Le péritexte, sésame pour des documents audiovisuels ?
Analyse de corpus textuels interprétant des émissions télévisées à des fins d'indexation.

Karine Lespinasse, Bruno Bachimont
klespinasse@ina.fr, bbachimont@ina.fr

Mots clés :, document audiovisuel, indexation, linguistique de corpus, recherche de document, sémantique textuelle, lexicométrie

L'INA (Institut national de l'audiovisuel) est un établissement public de l'État à caractère industriel et commercial, qui emploie 1 000 salariés. L'INA a pour mission l'archivage et l'exploitation des émissions télévisées et radiophoniques. Ses archives télévisées représentent plus de 2 millions de documents (actualités, téléfilms, magazines...), soit 300 000 heures de programmes. Elles existent sous différents formats (cassette vidéo, film…) mais sont en cours de numérisation. Quarante-cinq ans de télévision représente ainsi 60 km de rayonnage.

L'archivage pratiqué par l'INA vise à faciliter la recherche et la réexploitation des documents. Deux usages sont essentiellement ciblés, d'une part la consultation des archives audiovisuelles par des chercheurs, d'autre part l'utilisation d'archives pour la production de nouveaux programmes.

Nous proposons une définition du document comme consistant en un acte de publication d'un contenu inscrit sur un support. Ceci signifie qu'un document résulte de la volonté d'une institution ou d'une personne de rendre public et/ou accessible un contenu. Le document audiovisuel est donc *un document défini selon un usage* [1]*, qui a une ligne temporelle unique (à la différence du multimédia qui en a plusieurs) et qui mélange des images et du son*. Le point de vue institutionnel de l'INA de ce qu'est un document audiovisuel est celui de l'archive (une mémoire conservée). Il se distingue par exemple du point de vue du diffuseur pour qui l'émission est un élément de la grille de programme.

Des métadonnées permettent de concrétiser le point de vue documentaire lors du procédé d'indexation. Elles correspondent en pratique à la description en texte libre (analyse et résumé), et en texte contrôlé (mots-clés issus d'un thesaurus [2]) du document. Cette pratique de l'INA est en fait générique à la profession, sachant que les documentalistes de l'INA qui indexent assurent aussi la recherche des documents.

Pourquoi mettre en mots des images ? Parce que la langue offre un système fonctionnel, fondé sur des unités auto-normées (par exemple, discrétisables selon les espaces typographiques). Ceci permet d'une part la constitution d'index, ou la normalisation des unités discrètes ainsi définies, et d'autre part des manipulations informatiques (requêtes sur les index par exemple). En effet, il n'existe pas de dictionnaires d'images, alors que chaque unité lexicale est référencée dans un dictionnaire, avec une signification donnée [3].

En pratique, il existe une grande quantité de textes, en amont de la production (scénario, synopsis…) ou en aval (notices documentaires, critiques…). Notre position est que le texte apparaît comme la condition d'accès aux documents audiovisuels : les recherches documentaires sont effectuées sur le texte des notices. En outre, le texte, en proposant un sens explicite en langue, permet l'accès à la signification de l'image. C'est-à-dire il est l'explicitation du point de vue spécifiant en quoi les images sont pertinentes pour l'usage donné. Par exemple, une critique d'émission se donne comme le point de vue de l'intérêt que l'émission présente pour un certain public.

Or, aujourd'hui, du fait de l'évolution engendrée par le développement des technologies numériques, il est envisageable de définir de nouvelles pistes d'accès aux images et une

[1] - Au département Archives, les usages sont majoritairement liés aux demandes des journalistes et des producteurs audiovisuels.

[2] - Thesaurus : vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relations générique-spécifique), selon la norme ISO 5964-1.

