Gestion de corpus oraux annotés : Méthodes et outils

Jacobson Michel

Laboratoire de Langues et Civilisations à Traditions Orales
CNRS , 7 rue Guy Môquet, Bât. 23 – 94801 Villejuif Cedex, France
Tél.: ++33 (0)1 49 58 37 78 - Fax: ++33 (0)1 49 58 37796
Mél: jacobson@idf.ext.jussieu.fr - http://lacito.vjf.cnrs.fr /

ABSTRACT

In linguistics or in phonetics, the speech corpora comprise, in general, two kinds of resources: speech recordings and annotations. Management tools of such corpora must be able to manage these two kinds of resources. We'll present here a survey of tools and formalisms used in the creation of such corpora. We'll also present some criteria allowing us to make a choice between them. We will conclude by the presentation of a model connecting two management tools: a software which makes phonetics analysis and a software which makes requests on a textual linguistic annotation. We will illustrate the use of this model in an example showing how to enrich an annotation by the automatic addition of values computed on the speech signal.

1. INTRODUCTION

Les corpus oraux sont en général composés de deux types principaux de ressources, que sont les enregistrements et les annotations de ces enregistrements. On distingue aussi, traditionnellement, deux types d'annotation : a) les annotations qui portent directement sur l'analyse du signal proprement dît, tels que les transcriptions ou les traductions ; b) les annotations qui portent principalement sur l'analyse de la situation d'enregistrement (lieu, date, etc.) et que l'on qualifie de métadonnées.

Les enregistrements de parole qu'ils soient de nature audio (ou vidéo, mais nous n'en parlerons pas ici) ont donné lieu en informatique à de nombreux formats de codage (PCM, ADPCM, A-Law, etc.) ainsi qu'à de nombreux formats de fichiers (RIFF, AIF, SND, etc.). Aujourd'hui, les développements s'orientent en grande partie vers la mise au point de systèmes de diffusion (comme le « streaming »), ainsi que vers la définition d'algorithmes de compression tels que les normes Mpeg ou Ogg Vorbis.

Les systèmes d'annotation du signal ont donné lieu, eux aussi, à de nombreuses propositions de codage et de formats de fichier. On peut distinguer deux grandes familles de systèmes génériques de structuration de l'information, très largement utilisés, que sont les bases de données et les langages de balisage de textes.

Les outils de gestion de corpus oraux doivent pouvoir gérer ces deux ressources (enregistrements et annotations). Sous le terme de « gestion », nous regroupons tout un ensemble de fonctionnalités allant de l'édition, (création et modifications des ressources), à la consultation qualifiée de multimédia ou d'hypermédia (Nelson [6]), en passant par l'interrogation (les requêtes formulées pouvant porter autant sur l'annotation que sur l'enregistrement ou que sur les deux à la fois).

2. LES OUTILS EXISTANTS

Les outils et formalismes dédiés ou pouvant être utiles à l'annotation de corpus oraux se sont développés assez rapidement ces dernières années, reflétant en cela les capacités croissantes du Web et de l'Internet à supporter ce type de données[1].

2.1. Les formalismes

Parmi un certain nombre de formalismes récemment définis, à un niveau très générique, Unicode pour le codage des caractères, et XML pour la structuration de l'information sont devenus des standards. À un niveau plus spécifique des recommandations comme celles de la Text Encoding Initiative[2] (TEI) du Corpus Encoding Standard[3] (CES) ont massivement été suivies. Nous présentons brièvement ci-après quelques uns de ces formalismes sans soucis d'exhaustivité, afin d'illustrer la diversité des domaines auxquels la constitution des corpus oraux est confrontée.

Unicode[4] est un code caractère ayant pour vocation un codage universel pour tous les systèmes d'écriture du monde. Il est apparu dans sa première version en 1990. Aujourd'hui dans sa version 4.0 il comporte quelques 96382 caractères et a été assez largement adopté par les constructeurs et les organisations y compris par le Web. Ce code a été normalisé au sein de l'ISO sous le nom ISO/IEC 10646. Entre autres avantages pour les linguistes, il permet le codage de documents multilingues, d'utiliser de grands ensembles de caractères comme le

[1] En effet, les données audio/vidéo sont classiquement volumineuses et demandent des débits assez élevés.

