas question de donner une méthodologie fixe à suivre scrupuleusement, mais simplement de sensibiliser le chercheur sur les points juridiques et éthiques pertinents. Le guide effectue donc un état des lieux en ce qui concerne les « éléments de la situation en jeu » (paragraphe 3.2), les « corpus et les types des données », les « modes d’enregistrement » (paragraphe 3.2.1.1), les « supports d’enregistrement » (paragraphe 3.2.1.2). Il fait également le point sur la standardisation et la normalisation. Il explicite ce qu’induit les techniques d’enquête choisies pour le recueil et la production des données.

Le guide adresse les problématiques liées à l'analyse des données: quel type de transcription choisir ? Quels en sont les enjeux ? Une discussion est faite sur les limitations liées aux outils utilisés pour les phases d'analyse et les objets de ces analyses.

Enfin, le guide rappelle la situation des corpus oraux en tant qu’objets de patrimoine en France, puisque un retard important a été pris dans ce domaine par rapport à d'autres langues comme l'anglais. Le rôle des organismes en charge de l'archivage des ressources orales est expliqué, et le guide incite les chercheurs à unir leurs efforts afin de créer de grands corpus oraux pour le français.

Ce guide était provisoire à l'heure de la rédaction de ce mémoire, il n'en est pas moins un appel à ce que les chercheurs de différentes disciplines donnent leur opinion sur son contenu et collaborent. Il est donc utile d'y chercher régulièrement des mises à jour.

3. Bilan

Bien qu'un guide tel que celui publié lors de la journée à la BNF donne beaucoup de conseils, son but n’est pas d'apporter de l’aide concrète pour aider les chercheurs à répertorier leurs corpus. Par ailleurs, si des ressources significatives existent pour l'oral, telles que la base de données du LACITO, les chercheurs externes disposent de peu d'indications pour créer eux-mêmes leurs propres données. En général, les projets existants ne mettent pas toujours l’accent sur la réutilisabilité par d’autres chercheurs des logiciels et ressources produits.

La journée d’étude de la BNF est un premier pas important vers une réflexion commune visant à mettre en place des approches validées par plusieurs disciplines linguistiques pour la création et l’exploitation de corpus oraux. Néanmoins, l’absence d’acteurs de la communauté du Traitément du Langage Parlé, fait que les propositions de la première version du guide manquent de généralité pour répondre aux besoins de toutes les disciplines concernées. On sent que plus de cohésion entre différents projets qui s’attaquent à des problèmes similaires est important, et que la réflexion entamée à la BNF doit se poursuivre.
III Faciliter la normalisation pour les linguistes : le projet innovant de l’ED268

1. Présentation du projet

Pour répondre aux besoins d’archivage, de stockage, de diffusion et d’exploitation des données linguistiques qui peuvent être utiles à de nombreux chercheurs, un projet innovant financé par le Conseil Scientifique de l'Université Paris 3 - Sorbonne Nouvelle a été lancé. Le projet PI-ED268, dont Serge Fleury est le responsable, doit mettre en place une base de données pluridisciplinaire pour les membres de l’Ecole Doctorale 268 Langages et Langues. Cette base de données permettra de déposer et de partager des corpus oraux et vidéos.

L’objectif de ce projet est de proposer une réflexion vers une démarche de normalisation lors de la conception de bases de données de ressources linguistiques regroupant des données de langues et de natures différentes. Les propositions réalisées sont plus particulièrement orientées vers des corpus oraux, du fait des caractéristiques différentes entre les ressources orales et textuelles. En effet, la normalisation de corpus oraux pose davantage de contraintes que celle de corpus textuels, puisque l’on est confronté à des enregistrements sonores accompagnés de données textuelles pour les transcriptions et annotations. En outre, ces transcriptions et annotations sont rédigées à l'aide d'outils différents, qui ne sont pas nécessairement compatibles entre eux.

