Encodage SGML de corpus: application à l'étude d'un débat parlementaire

Serge Heiden

Encodage SGML de corpus
Application à l'étude d'un débat parlementaire

Serge Heiden
Analyses de corpus, usages et traitements
UMR8503

1 Introduction

Dans certains domaines de la science, telles la physique ou la psychologie expérimentale, on admet que l'outil d'observation puisse influencer la donnée observable. L'analyse de corpus informatisée, plus particulièrement la lexicométrie, n'échappe pas à ce phénomène, ce qui peut poser un problème de méthode relativement important. Faut-il systématiquement adapter un texte à l'outil qui permettra son traitement automatique pour assister son analyse ? Quels sont les risques à ne pas le faire ? Quelles stratégies adopter ? Y a-t-il un format de stockage ou de représentation du texte général qui permette de transformer aisément le texte pour tel ou tel outil d'analyse ? Etant donné la multitude d'outils d'analyse et de formats de stockage des données, il importe de se questionner sur la démarche d'encodage d'informations dans un texte à des fins d'analyses lexicométriques et de proposer des éléments de réponse aux questions posées systématiquement par l'étude du discours à l'aide d'outils traitant des corpus textuels sous forme électronique.

Dans cet article, après une analyse de la démarche et des enjeux de l'encodage de corpus, nous proposons un format et des outils d'encodage qui satisfont aux contraintes de la méthode.

Cet article est organisé comme suit : dans la section 2, nous définissons trois types d'informations fondamentaux pour les traitements lexicométriques. Dans la section 3, nous exposons une méthode normalisée d'encodage mise au point pour la base de textes de notre laboratoire en liaison avec les différents outils de traitement afférents. Enfin, dans la section 4, nous illustrons la démarche proposée en l'appliquant à un corpus de débats parlementaires pour montrer quelques exemples d'exploitation.

2 Définitions et enjeux

La mise en format d'un corpus est l'étape préalable à toute analyse lexicométrique. Elle conditionne l'observable du discours. Un des avantages cruciaux de l'analyse lexicométrique statistique est sa rapidité à fournir des éléments d'interprétation pour un corpus. Quels types d'informations doit-on encoder dans le corpus pour obtenir ces supports d'interprétation ?

Nous nous intéresserons, dans cet exposé, à trois types d'informations fondamentaux :

- les unités lexicales ;
- la délimitation de parties dans un texte ;
- les références éditoriales permettant de situer précisément l'occurrence d'un événement textuel dans le corpus.

Précisons que dans cet article, nous nous intéressons à une notion élargie des outils d'analyse lexicométrique. Les outils visés procèdent aussi bien à des mesures statistiques sur divers événements textuels qu'à la navigation dans le corpus, comme le fait par exemple un concordanceur, ce qui permet l'aller-retour permanent entre les résultats de synthèse et le corps du texte.

Nous nommerons l'opération qui consiste à créer une représentation de certaines informations à l'intérieur du texte lui-même une opération d'*encodage*, l'interprétation de cette représentation à des fins d'exploitation étant nommée *décodage*. Nous ne nous intéresserons qu'aux techniques où l'encodage s'exprime à l'aide de caractères au fil du texte de manière

explicite. Egalement, nous nommerons *balise* toute marque exprimant un encodage d'informations particulières.

Unités lexicales

La forme et le contenu d'un texte sont habituellement représentés par les unités lexicales. Dans la suite de cet article, nous les nommerons unités graphiques en référence à leur identification fréquente par la délimitation de caractères séparateurs. La forme graphique est une unité de base dans les traitements (pour le calcul de vocabulaires, de concordances, etc.) Elle est fondamentalement liée au système d'encodage des caractères, comme par exemple le code ASCII, qui lui-même dépend souvent du système d'exploitation de l'outil (notamment en ce qui concerne les caractères diacrisés). La délimitation des unités est toujours implicite. Par exemple, dans un système comme Frantext (1),

(1) Base textuelle conçue à Nancy par Jacques Dendien à partir des textes ayant servi à l'élaboration du Trésor de la Langue Française.

on emploiera le caractère espace « » pour distinguer l'unité « C' » de l'unité « est » dans un texte. On y considère en effet implicitement le caractère « ' » comme constituant de mot, comme dans « aujourd'hui » par exemple. Il y a parfois des délimitations d'unités graphiques qui ne reposent pas sur l'encodage de caractères seuls. Ainsi, pour délimiter et traiter le mot composé « carte de séjour » en tant qu'unité, un mécanisme supplémentaire au simple repérage de caractères dits « délimiteurs » (comme l'espace) doit être mis en œuvre.

Partitionnement

L'analyse lexicométrique repose souvent sur l'identification de parties dans un corpus (par exemple chaque œuvre du corpus, chaque période dans une partition diachronique, etc.). Ainsi, dans un système d'analyse contrastive comme Hyperbase (2),

(2) Le logiciel Hyperbase est un logiciel d'analyse lexicométrique permettant de traiter n'importe quel texte et mettant en œuvre une interface de type hypertextuelle. Il a été conçu par Etienne Brunet.

le partitionnement est délimité par des en-têtes de la forme *&&& ... &&&*. Le texte situé entre ces deux « balises » dénomme chaque partie que l'on va traiter et comparer avec les autres.

Dans un autre système comme celui du codage Machinal(4),

(4) Le format Machinal a été conçu comme support de représentation de l'encodage des corpus de notre laboratoire à partir de 1985. Il est défini par un document d'une soixantaine de pages (ref). Ce format est utilisé, avec divers aménagements, par les logiciels Lexico (André Salem), Pistes (Pierre Muller) et Saint-Chef (Majid Sékhraoui).

toutes les balises sont encadrées de chevrons « < » et « > ». La portée d'une balise (c'est-à-dire son domaine d'application) correspond alors à l'ensemble du texte se trouvant à sa suite. Les balises comprennent un nom séparé d'une valeur par un signe égal « = ». Pour créer une partition, il suffit donc de faire se succéder des balises de noms identiques mais avec une valeur différente. Certains noms sont réservés et ont une interprétation prédéfinie. Les noms comprenant des chiffres sont utilisés quand les balises prédéfinies ne suffisent pas à la démarche du chercheur. Ils ont alors une interprétation propre (et limitée) à la recherche. Un corpus peut encoder potentiellement plusieurs partitions, et une partie n'est pas toujours composée d'un seul bloc de texte contigu. C'est le cas, par exemple, de l'ensemble des répliques d'un acteur dans une pièce de théâtre.