[3] - Pour le débat sur la signification de l'image, nous renverrons à (Metz 1968).

nouvelle intégration du texte au document audiovisuel. Sur un support numérique, il est en effet possible d'inclure des métadonnées, autant d'informations supplémentaires mises en parallèle avec le flux audiovisuel (texte de la bande son ajusté à l'image, informations juridiques, mots-clés par segments de documents…) [4]. Il est également envisageable de redéfinir l'indexation sur un ou des documents, (ré)éditables selon une thématique ou un usage autre. Exemple : une recherche sur les allocutions du président François Mitterrand concernant l'Union européenne ou pour cerner l'émergence du concept de société civile entre 1981 et 1995.

Notre problématique de travail, dans le cadre d'une thèse en sémantique et dans une visée exploratoire, entend proposer une typologie de textes en fonction de leur relation aux documents audiovisuels, et une méthode de constitution et de traitement de corpus de texte en cohérence avec ces documents. Nous posons la question suivante : quelle modélisation sémantique est possible, afin de créer ou d'enrichir une terminologie et d'assister le processus d'indexation ?
Nous affrontons donc les problèmes suivants : quels corpus de textes choisir, selon quels usages et pour faire quoi ; comment les constituer ; comment utiliser les résultats ? Deux expérimentations apportent des éléments de réponse. La première a consisté en des tests lexicométriques sur des notices documentaires relatives à la politique. La deuxième, en cours, porte sur l'analyse des réseaux sémantiques à travers les transcriptions de la bande sonore de 63 magazines politiques (« L'Heure de Vérité »).

Dans la première expérimentation, le choix d'une analyse statistique, dans la lignée de (Lebart et al., 1998), s'est imposé d'une part pour sa robustesse : les notices sont bruitées (fautes de frappe, syntaxe «prise de notes», etc.) qui risquent de gêner des analyseurs linguistiques. D'autre part, le corpus offre des régularités a priori intéressantes d'un point de vue lexicométrique, alors que nous ne savions pas encore s'il présentait des régularités distributionnelles permettant certains traitements linguistiques. Bien que les sous-domaines soient nombreux et éclatés, et que les notices soient des textes assez courts (124 mots en moyenne, avec un écart-type de 118 dans ce premier corpus), elles présentent des éléments répétitifs liés à des thèmes (sports, théâtre...) et des genres (actualités, magazines...). Le domaine retenu pour sa cohérence thématique supposé a été celui de la politique intérieure, sur une décennie, de 1980 à 1990 [5]. Le corpus sélectionné dans la base documentaire des archives de l'INA comporte 1,3 million de « mots », soit 10 394 notices, plus de 10 Mo de données textuelles (cf. un exemple de notice, figure1). L'outil de statistique lexicale retenu est Lexico 2.05, développé par l'UPRES SYLED (Université Paris 3 - Sorbonne nouvelle).
Il est apparu premièrement que les hapax représentent 75 % des formes [6], car les entités nommées occupent une place prépondérante (20% des 100 premières formes pleines, i.e. hors mots-outils). D'où une difficulté pour les reconnaître statistiquement. Deuxièment, les résultats sont assez bruités : du fait des abréviations et pratiques « maison », de la présence d'émissions rétrospectives multi-thématiques dans une notice, du domaine large (politique intérieure *dans tous les pays),* d'usage erroné de mots-clés par des documentalistes (*ELECTION* (politique) employé pour *élection de Miss Monde*).
Néanmoins, ces premiers résultats ouvraient des perspectives dans plusieurs directions : l'affinement de méthodes de repérage des entités nommées (elles figurent en majuscules), afin de les associer à la reconnaissance automatique de visages ; le repérage de variantes de mots-clés et le typage de relations sémantiques ; des explorations par patrons lexicaux plutôt que par dénominations, voir (Lespinasse et al. 2000).
Il est apparu également que les résumés sont peu spécialisés et polysémiques, ce qui apparaît lors des traitements automatiques. Partant des unités lexicales *en corpus,* notre effort doit donc porter sur une méthode de définition d'autant plus systématique que ces unités seront manipulées informatiquement. Par ailleurs, le bruit et la pauvreté lexicale des notices encouragent à la recherche de ressources textuelles de types

[4] - Pour une présentation plus ample, consulter (Bachimont 1998).