Un des résultats escomptés de ce projet est de favoriser les collaborations pluridisciplinaires. A cette fin, des recommandations d'outils permettant de supporter une démarche de normalisation sont faites, et une aide est proposée aux chercheurs pour construire leurs corpus.

a) Collaboration pluridisciplinaire

Ce travail est donc réalisé dans un cadre pluridisciplinaire faisant intervenir les différents domaines de la linguistique et le traitement automatique des langues. La mise en commun de corpus récoltés pour ces disciplines est en effet indispensable afin de pouvoir tenir compte des besoins propre à chaque discipline. Ce point a naturellement favorisé une collaboration interdisciplinaire au sein de l’équipe pour ne pas orienter de manière irrémédiable cette base de corpus vers une discipline particulière. Dans un futur proche, on espère ainsi le rapprochement de plusieurs disciplines pour des travaux communs sur de mêmes corpus. Des expériences ont déjà été tentées dans ce sens, notamment par le biais de conférences appelant différentes communautés à travailler sur un corpus unique, comme la journée de la BNF. De plus, nous travaillons en étroite collaboration avec le LACITO (plus particulièrement Michel Jacobson), ce qui nous permet d’enrichir notre travail en bénéficiant de l’expérience des travaux réalisés dans ce laboratoire.

b) Utilisation et création d’outils

Pour commencer notre travail, nous avons regroupé 13 corpus intégrant une annotation linguistique, provenant de formats différents (audio et vidéo) et de disciplines variées: acquisition du langage, syntaxe et sémantique, sociolinguistique, phonétique et phonologie. Précisons qu’il est nécessaire de prendre en compte trois types de données : les données...
brutes des ressources (flux audio et vidéo), les annotations linguistiques (qui accompagnent et décrivent les données brutes), et les métadonnées que nous devons créer pour décrire ces différents fichiers de manière précise, sur le même principe qu’une fiche de bibliothèque qui décrit les caractéristiques d’un ouvrage.

Les travaux d’annotation fournis avec nos corpus ont été effectués en utilisant plusieurs outils. Cependant, les outils d’annotation linguistique (par exemple Transcriber), de traitement du son (par exemple Praat), de traitement de l’image (par exemple Anvil / Clan), ou de création de textes (par exemple Word) ne sont pas toujours compatibles entre eux car ils ont souvent des formats propriétaires. Notre objectif étant de mettre à disposition des ressources accessibles et réutilisables par tous, il nous était donc indispensable de trouver un format de normalisation au niveau de la structure de ces annotations. XML est un métalangage libre de droit qui nous permettait l’insertion d’informations très complètes dans les documents décrits. Des passerelles peuvent ainsi être facilement construites pour passer des formats propriétaires générés par les logiciels mentionnés, au format XML. Par exemple Praat2XML, disponible sur Internet, permet la conversion du format propriétaire de Praat en une sortie XML exploitable.

L’encodage des caractères au sein de l’annotation peut également apporter son lot de problèmes. L’utilisation du codage Unicode, qui attribue un même indice pour chaque caractère graphique, indépendamment de la plate-forme informatique, du logiciel ou de la langue, permet d’éviter ces difficultés s’il est utilisé dans les formats de sortie des outils. Pour approfondir ce dernier point, une réflexion sur la TEI s’avère utile pour fournir des outils intégrant cette normalisation de l’annotation. Nous allons donc indiquer quelques outils qui supportent la TEI et génèrent une sortie XML. De plus, ces outils peuvent tous être utilisés gratuitement.
i. Praat

Praat\textsuperscript{35} est un outil qui permet une analyse acoustique de signaux sonores en plus de l’annotation linguistique. Il permet une transcription fine, et offre huit niveaux d’analyse indépendants avec un ancrage temporel nécessaire. Praat peut produire des sorties XML avec des DTD propriétaires. La figure 12 représente la fenêtre d’édition du signal affichant de haut en bas le signal sonore, le spectrogramme, la fréquence fondamentale mesurée à partir du signal, et l’annotation.