Enfin, d'autres systèmes comme Alceste (3)

(3) Ce logiciel implémente la méthodologie « Alceste » d'analyse statistique des distributions d'unités lexicales dans les unités de contexte élémentaires d'un corpus. Il a été conçu par Max Reinert.

proposent une approche a posteriori du partitionnement. L'espace minimal de rencontre des unités lexicales est défini par la notion d'unité de contexte élémentaire qui correspond, en première approximation, à la phrase mais peut aussi s'étendre à des unités plus larges suivant la qualité de la classification obtenue à partir des unités graphiques qu'elles contiennent. Les unités de contexte élémentaires peuvent être encodées a priori dans le corpus ou bien déterminées implicitement par un algorithme analysant certains caractères de ponctuation forte.

Références

L'interprétation de synthèses de recherche d'événements textuels (type concordance) et de navigations dans un corpus (du type hypertexte par exemple) reposent sur l'identification précise du lieu d'apparition de ces événements. Par exemple, dans le système Hyperbase, les références de numéros de pages sont encodées dans le corpus à la suite du caractère *$* comme dans *$2* pour commencer la page 2 d'un texte. Ces références ne sont pas toujours d'ordre éditorial (titre de l'œuvre, numéro de page, de section, etc.) mais peuvent dépendre de marqueurs propres à la démarche d'analyse, comme dans le cas d'un corpus composé d'extraits d'énoncés par exemple. Dans un système comme Alceste, les références sont encodées à l'aide du caractère spécial * préfixant toute information de référence. Tous les caractères suivant l'astérisque jusqu'au caractère séparateur suivant forment une référence (ou mot étoilé). Dans la méthodologie Alceste, la classification des unités de contexte élémentaire par rapport à leur vocabulaire permet a posteriori de considérer certaines de ces références comme des données de partitionnement. Dans le format Machinal, les références sont encodées sous la même forme que les informations de partitionnement. Par exemple, la balise *<Epg=10>* désigne un début de page (et son numéro) dans l'édition de référence.

La qualité de l'identification des informations par les outils d'analyse dépend de la précision de l'effort d'encodage. Ainsi, toute analyse repose sur la qualité de l'identification des unités graphiques. Afin de bien repérer ces unités, il est parfois nécessaire d'adapter le texte. Deux opérations sont représentatives de ce type d'adaptation : d'une part, le passage en minuscules des majuscules de début de phrase est une opération fréquente pour homogénéiser les classes de formes du vocabulaire. Dans des textes dont la longueur moyenne des phrases oscille entre 10 et 20 mots, cette opération peut influencer les décomptes de formes se trouvant fréquemment en début de phrase. Par ailleurs, l'agglutination des formes pour le repérage de mots composés autorise la comparaison des unités lexicales correspondant aux figements du corpus et permet de distinguer dans les analyses les composants de figements de leur forme homographes qui sont indépendantes lexicalement. Par exemple, l'unité graphique un « pied d'égalité » ne sera pas confondue avec une partie du corps.

En ce qui concerne le partitionnement du corpus, une erreur de délimitation peut prendre des proportions importantes dans la mesure où elle peut concerner beaucoup d'événements textuels à la fois, et il n'est pas toujours aisé de vérifier sa conformité.

Les enjeux des erreurs de gestion des références sont du même ordre bien que leur utilisation documentaire soit souvent plus illustrative que comparative.

La précision de l'encodage d'informations dans un texte est coûteuse en temps et ceci est contradictoire avec certaines motivations à l'origine de l'utilisation d'outils lexicométriques :

la rapidité et la couverture de l'analyse. Par ailleurs, la généralisation d'outils de marquage morpho-syntaxiques, de lemmatiseurs… nous offre l'opportunité d'observer des événements textuels plus fins ou plus généraux que la simple surface du texte (de l'ordre des unités graphiques). Ces besoins d'encodage d'informations de plus en plus fin (partie du discours, lemme, traits sémantiques, …) imposent une approche méthodologique différente. En effet, le marquage fin des unités lexicales tend à systématiser l'encodage de la segmentation des unités graphiques.

Enfin, mentionnons que la quantité de textes disponibles, leur réutilisation fréquente dans les études, la facilité des échanges et l'évolution rapide des logiciels imposent une normalisation du support de représentation des textes, c'est-à-dire une normalisation de la relation codage/interprétation.

3 Une proposition de méthode

Notre meilleure réponse au compromis temps d'encodage/rapidité d'analyse est l'utilisation d'un format permettant un encodage rapide et progressif, dont le résultat est réutilisable. Ainsi, nous proposons l'usage d'un format d'encodage unique pour l'ensemble des informations à annoter dans le corpus, quel que soit leur niveau de granularité. En effet, nous proposons un format uniforme pour encoder : la délimitation et les propriétés du corpus, de chaque texte et de leur contenu (chapitre, section, paragraphe, phrase, unité graphique, caractère). Par exemple dans ce format, l'étiquetage morpho-syntaxique devient une propriété de l'élément d'encodage des unités graphiques. Par ailleurs, l'uniformité de l'encodage est un gage d'extensibilité pour l'évolution future des logiciels d'analyse.

Remarquons que nous plaçons sur le même plan la mise en conformité des unités graphiques suivant les objectifs du traitement et l'ajout d'informations de partitionnement, alors qu'habituellement, on associe plus facilement à cette dernière opération la notion d' « ajout » ou encore d' « encodage » d'informations en opposition au simple « toilettage » du texte de la première opération. Dans les deux cas, il s'agit d'expliciter à la machine les objets ou événements à traiter en fonction de la partie du logiciel concernée.