[2] http://www.tei-c.org/

[3] http://www.cs.vassar.edu/CES/

[4] http://www.unicode.org

chinois, ou de mélanger les sens d'écriture. C'est aussi la première fois que l'alphabet phonétique international est normalisé, ce qui relègue au passé pléthore de bricolages maisons de polices de caractères plus incompatibles les unes avec les autres, ainsi que des propositions *ad hoc* telle que SAMPA.

XML[5] (Extensible Markup Language) est un nouveau formalisme qui a pris la suite de son ancêtre SGML pour le codage de documents structurés. XML associé à Unicode est aussi devenu un standard pour l'échange de documents entre outils et plates-formes. XML n'est pas seulement un langage de balisage de texte, il s'agit de toute une famille de technologies, telles que RDF pour la définition de métadonnées, XSL pour la définition de feuilles de styles, Xlink pour la définition des liens, Xquery comme langage de requêtes, etc.

Nous listons ci-dessous quelques exemples d'applications de XML à des domaines particulièrement intéressants pour la gestion de corpus oraux :

TEI (Text Encoding Initiative) est une recommandation pour le codage de documents en sciences humaines avec notamment un chapitre sur les corpus oraux. A l'origine décrite en termes SGML, cette recommandation l'est maintenant aussi en XML.

SMIL[6] (Synchronized Multimedia Integration Language) est une application de XML pour le codage d'animations multimédia mélangeant ressources textuelles, iconographiques, audio et vidéo dans un scénario de présentation.

OLAC[7] (Open Language Archives Community) est une organisation de détenteurs de ressources linguistiques qui partagent la définition d'un ensemble de descripteurs de ressources (métadonnées basées sur du 'Dublin-Core'[8] enrichi et re-spécifié pour le domaine des ressources linguistiques), ainsi que sur une manière d'échanger ces informations en utilisant le protocole défini par l'OAI[9] (Open Archives Initiative).

2.1. Les outils

Pour la création et la maintenance des annotations, des outils génériques tels que les Systèmes de gestion de bases de données relationnelles (SGBDR) sont utilisés depuis assez longtemps et des interfaces avec le Web ont été mises au point pour l'interrogation et la diffusion des données. Plus récemment les développements autour du formalisme XML ont donné naissance à des outils génériques équivalents comme des éditeurs XML intégrant des technologies associées (XSL, XML-Schemas, etc.) et remplissant les mêmes fonctions de création et de maintenance pour des annotations. Des environnements de diffusion de données XML par le Web ont aussi fait leur apparition avec des outils comme Cocoon[10] ou Tomcat[11].

Des outils plus spécifiques ont été développés notamment pour aider à lier les deux types de ressources : audio/vidéo et annotations. Ces logiciels permettent, suivant le point de vue où l'on se place, soit d'ancrer temporellemment des annotations, soit d'annoter des moments du signal. Certains logiciels utilisent le formalisme XML pour coder l'annotation et ajouter à celle-ci des éléments ou attributs marquant des événements temporels de commencement, de fin ou de durées. Nous pouvons citer dans cette famille des outils comme soundIndex décrit dans Jacobson [5], Transcriber[12], ou ELAN du projet DOBES[13] sur les langues en danger, etc.

D'autres outils utilisent soit un formalisme propriétaire indépendant de l'enregistrement, soit profitent des possibilités d'un format de fichier particulier comme RIFF ou AIF pour y placer des annotations. Dans cette famille nous pouvons citer Praat[14] (avec son format de fichier d'annotation TextGrid) ou Audacity (avec son système de fichiers labels), mais aussi la plupart des éditeurs de son classiques (SoundForge, GoldWave, etc.)

Une autre catégorie d'outils facilitant le travail d'annotation de corpus oraux est apportée par des éditeurs spécialisés. Il existe, par exemple, des outils facilitant la saisie de textes interlinéaires et la saisie des gloses. La SIL[15] a depuis longtemps diffusé assez largement des outils de ce type (IT, Shoebox, LinguaLinks). Plus récemment un outil ITE (Interlineair Text Editor) a été développé au sein du LACITO afin de remplir ces deux fonctions tout en utilisant le formalisme XML.

3. LES CRITÈRES DE CHOIX

Avant de définir des critères de choix pour les outils et formalismes, il convient de définir les besoins que l'on souhaite couvrir pour la gestion des corpus.