Figure 12: Copie d’écran du logiciel Praat.

\textsuperscript{35} Praat : http://www.fon.hum.uva.nl/praat
Transcriber\textsuperscript{36} est un logiciel d’annotation linguistique spécifiquement conçu pour la transcription lexicale des dialogues. Il propose un seul niveau d’analyse, et génère un format XML avec une DTD spécifique.

\textbf{Figure 13: Copie d’écran du logiciel Transcriber.}

\textsuperscript{36} Transcriber : http://www.etca.fr/CTA/gip/Projets/Transcriber/
iii. ELAN

Elan est un logiciel d’annotation linguistique très complet: il autorise notamment l’annotation de la vidéo, offre plusieurs niveaux d’analyse (dépendants / indépendants, nécessité de spécifier l’ancrage temporel pour chaque niveau d’analyse) et permet d’exporter en format XML avec une DTD spécifique.

Figure 14: Copie d’écran du logiciel ELAN.

37 ELAN: (http://www.mpi.nl/tools/elan.html)
iv. SoundIndex

SoundIndex \(^{38}\) est un logiciel qui permet plusieurs niveaux d’analyse de données audio. Il est possible de spécifier l’ancrage temporel des transcriptions pour chaque niveau d’analyse, ce qui n’est cependant pas obligatoire. SoundIndex génère une sortie XML sans avoir recours à une DTD spécifique. On peut regretter certains problèmes liés à l’ergonomie du logiciel, qui ne permet par exemple pas la possibilité d’annuler les actions effectuées.

Figure 15: copie d’écran de SoundIndex

\(^{38}\) SoundIndex : http://michel.jacobson.free.fr
2. Vers une aide plus adaptée pour les chercheurs

Une réflexion sur la normalisation de l’encodage et de la description de corpus de langues devient nécessaire dans la mesure où l’on entre dans une démarche d’exploitation, de conservation et de diffusion des données. En effet, il est important d’utiliser des normes de description utilisées par le plus grand nombre de communautés, ainsi qu’un format de représentation universel.

La première étape du travail d'un linguiste consiste souvent en la constitution d’un corpus. Cette activité est au croisement d’aspects théoriques essentiels pour la vérification des hypothèses scientifiques d'une étude particulière. Mis à part ces aspects théoriques, un corpus doit également répondre à des exigences méthodologiques. La possibilité de pouvoir intégrer différents corpus dans une seule base de données, comme celle proposée par le PI-ED268, permet aux chercheurs désireux de vouloir y contribuer de le faire. Il leur faut cependant être attentifs aux informations qu’il leur sera impératif de fournir pour que leur corpus puisse être pleinement exploité, notamment: nom des locuteurs, type exact de données recueillies, noms et rôles des différents contributeurs, format des données, etc.

Ainsi, ce projet permet de proposer des approches de normalisation pour l’encodage de corpus oraux, ainsi que des méthodes de constitution d’un corpus utilisant des outils performants permettant la normalisation pendant les phases d’enregistrement, d’annotation, de transcription, puis d’analyse. Ce projet favorise également les collaborations inter et pluridisciplinaires, du fait même qu’il se centre sur les besoins de plusieurs disciplines en bénéficiant de la collaboration de plusieurs représentants de ces disciplines. Enfin, il a pour objectif d’aider les chercheurs à constituer leurs corpus à l’aide d’un outil convivial qui permet de suivre les normes actuelles en vigueur pour la normalisation de données linguistiques.

a) MakeMetadata (MKM)

L’outil MakeMetadata (MKM ci-après) permet de créer et de modifier des métadonnées de ressources linguistiques pluridisciplinaires selon les normes en vigueur pour la normalisation de ressources linguistiques sous format électronique. Il se base sur la norme de métadonnées du Dublin Core et le standard de métadonnées appliquées à la linguistique OLAC. MKM produit des fichiers XML valides et normalisés.