3.1 Une norme d'encodage : SGML

La première fonction d'un format est de permettre le codage de toutes les informations exploitées par les différents formats précédents : segmentation en unités graphiques, partition, référence. Nous proposons une application de la norme d'encodage de documents internationale ISO (International Standard Organisation) correspondant à l'encodage SGML (Standard Generalized Markup Language) pour réaliser ces différentes fonctions.
L'intérêt d'un encodage de type SGML réside fondamentalement dans les possibilités de partage et de réutilisabilité de l'effort d'encodage. De plus, la normalisation ISO de ce format universel d'échange de textes et de corpus repris par les standard d'encodage TEI (Text Encoding Initiative), EAGLES (Expert Advisory Group on Language Engineering Standards) et CES (Corpus Encoding Standard) est un gage de stabilité. SGML est par ailleurs déjà très répandu sur Internet sous la forme de la norme HTML (Hyper Text Markup Language) qui est elle-même une application SGML (sa DTD (Document Type Declaration) est celle des pages WWW (World Wide Web)). Mentionnons également que de nombreux éditeurs de texte SGML sont disponibles ainsi que des outils de vérification de conformité à une DTD particulière et de transcodage entre DTD.
Quelle forme prendra le codage des informations ? Les règles de base du codage SGML sont les suivantes :

- toute information méta-textuelle, des informations bibliographiques générales au texte à l'étiquetage fin des unités lexicales, en passant par l'encodage de la structure du document, est représentée par une balise nommée et écrite entre chevrons « <…> ». Il y a deux types de balises : les balises ponctuelles et les balises à contenu.
- une balise peut être ponctuelle, comme par exemple la balise «
» qui marque un saut de ligne dans le texte de référence.
- une balise peut porter sur une portion quelconque du document, auquel cas elle délimite sa portée (son contenu) à l'aide d'une balise ouvrante et d'une balise fermante de nom identique à la première mais préfixé par le caractère « / ». Par exemple, l'encodage : **<**I**>**passage de l' échelle**<**/I**>** peut s'interpréter comme la mise en italique d'une portion de texte.
  Dans le cas de balises à contenu, différents niveaux de balises peuvent alors s'imbriquer. Par exemple : <DIV><HEAD>La vie en rose</HEAD><PAGE>texte de la page 1 /…/ </DIV> /…/ où la balise <DIV> marque une division structurelle du document, la balise <HEAD> l'entête de cette division et la balise <PAGE> un saut de page à cet endroit dans le texte. On notera que la notion de passage à la ligne doit s'encoder explicitement (si elle est nécessaire à l'analyse).
- toute balise peut posséder plusieurs propriétés, en plus de son nom, sous la forme d'une relation nom_attribut=valeur_attribut. Par exemple, nous aurions pu non seulement encoder le saut de page dans l'exemple précédent mais aussi le numéro de la nouvelle page à l'aide d'une balise attribuée de la forme <PAGE N=10>. N étant un attribut numérique de la balise PAGE dont la valeur est ici 10 .
- un mécanisme d'abréviation permet d'inclure dans le texte à la fois les trois caractères spéciaux induits par le système d'encodage : « < » est encodé par « &lt ; », « > » par « &gt ; » et « & » par « &amp ; » (contrairement aux formats précédents qui empêchent l'usage littéral de certains caractères réservés à l'encodage), de condenser l'écriture et de traiter le cas des symboles spéciaux propres à toute langue : « &oelig ; » pour « œ », « &szlig ; » pour *β*, …
- enfin, pour un document donné, l'ensemble du système de balises et des attributs repose sur une déclaration dans un entête : c'est la Déclaration de Type de Document ou DTD. Le mécanisme des chevrons et des attributs de balise est universel ; la DTD quant à elle spécifie les noms de balises et d'attributs et la manière dont elles s'imbriquent dans le document.

3.2 Une application de cette norme : le format LML

Comment sont encodées les informations dans le texte ? Les partitions et les références sont codées naturellement par la conjonction des imbrications de balises et par les valeurs d'attributs que nous venons de voir. La segmentation en unités graphiques repose sur le même type d'algorithme que dans la segmentation classique mais pour certains cas spécifiques des balises de segmentation prennent priorité sur l'algorithme, par exemple dans « Ces <W>cartes de séjour</W> ne sont pas /…/ » les balises forceront la segmentation de « carte de séjour » en une seule unité lexicale bien que l'espace soit un délimiteur implicite.

Notre application LML (pour Lexico Markup Language) de la norme SGML correspond à la fois à une définition de structuration de balisage, la DTD LML, et à une bibliothèque d'outils de mise en œuvre (vérification, projection, extraction…). Le format LML tente de représenter le maximum d'informations méta-textuelles d'un corpus dans l'encodage. Il permet d'encoder les classes d'informations méta-textuelles suivantes :

- <B>, <I>, <FONT>… : la typographie d'origine (gras, italique, police, ...) ;

-
, <PAGE>… : la structuration éditoriale d'origine (pagination, saut de ligne, ...) ;
- <DIV>, <P>… : la structuration logique d'origine (chapitre, section, paragraphe, ...) ;
- <SP>, <ENONCE>… : une structuration logique propre à l'étude (locuteur, portion de texte, ...) ;
- <S>, <W>, <C> : le choix de la segmentation en phrases, formes et ponctuation ;
- <DIV TYPE=chapitre N=2>, <W CAT=verbe LEM=avoir>… : l'annotation de tous ces éléments avec des propriétés (comme le numéro d'un chapitre, la catégorie morpho-syntaxique d'une forme, son lemme, ...) ;
- <HEADER AUTHOR='Vivant Denon'… >… : les informations bibliographiques du document (nom de l'auteur, titre, date d'édition, nom de l'éditeur, origine de la numérisation, nom des codeurs, des catégoriseurs utilisés, des correcteurs, ...) ;
- <IMG>, <A> : les informations multimédia et hypertextuelles (images incluses, liens HTML, ...) ;
- © , ¥ … : le codage des caractères spéciaux (©, ¥, ß, ¿, ¶,½, ...).