3.1. L'analyse des besoins

La nature des corpus oraux, nous oriente vers deux types de préoccupations : la gestion des enregistrements qui sont de natures audio et temporelle et la gestion des

[5] http://www.w3c.org/TR/2000/REC-xml-20001006/

[6] http://www.w3c.org/TR/2001/REC-smil20-20010807/

[7] http://www.language-archives.org/

[8] http://dublincore.org/

[9] http://www.openarchives.org/

[10] http://cocoon.apache.org/

[11] http://jakarta.apache.org/tomcat/

[12] http://www.etca.fr/CTA/gip/Projets/Transcriber/

[13] http://www.mpi.nl/DOBES/

[14] http://www.fon.hum.uva.nl/praat/

[15] http://www.sil.org

annotations qui sont de natures textuelles, linéaires et hiérarchiques.

La gestion des enregistrements audio nécessite, suivant les cas :

- des outils de numérisation pour convertir les données analogiques en numérique ;
- des outils de restitution du signal comportant les fonctions classiques de « playback » ;
- des outils d'édition comportant les fonctions classiques de « presse papier », et des fonctions de transformations (filtrage, formatage, etc.)
- des outils d'analyse comportant des fonctions de mesure phonétique et de visualisation ;
- des outils d'extraction comportant des fonctions d'adressage pour appliquer toutes les opérations précédentes sur des parties de signal.

La gestion des annotations requière elle aussi un certain nombre d'outils :

- des outils d'édition pour créer et modifier les annotations ;
- des outils de visualisation pour consulter les annotations ;
- des outils d'interrogation pour formuler des requêtes sur les annotations ;

En plus de ces deux aspects, nous avons aussi besoin d'outils permettant de spécifier les liens entre les deux types de ressources (relations annotation / objet annoté, et liens temporels). Enfin, nous avons besoin que les outils de gestion communiquent les uns avec les autres afin, par exemple, que des modifications apportées à un type de ressource puisse modifier l'autre en conséquence.

3.1. Les enjeux stratégiques

Le choix des outils et des formalismes adéquats pour satisfaire les besoins précédemment listés, seront guidés par quelques grands principes généraux.

Les formalismes choisis doivent :

- avoir un pouvoir d'expression le plus grand possible, afin de pouvoir exprimer le plus de choses possibles ;
- être libres, publics, ouverts afin d'être indépendants et éventuellement de pouvoir influer sur la définition du format lui-même ;
- être le plus normalisés possible, ceci afin de garantir la pérennité de leur interprétation et donc la conservation à long terme des données. De plus l'échange à grande échelle des données n'est possible que si celle-ci sont normalisées ;
- à défaut d'être normalisés l'utilisation de formalismes standardisés et largement acceptés dans la communauté est une garantie permettant, au minimum, la comparaison des corpus.

Le choix des outils, lui aussi devra être guidé par le même type de préoccupations stratégiques, en ajoutant pour ceux-là des préoccupations économiques, ergonomiques et juridiques.

4. ARCHITECTURE DES TRAITEMENTS

4.1. Description des besoins

La tâche que l'on souhaite effectuer consiste à enrichir l'annotation d'un document XML avec les valeurs des trois formants (F1, F2 et F3) des voyelles qu'il comporte. Nous retiendrons comme mesure des formants la valeur ponctuelle mesurée au centre de la plage temporelle déclarée explicitement dans le document pour les segments concernés.

4.2. Description des choix

Pour effectuer cette tâche, nous avons fait un certain nombre de choix, guidés par les critères vus précédemment : XML et Unicode ont été choisis pour le codage des annotations. XSLT a été choisi comme langage de requête. L'outil choisi pour exécuter les requêtes est Xalan (processeur XSLT). Enfin les calculs sur le signal sont faits avec le logiciel Praat. Il été nécessaire d'avoir recours à deux outils distincts car, à notre connaissance, il n'existe pas d'outils permettant de faire porter des requêtes à la fois sur des données textuelles structurées et sur de données binaires telles que des données audio.

Alors que le choix de XML et XSLT se justifie entièrement par les critères que nous avons définis plus haut, le choix de Praat provient d'un mélange de raisons théoriques, pragmatique et opportunistes : l'utilisation du code source de Praat est soumise à une licence libre (GNU General Public Licence), ce logiciel est assez largement utilisé dans la communauté des phonéticiens, il possède un langage de script et une interface en ligne de commande, enfin il existe des implémentations sur de nombreuses plates-formes.

La distribution des tâches sur deux types d'outils nous a conduit à développer un logiciel permettant d'établir une interopérabilité entre eux. Ce logiciel d'interface à été développé en Java. Il permet au processeur de styles (Xalan) d'appeler le logiciel Praat pour effectuer des mesures ou des tests sur le signal et d'utiliser le résultat de ceux-ci dans le cadre des propres opérations.