[5] - Lire (Habert 1985) qui évalue l'intérêt et les limites de l'analyse lexicométrique sur des discours politiques.

[6] - Forme : occurrence composée des mêmes caractères non délimiteurs d'occurrence, les délimiteurs comprennent : _ - : ; / . ? ! * " + = ( ) { } $ et l'espace.

différents, nous renvoyant à la nécessité d'en définir une typologie.

La deuxième expérimentation sur la transcription de plusieurs « Heure de Vérité », magazine de politique, fondé sur le principe de l'interview d'un invité, a donc été mise en place.
Ce corpus est particulier : les notices documentaires relèvent de l'écrit interprétatif (si l'on admet que la mise en mots de l'audiovisuel est un processus interprétatif) alors que les transcriptions appartiennent à un écrit oralisé.
La bande-son a été saisie puis balisée selon une DTD XML. Nous supposons que des structures récurrentes du type défini comme « l'Heure de Vérité », apparaîtront, qui pourront être alignées avec la video et confrontées avec d'autres magazines politiques. Par ailleurs, le balisage incluant le repérage des entités nommées et d'éléments tels que les fonctions de personne, une extraction terminologique avec Lexter, développé à EdF, serait facilitée. Son fonctionnement par découpage de patrons syntaxiques est détaillé dans (Bourigault 1993). Il reste à démontrer que l'outil est adapté à ce type de texte, et qu'il est possible ou pas de faire apparaître des unités lexicales et des relations sémantiques modélisables.

En conclusion, il est apparu qu'une difficulté incontournable de l'approche envisagée réside dans le faible nombre de textes électroniques disponibles. En effet, si nous avons constaté qu'il existe effectivement de nombreux textes liés à des émissions télévisées, très peu sont conservés sous format électronique.
Peu de travaux se sont, d'ailleurs, intéressés au rapport texte-audiovisuel (Srihari, 1994). D'où la nécessité d'établir au préalable une typologie des textes sur l'image, ou du moins de présenter des éléments définitoires de profil de textes par rapport aux documents.

Bibliographie :

- Bachimont B., (1998), "Bibliothèques numériques audiovisuelles. Des enjeux scientifiques et techniques", *Document numérique*, 2 : 3-4, pp219-242.
- Bourigault D., (1993), "Analyse syntaxique locale pour le repérage de termes complexes dans un texte", *TAL*, 34 : 2, 105-117.
- Habert B., (1985), L'analyse des formes "spécifiques", bilan critique et propositions. *Mots*, (11) : pp127–154.
- Jacquemin C., (1997), *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, Habilitation à diriger des recherches en informatique, Université de Nantes, Nantes.
- Lebart L., Salem A. & Berry L., (1998), *Exploring textual data*, Kluwer Academic Publishers, Dordrecht.
- Lespinasse K., Habert B. & Bachimont B., (2000), *Le péritexte, un sésame pour les données audiovisuelles? L'analyse exploratoire d'un corpus hétérogène de notices documentaires interprétant des documents audiovisuel,* JADT 2000, 5es Journées Internationales d'Analyse statistique des Données Textuelles, Lausanne.
- Metz C., (1968), *Essais sur la signification au cinéma*, Klincksieck, Paris.
- Rastier F., co-auteurs : Cavazza M. & Abeillé A., (1994), *Sémantique pour l'analyse*, PUF, Paris.
- Srihari R. K., (1994), "Computational models for integrating linguistic and visual information : a survey", *Artificial Intelligence Review*, p.1-24.

Figure 1 : Une notice documentaire courte.

Mise en rapport d’une notice et du flux audiovisuel

ÉMISSION record8809

YOUGOSLAVIE
POLITIQUE INTÉRIEURE
NATIONALISME
ETHNIE
BOSNIE
EXODE
JOURNAL TÉLÉVISE
230005

Réveil des nationalismes et des ethnies en Bosnie, sous-ensemble de la Yougoslavie.
- DP Départ d’une colonne de réfugiés.
- PL Distribution de nourriture par les soldats.

Manif YOUGOSLAVIE
FR3
SOIR3