Dans la section qui suit, nous présentons une mise en œuvre de l'encodage d'un corpus à l'aide de ce format et quelques exemples d'exploitation possibles.

4 Exposition de la méthode dans un exemple

4.1 Origine du corpus RESEDA et démarche

Les données d'origine du corpus RESEDA nous ont été fournies sous la forme de plusieurs fichiers informatiques de texte brut pour les débats de toutes les séances d'une journée à l'assemblée nationale, c'est-à-dire pour le contenu de plusieurs exemplaires du journal officiel.
Nous avons abordé l'analyse de ce texte selon une approche contrastive entre les partis politiques et entre certains locuteurs. L'effort d'encodage a donc porté sur l'association d'une personne et de son parti à chaque prise de parole dans le cadre des seuls débats portant sur la loi « Entrée et séjour des étrangers en France et droit d'asile ». L'usage de l'approche LML se justifie par la richesse des débats :
- 8 séances parmi 43 pendant 5 jours de débats ;
- 7028 prises de paroles sur l'ensemble des débats ;
- 323 locuteurs différents classés dans 6 partis politiques (SOC, RPR, …) et 2 catégories à part (président de séance, …) ;
- plus de 1578 commentaires ou didascalies à ignorer entre les prises de parole. Par exemple : *(Exclamations sur les bancs du groupe du Rassemblement pour la République)*.

4.2 Encodage

Afin de rendre ce corpus réutilisable, nous avons opté pour un encodage de l'ensemble des débats, incluant ceux qui ne concernaient pas l'étude, tout en affinant l'encodage et sa vérification aux seuls débats intéressants RESEDA. De même, certains locuteurs, comme le président de séance, ou certains commentaires qui ne sont pas pris en compte dans l'étude ont été délimités tout en restant présents dans le corpus encodé. Des procédures d'extraction nous permettent de n'appliquer les outils de mesure qu'aux seules unités intéressant l'étude.
De plus, le maintien de toutes les informations éditoriales dans la phase d'encodage nous permet d'obtenir une édition du corpus la plus fidèle possible aux documents d'origine (comme on pourra le vérifier à la figure 4).

Voici un extrait des débats que nous avons encodés :

```
ENTRÉE ET SÉJOUR DES ÉTRANGERS EN FRANCE ET DROIT D'ASILE

Discussion, après déclaration d'urgence,
d'un projet de loi
 M. le président. L'ordre du jour appelle la discussion, après déclaration
d'urgence, du projet de loi relatif à l'entrée et au séjour des étrangers
en France et au droit d'asile (n°s 327, 451, 483).
 Avant que la discussion ne s'engage, je ferai deux observations.
 Nous avons prévu un temps assez important pour ce débat puisque le vote
final interviendra mardi en huit. Je souhaite que nous disposions du temps
nécessaire pour examiner le texte et les amendements,...
 M. Jean-Louis Debré. Voulez-vous que nous déposions d'autres amendements ?
 M. le président. ... sans qu'il y ait pour autant d'obstruction.
 En second lieu, et surtout, comme il s'agit d'un sujet sensible, je
souhaite que le débat se déroule dans le climat de dignité et de respect
nécessaire.
 M. Rudy Salles. Ça dépend du Gouvernement !
/…/
```

Notre stratégie d'encodage repose sur un cycle associant à un travail de substitution de chaînes de caractères dans le texte brut du corpus un travail progressif d'exploitation de la structure et des propriétés SGML.
Dans la phase de substitution de chaînes de caractères, nous avons exploité certaines régularités de saisie, comme par exemple la structure récurrente « `M.` **`prénom nom, titre officiel*.`** » que l'on trouve systématiquement au début de chaque prise de parole et en début de ligne. Par exemple, nous avons substitué à la chaîne de caractères « `M. Jean-Louis Debré. Voulez-vous` » la chaîne « `<SP WHO="Jean-Louis Debré"><P> Voulez-vous` » où la balise `<SP>` (pour SPeaker) marque le début de la prise de parole, l'attribut `WHO` (qui) de cette balise encodant le nom du locuteur et la balise `<P>` (pour Paragraph) le début d'un paragraphe.
Après vérification et normalisation de la structure SGML du corpus, nous avons alors pu exploiter la structuration encodée dans la phase précédente. Par exemple, pour chaque prise de parole, nous avons utilisé une liste d'association de la forme nom_de_personne – nom_de_parti-politique (associant par exemple « Jean-Louis Debré » à « RPR »), pour projeter l'encodage du parti politique dans chaque élément SGML prise de parole `<SP>` sous la forme de la valeur d'un attribut `PARTI` en fonction de la valeur de son attribut `WHO`.
Voici le résultat de l'encodage de l'extrait :

```
<DIV LEVEL="2" TITLE="ENTRÉE ET SÉJOUR DES ÉTRANGERS EN FRANCE ET DROIT
D'ASILE">
<TITLE>ENTRÉE ET SÉJOUR DES ÉTRANGERS EN FRANCE ET DROIT D'ASILE</TITLE>

Discussion, après déclaration d'urgence, d'un projet de loi

<SP WHO="le président" PARTI="NA"><P> L'ordre du jour appelle la
discussion, après déclaration d'urgence, du projet de loi relatif à
l'entrée et au séjour des étrangers en France et au droit d'asile <REM>n°s
327, 451, 483</REM>.
<P>Avant que la discussion ne s'engage, je ferai deux observations.
<P><c V="Nous">nous</c> avons prévu un temps assez important pour ce débat
puisque le vote final interviendra mardi en huit. <c V="Je">je</c> souhaite
que nous disposions du temps nécessaire pour examiner le texte et les
amendements,...
</SP>
```

```
<SP WHO="Jean-Louis Debré" PARTI="RPR"><P> Voulez-vous que nous déposions
d'autres amendements ?
</SP>
<SP WHO="le président" PARTI="NA"><P> ... sans qu'il y ait pour autant
d'obstruction.
<P>En second lieu, et surtout, comme il s'agit d'un sujet sensible, je
souhaite que le débat se déroule dans le climat de dignité et de respect
nécessaire.
</SP>
<SP WHO="Rudy Salles" PARTI="UDF"><P> Ça dépend du Gouvernement !
</SP>
/…/
```