MKM est disponible sous deux formats différents: un programme exécutable pour Windows39, et un ensemble de scripts Perl/Tk pouvant être utilisés sur de nombreux systèmes d'exploitation40. Un mode d’emploi est disponible41, il peut être trouvé en annexes de ce mémoire.

40 MKM en version Perl/tk est disponible sur demande
L’interface du programme comportent différents onglets :
- HOME : donne des informations sur le programme

Figure 16: Copie d’écran de l’onglet "HOME" du MKM.

- HELP-MKMETADATA : module d’aide sur le programme

Figure 17: Copie d’écran d’un extrait de l’onglet "HELP-MKMETADATA".
HELP-DC-OLAC : module d'aide pour les éléments du Dublin Core et d'OLAC qui spécifie notamment leurs relations. Des éléments d'aide peuvent être obtenus de manière contextuelle en cliquant à droite des éléments dans les onglets suivants.

MKMETA1, MKMETA2, MKMETA3, MKMETA4, MKMETA5 et MKMETA6 : description de la totalité des métadonnées du Dublin Core et d'OLAC que l'utilisateur doit spécifier. Cela se fait soit en cochant les cases correspondant au corpus décrit, soit en sélectionnant un code de langue, soit en remplissant directement l’information dans une fenêtre (titre et description).

Figure 18: Copie d'écran d'un extrait de l'onglet "HELP-DC-OLAC".

Figure 19: Copie d'écran de l'onglet "MKMETA1" du MKM.
La figure 20 montre un exemple de spécification de données par l'utilisateur.

Figure 20: Fenêtre pour écrire le texte associé au titre ou à la description.

-RESULT : permet d'effectuer un certain nombre de tâches sur le fichier de métadonnées réalisé: validation du document XML, génération d'une version pour la présentation de ces métadonnées (sous format HTML, graphique PNG, transformation XSLT avec un feuille de style OLAC), sauvegarde des métadonnées.

Figure 21: Copie d'écran de l'onglet "RESULT" du MKM.
-EXPORT : permet d’exporter le fichier de métadonnées au format HTML (plusieurs types de sortie peuvent être spécifiés).

Figure 22: Sortie HTML numéro 1 possible avec l’onglet "EXPORT" du MKM.

Figure 23: Sortie HTML numéro 2 possible avec l’onglet "EXPORT" du MKM.
Ainsi, l’outil MakeMetadata permet aux utilisateurs linguistes de créer des fichiers de métadonnées associés à leurs corpus de manière relativement simple. Il faut noter qu’un fichier de métadonnées devra être créé pour chaque fichier appartenant au corpus. Il faudra donc un fichier de métadonnées pour la ressource sonore et un autre fichier pour l’annotation textuelle.
b) Catalogue du PI-ED268

Afin de connaître les ressources disponibles dans la base de données, un moteur de recherche pour le catalogue de métadonnées du Projet Innovant PI-ED268 a été mis en place pour faciliter les requêtes.⁴²

Ce moteur permet à la fois de consulter les ressources disponibles en fonction d’un ensemble de métadonnées issues du Dublin Core, et d’effectuer une recherche sur le contenu de ces métadonnées.

L’interface du moteur permet de télécharger le catalogue, et de spécifier des requêtes à l’aide de boîtes de dialogues. Les requêtes sont spécifiées en sélectionnant des métadonnées du Dublin Core dans une liste déroulante, en saisissant directement une requête XPath, ou en saisissant directement une expression à chercher dans le contenu des métadonnées. Une fois la recherche effectuée, les résultats s’affichent en dessous des liens du projet.

Une fois démarrée, la page du moteur de recherche du catalogue se présente telle qu’illustre sur la figure 24.

Figure 24: Moteur de recherche du catalogue du PI-ED268.

Dans l’exemple suivant, en sélectionnant l’élément `dc:description` dans la liste déroulante, on obtient la totalité des descriptions contenues dans le catalogue, comme l’illustre les figures 25 et 26 qui en montre un extrait.