4.3. Organisation des traitements

Mesure sur le signal

La mesure des valeurs moyennes des formants se fait par un script Praat. Ce script contient quatre paramètres qui sont : le nom du fichier audio, les temps de début et de fin de la plage à examiner et le numéro du formant.

Enrichissement de l'annotation

Le fichier d'annotation est un fichier XML qui comporte une référence à l'emplacement du fichier d'enregistrement audio, suivie par une liste de syllabes isolées. Pour chacune de ces syllabes, le noyau vovalique est transcrit et ancré temporellement.

```
<noyau start="0.154" end="0.238">a</noyau>
```

Figure 1 : Annotations d'un noyau syllabique avant la mesure des formants.

La requête permettant d'enrichir le corpus en ajoutant les valeurs des formants des noyaux, est exprimée dans une feuille de style. Celle-ci comporte une règle qui en substance dit que lorsqu'une voyelle est rencontrés lors du parcours de l'arborescence XML, les valeurs calculées de ses formants sont ajoutés sous forme d'attributs. Le résultat devrait donc, après calcul, être de la forme :

```
<noyau F1="238" F2="838" F3="1038"
        start="0.154" end="0.238">a</noyau>
```

Figure 2 : Annotations d'un noyau syllabique après la mesure des formants.

La requête qui demande le calcul des formants est codée dans une syntaxe (Xpath). Cette requête demande en substance à appeler le logiciel d'interface en lui passant les quatre paramètres dont le script Praat a besoin.

Après exécution de la tâche, le processeur de style récupère le résultat du calcul puis recopie cette valeur dans l'annotation sous la forme d'un attribut portant le nom du formant.

```
<xsl:attribute name="F1"
select="java:maquette.get($cmdLine, $filename, 1
        {@start},{@end})"/>
</noyau>
```

Figure 3 : requête en XSLT pour calculer le formant F1.

4. PERSPECTIVES

Le logiciel d'interface créé peut être utilisé à peu près de la même manière pour tester une hypothèse phonétique que pour faire une mesure. C'est le script Praat, qui suivant la demande délivrera un résultat que le processeur de style devra interpréter tantôt comme une valeur numérique, tantôt comme une valeur booléenne, tantôt enfin comme une simple chaîne de caractères.

Il est possible aussi de récupérer les valeurs calculées pour les tester à l'intérieur même de la feuille de styles. Par exemple, pour établir un classement ou une typologie. On peut aussi utiliser les valeurs de mesure pour réorganiser la liste des syllabes par ordre croissant de F1 puis secondairement par ordre croissant de F2, puis enfin de F3.

L'absence de formalisme et d'outils logiciel permettant de répondre simplement et directement à notre problème illustre bien le cloisonnement qui existe entre les développements faits dans le domaine du traitement des informations textuelles et ceux faits dans celui des données audio ou vidéo. Les perspectives à court et moyen terme, semblent être comme nous l'avons montré, l'exploitation de l'interopérativité facilité par l'ensemble des critères que nous avons évoqué plus haut, à savoir l'utilisation de formats libres, ouverts et standardisés et d'éviter le plus possible les formats propriétaires. Un progrès notoire pourrait être apporté par la définition d'une interface de programmation (API) ou d'une bibliothèque de fonctions standardisée pour exprimer des requêtes sur du signal, comme il en existe pour la navigation dans un document XML.

BIBLIOGRAPHIE

[1] P. Andries. Entretien avec Ken Whistler, directeur technique du consortium Unicode. *Document numérique*. Vol 6, n° 3-4, 2002.

[2] S. Bird et G. Simon. Seven dimensions of portability for language documentation and description. *Language*. Vol 79, n° 3, 2003.

[3] T. Bray and Co. (Eds.). (2000). *Extensible Markup Language (XML) 1.0 (Second Edition)*. W3C Recommendation, 6 octobre 2000 (http://www.w3.org/TR/REC-xml).

[4] M. Jacobson. Les outils modernes pour la transcription de corpus de parole. *Parole*. Vol 22, 23, 24, 2002.

[5] M. Jacobson, B Michailovsky and J. B. Lowe. Linguistic documents synchronizing sound and text dans le numéro spécial « Speech Annotation and Corpus Tools » de *Speech Communication*, n°33, 2001.

[6] T. H. Nelson. *Computer Lib/Dream machines*. Mindful Press. 1974.