Dans cet exemple, la balise DIV délimite une discussion sur un projet de loi dans l'ensemble des débats (les divisions de premier niveau, dont l'attribut LEVEL prend la valeur 1, sont réservées aux séances et contiennent les divisions de niveau inférieur). Le titre de cette discussion a été encodé à la fois sous la forme d'un contenu de la balise TITLE et sous la forme de la valeur d'un attribut de la balise DIV pour permettre des extractions de divisions par valeurs d'attributs. Chaque prise de parole est délimitée par une balise SP dont l'attribut WHO encode le nom du locuteur et l'attribut PARTI le nom de son affiliation politique. On peut constater que les paragraphes (P) ne se referment pas, bien qu'ils aient un contenu. Leur fermeture est implicite et assurée par la DTD (les paragraphes ne pouvant pas contenir de paragraphe, l'ouverture de l'un provoque la fermeture du précédent). La succession de numéros « 327, 451… » a été encodée comme une didascalie (REM) par erreur. En effet, les 1578 didascalies du corpus ont été encodées en une seule opération de substitution de parenthésages. Manifestement cette mise entre parenthèse ne forme pas une didascalie. Cependant, dans la mesure où nous ignorons les prises de parole du président de séance dans nos analyses, cette scorie ne porte pas à conséquence. Enfin, signalons la présence de quelques adaptations d'unités graphiques du type majuscule de début de phrase (« Nous » a été transformé en « nous » à l'aide de la balise C). On pourra vérifier la possibilité de rétroconversion vers le texte brut d'origine en supprimant tout le marquage (qui est situé entre chevrons).

Après avoir délimité et partiellement enrichi notre corpus, nous avons alors pu en extraire diverses informations à l'aide des outils d'exploitation LML qui interprètent la structure SGML pour vérifier nos hypothèses.

4.3 Exploitation et résultats

L'exploitation du corpus a consisté à procéder aux extractions de texte limité aux seuls débats RESEDA et aux prises de parole. Pour cela, nous avons utilisé les outils LML en les appliquant au flux SGML du corpus RESEDA :

- normalisation du corpus balisé et transformation en un flux SGML (par l'outil `mknsg`) ;
- extraction de ce flux de n'importe quel élément SGML ayant certains attributs (cet outil nous a servi à l'extraction des seuls débats RESEDA et des seules prises de parole, c'est-à-dire au *partitionnement*) (outil `lmlget`) ;
- suppression du flux de n'importe quel élément SGML ayant certains attributs (cet outil nous a servi à la suppression des didascalies et de certains commentaires d'encodage au fil du corpus) (outil `lmlrm`) ;
- insertion dans le flux d'éléments SGML supplémentaires pour la construction des *références* lors de l'intégration du corpus dans les bases de l'outil d'analyse (le *Lexploreur*) (outil `lmlinsref`) ;
- segmentation en *unités graphiques* (outil `mach2tttxt`) ;

- transformation du format LML vers un autre format (le format HTML par exemple) (outil `lmlcost`).

Les possibilités d'extraction très fines (à la prise de parole près), et les possibilités de filtrage et de tri des résultats nous ont permis, dans un premier temps, de produire des graphes de synthèse des débats mettant à jour la notion d'interrupteur.
Pour l'analyse de cette notion (voir l'article infra de Simone Bonnafous et de Dominique Desmarchelier), nous avons d'abord composé une synthèse générale de la succession des prises de parole ayant eu lieu lors de la première séance des débats relatifs à la loi RESEDA (datée du 4 décembre 1997) (voir figure 1). Dans ce schéma, l'axe des abscisses représente l'ordre des prises de parole dans le temps, de la gauche vers la droite. L'axe des ordonnées rend compte du nombre de mots prononcés lors de chaque prise de parole. Chaque prise de parole est représentée par un segment vertical étiqueté à son sommet par les initiales du locuteur. Plus ce segment est haut, plus la prise de parole est longue. Le motif de hachurage du segment représente le parti politique du locuteur (voir la légende). Dans ce schéma, le « fouillis » des interrupteurs représenté par les petits segments et le recouvrement illisible des initiales a permis d'identifier rapidement les locuteurs principaux : Jean-Pierre Chevènement (J-PC), Gérard Gouzes (GG), Jean-Yves Le Déaut (J-YLD)…
Pour caractériser graphiquement cette notion d'interrupteur, nous avons construit un graphe représentant une synthèse de la participation de chaque locuteur aux débats (voir figure 2). Dans ce schéma, l'axe des abscisses représente le nombre de prises de parole du locuteur. L'axe des ordonnées représente le nombre total de mots qu'il a prononcé lors des débats. Le symbole marquant sa position (triangle, carré, étoile, losange, …) représente le parti politique du locuteur. La compression des deux dimensions (par le logarithme en base 10 des valeurs réelles) nous a permis de dégager un « nuage » et surtout un axe principal des locuteurs ayant eu un rôle d'interrupteur. La forme elliptique représentée dans le schéma et son axe de symétrie rendent compte de la zone de repérage des interrupteurs. L'intérêt de cette zone est de permettre l'évaluation rapide de la participation des locuteurs qui en sont exclus mais qui en sont proches, c'est-à-dire dont le rôle n'est pas interprétable facilement a priori. L'archétype de l'interrupteur est Christian Estrosi (CE) qui intervient 15 fois mais ne prononce jamais plus de 10 mots. Pour valider cette hypothèse d' « axe d'interruption », nous avons à nouveau utilisé les outils d'extractions LML pour composer un corpus de toutes les prises de parole de chaque locuteur. Ce corpus est ordonné selon le locuteur par le nombre de mots prononcés, puis par le nombre de prises de parole (voir figure 3). On constate alors qu'aux limites de l'ellipse des interrupteurs, les locuteurs ont eu un rôle mixte, à la fois d'interruption et occasionnellement de discours prolongé. On peut prendre le cas d'André Gerin qui est intervenu 8 fois en prononçant en moyenne 12 mots, mais qui est aussi intervenu une fois avec 232 mots. Dans tous les cas, les locuteurs hors de l'« ellipse d'interruption », y compris ceux ayant pris peu la parole (CT, Fco, Fd'A, AG), ont prononcé un discours programmé officiellement dans les débats.
Indépendamment de ces synthèses, le format LML nous a permis, de plus, de produire un Hypertexte des débats dont la mise en page est la plus proche possible du fac-similé original du journal officiel. Cependant, à la différence du J.O. cet Hypertexte permet à la fois la lecture du déroulement des sessions (figure 4) et celle du corpus des prises de parole de chaque locuteur (figure 5). Cet Hypertexte des débats est composé de 3 zones :
- A droite, chaque contenu d'une page du J.O. ou le corpus total des prises de parole d'un locuteur. A la différence de l'édition originale, dans une page de J.O. de cet Hypertexte, le marquage de chaque prise de parole par le parti politique du locuteur dans le corpus a été reproduit à la suite du nom et en couleur. Chaque prise de parole est, de plus, numérotée chronologiquement et son numéro renvoie, par un lien hypertextuel, à la prise de parole