Figure 25: Exemple des métadonnées disponibles dans le moteur de recherche du catalogue du PI-ED268.

Figure 26: Résultat de la recherche pour `dc:description`.
En revanche, en précisant dans la fenêtre de droite que l’expression recherchée est « bise » à l’intérieur du contenu des métadonnées (l’élément dc:description ici), alors seules les ressources répondant à cette requête apparaîtront dans les résultats. La requête peut être réalisée en spécifiant la ou les métadonnées dans lesquelles la recherche doit être effectuée, ou bien en considérant l’ensemble des métadonnées.

Figure 27: Exemple de recherche en spécifiant une recherche de l’expression "bise" au sein des métadonnées "dc:description".

39

Figure 28: Exemple de recherche en spécifiant uniquement une recherche de l’expression "phonèmes" sur l’ensemble des métadonnées.
c) Bilan sur le PI-ED268

Le projet PI-ED268 ne s’adresse dans un premier temps qu’aux membres de l’Ecole Doctorale 268 (Langages et Langues) et ne permet que la description de corpus oraux. Une fois que ce prototype aura fait ses preuves, il sera intéressant de partager cette expérience et d’enrichir la base de données avec d’autres groupes de recherche et d’autres communautés de linguistique. Il sera alors souhaitable de pouvoir élargir la base aux corpus écrits.

Un outil tel que MakeMetaData est une première étape vers la simplification de la normalisation des corpus, car il permet à des utilisateurs linguistes (non informaticiens) de décrire aisément leurs ressources de façon normalisée en utilisant l’ensemble des champs proposés par le DC et OLAC. Ce type d’outil offrant la combinaison de la facilité d’emploi avec de grandes possibilités d’annotation nous semble indispensable pour inciter des linguistes à adopter une démarche de normalisation de leurs corpus.
Conclusion

Dans cette partie, j’ai souligné l’importance de la normalisation des ressources linguistiques orales. Cette normalisation est en effet particulièrement cruciale pour les communautés de recherche concernées, puisqu’elle permet non seulement la pérennisation des ressources, mais également leur diffusion et donc leur réutilisation. Il est donc important que les chercheurs mesurent l’importance de changer leurs méthodes de travail en amont pour prendre en compte cette normalisation, ce qui leur permettra de faire partager leurs propres ressources et de bénéficier de celles des autres.

J’ai également montré le rôle central que peuvent jouer les métadonnées dans une démarche de normalisation de ressources. Des normes internationales reconnues existent désormais, et il est important qu’elles soient connues et utilisées par les chercheurs dans les différentes disciplines de la linguistique. Néanmoins, cela ne sera possible que si des outils faciles d’emploi et orientés vers les préoccupations des linguistes sont mis à disposition. On assiste heureusement depuis quelques temps à des rencontres d’acteurs de la langue visant à diffuser l’information, et à sensibiliser à l’usage de certaines normes et outils. Des projets tels que ceux initiés au LACITO ou à l’ED-268 essaient d’allier les nombreuses possibilités d’annotation et la facilité d’utilisation des logiciels produits.

Après ce constat, il devient donc possible de se tourner vers une utilisation concrète de ces normes et de ces outils pour la constitution de ressources linguistiques réutilisables. Dans la deuxième partie de ce mémoire, nous présentons un travail sur l’annotation des disfluences présentes dans des transcriptions orales de parole spontanée prenant en compte les caractéristiques de la normalisation que nous venons de présenter.
Bibliographie


Burnard Lou 1996, TEI Lite : An Introduction to Text Encoding for Interchange, article paru dans le Cahier GUTenberg, numéro 24, juin 1996, pp 23-151


Annexe
Mode d’emploi MakeMETADATA

Propositions de normalisation pour une base de corpus multimédia à l’ED268.
Bénard F. et C. Gendrot 2005, RJC 2005