correspondante dans le corpus du locuteur. Le corpus des prises de parole de chaque locuteur est composé de manière similaire en précisant en tête de chaque prise de parole le numéro de page du J.O en plus du numéro chronologique. Dans un corpus, le numéro de page est lui-même un lien Hypertexte vers la prise parole correspondante dans la page du J.O. Les pages de débats et les corpus de locuteurs se référencent donc mutuellement par des liens hypertextuels, permettant le va-et-vient entre les débats et l'argumentation de chaque intervenant.

- A gauche en haut, la liste des locuteurs triée alphabétiquement. Le nom de chaque locuteur est un lien hypertextuel vers le corpus de ses interventions ;
- A gauche en bas, la liste des numéros de page du J.O. Chaque numéro est un lien hypertextuel vers la page du J.O.

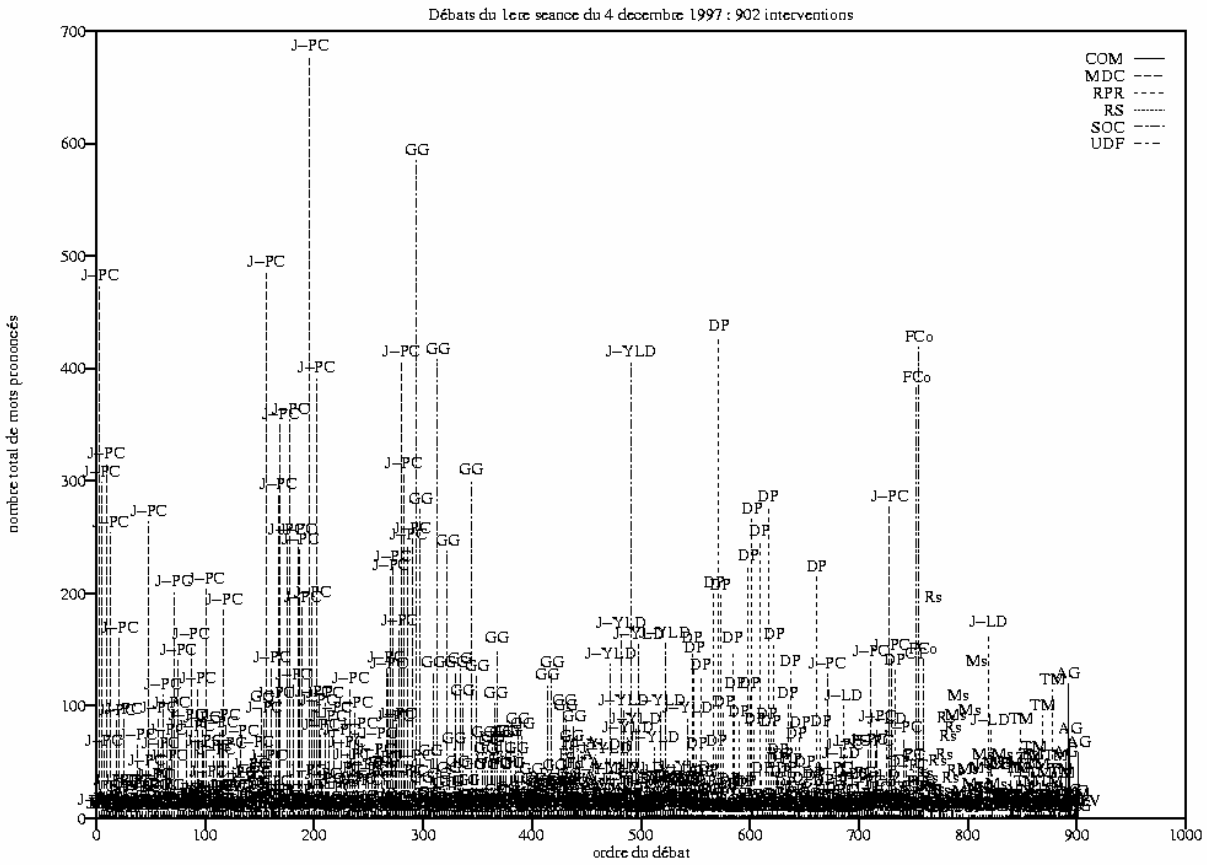

Figure 1.
Graphe de synthèse de la succession des prises de parole de la première séance des débats à l'assemblée nationale du 4 décembre 1997.

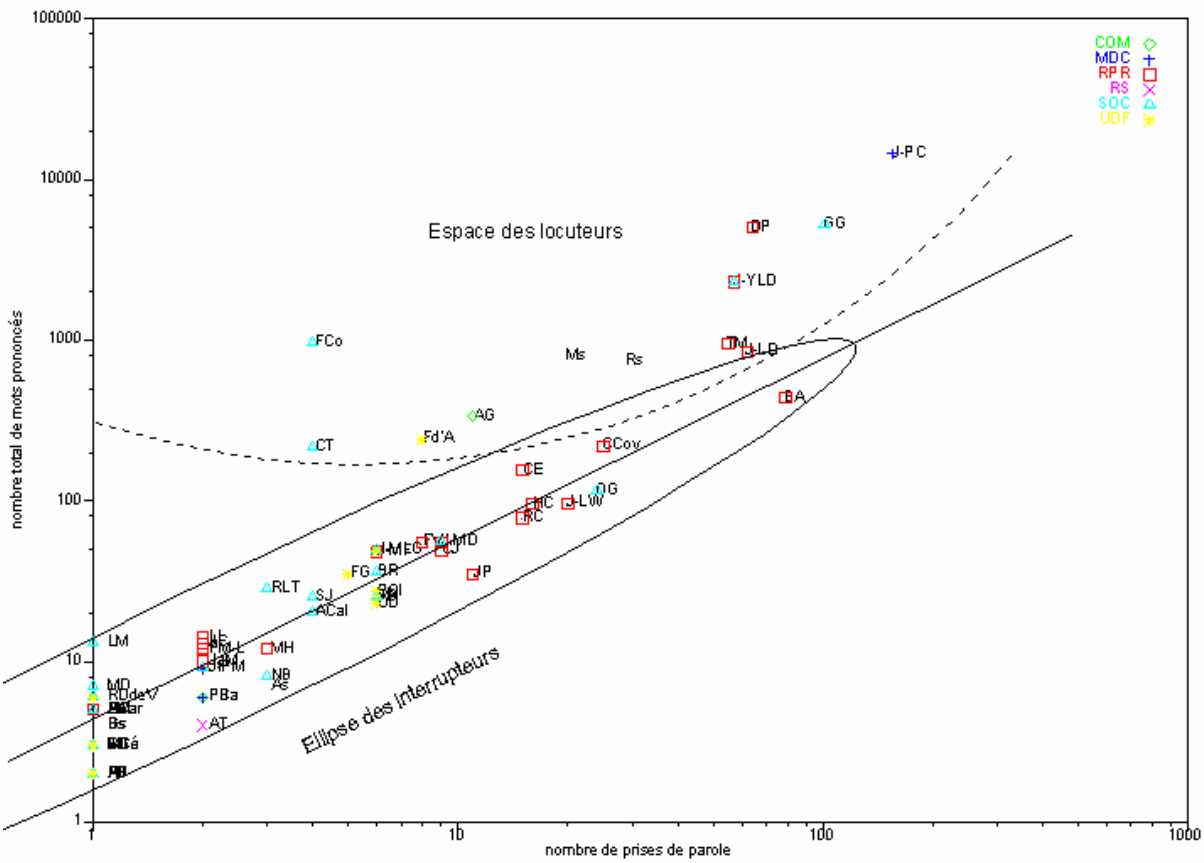

Figure 2.
Synthèse de l'ensemble des prises de parole de chaque locuteur de la première séance du 4 décembre 1997.

Figure 3.

Corpus des interrupteurs (le début) trié selon le locuteur par nombre de mots prononcés puis par nombre de prises de paroles (indiqués entre parenthèses et dans l'ordre inverse). Pour un locuteur donné chaque prise de parole est séparée de la suivante par un caractère « / ».

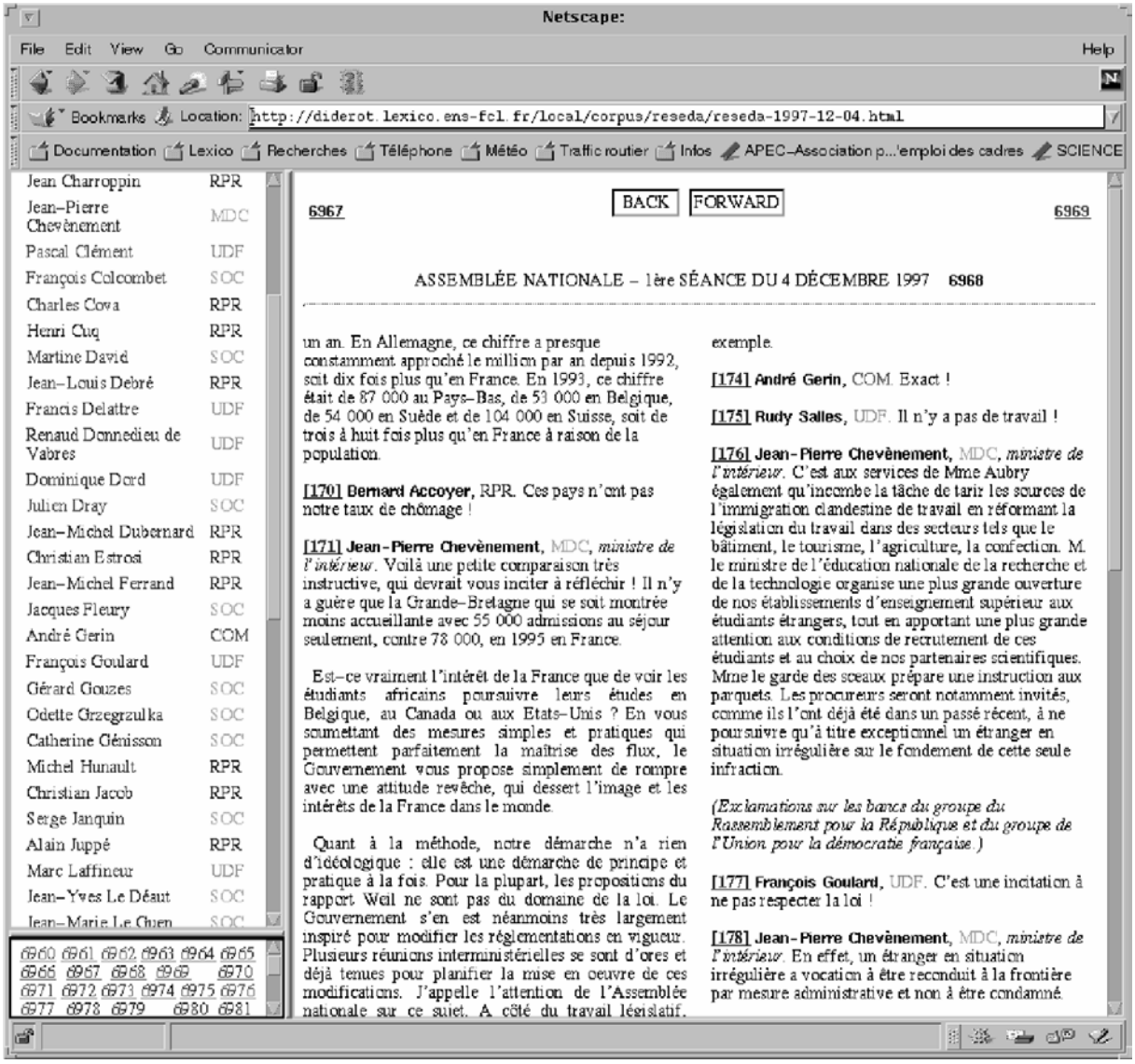

Figure 4.
Hypertexte de la session du 4 décembre 1997 (J.O. daté du 5/12/97), page 6968

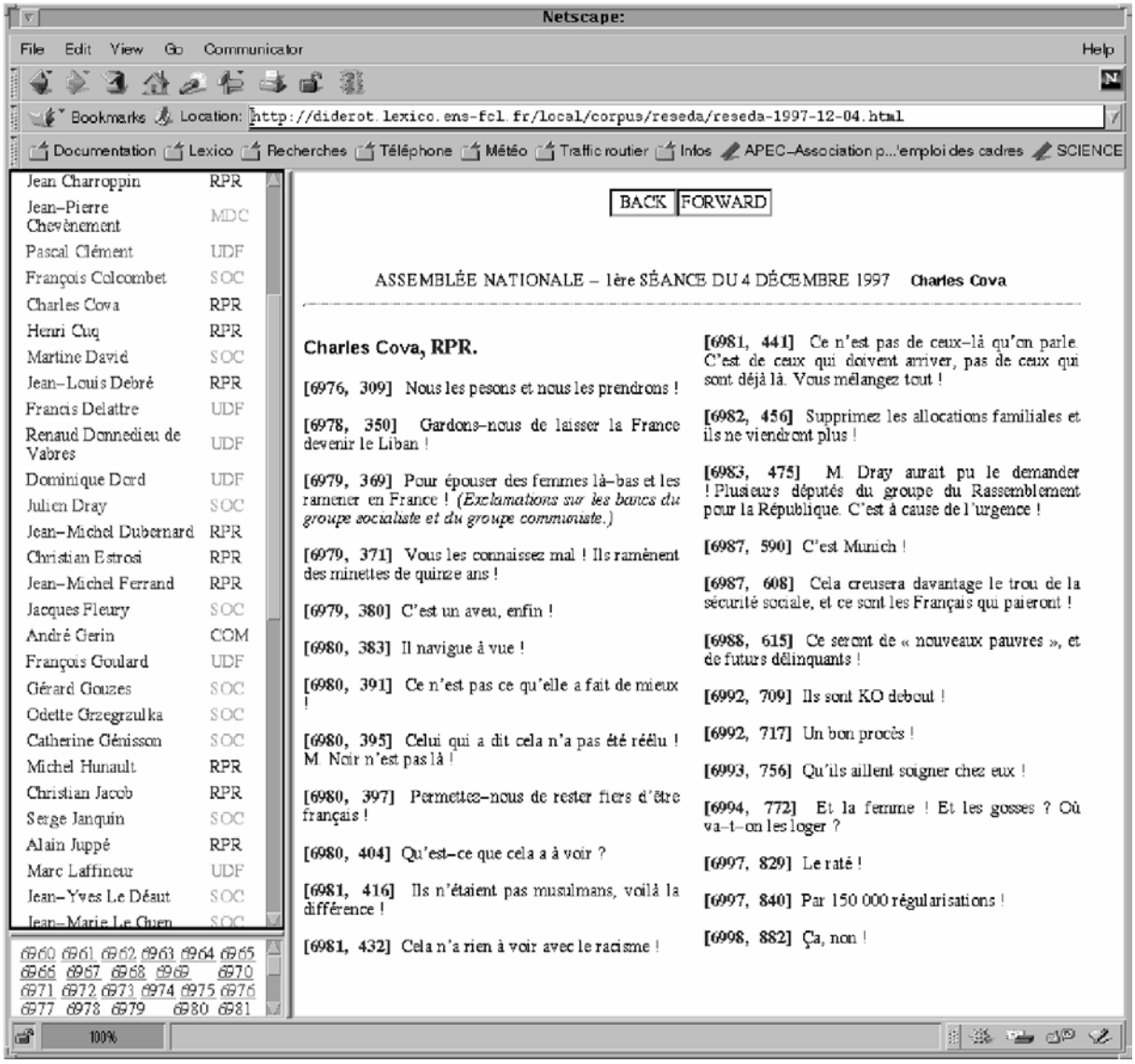

Figure 5.
Hypertexte des prises de parole de Charles Cova pour la journée du 4 décembre 1997.

5 Conclusion

Dans cet article, nous avons définis trois types d'informations fondamentaux en analyse de corpus et présenté les enjeux reliés à l'effort de leur encodage. Nous avons alors proposé une application de la norme SGML sous la forme d'un format simple dont l'application peut être progressive et contrôlable, le format LML. Enfin nous avons montré que la maîtrise d'un format uniforme pour tous les niveaux de granularité des informations encodées (corpus, texte, section, paragraphe, phrase, forme, …) permet la mise en œuvre de nouvelles formes d'exploitation des corpus.

La démarche présentée s'inscrit dans une tendance informatique qui cherche à rendre l'information d'encodage explicite et contrôlable. Le format LML est une première étape vers une application du standard XML, qui offrira la souplesse nécessaire au cycle permanent d'encodage/décodage ayant lieu dans toute recherche sur corpus.