Le péritexte, un sésame pour les données audiovisuelles ? L'analyse exploratoire d'un corpus hétérogène de notices documentaires interprétant des documents audiovisuels

Karine Lespinasse, Benoît Habert, Bruno Bachimont

{klespinasse, bbachimont}@ina.fr, habert@limsi.fr

Résumé

In the INA (*Institut national de l'audiovisuel*, the National Broadcasting Institute) dozens of indexers work on audiovisual document retrieval. Their task consists in building up the database by watching and describing the programs broadcast by public channels. This description is a textual interpretation of the program and includes both an univocal factual aspect (eg. dates, names) and a semantic aspect (summary, keywords from a thesaurus). Our assumption is that an audiovisual document, defined as a document mixing images and sound on a unique timeline, does not offer any basic elements such as words in a text, which can be extracted and whose meaning exists *a priori*, but rather allows an infinite number of interpretations. Because indexing at INA requires the use of natural language and an interpretation of the content of the document, we are confronted to the problem of variability and cost in the indexing process. Our objective is therefore to assist the indexers' tasks that can be computer-aided, eg. searching the database and reindexing semi-automatically according to the users' new needs. To reach this point, we have explored the existing indexing files in order to acquire the most frequent terms useful for indexation with their semantic relations, in order to extract units for the enlargement of the thesaurus. The first statistical results in the field of politics showing regularities, the next step will use natural language processing methods and tools.

Résumé

A l'INA, Institut national de l'audiovisuel, plusieurs équipes de documentalistes participent à la constitution d'une base de documents audiovisuels. Elles visionnent, décrivent les émissions télévisuelles des chaînes publiques. Cette description consiste en une interprétation textuelle de l'émission, qui comprend d'une part des aspects factuels, univoques, comme des dates et des noms propres, et d'autre part, un aspect sémantique, sous la forme de mots-clés issus d'un thesaurus et d'un résumé. Nous prenons comme hypothèse de travail que le document audiovisuel, défini comme un document mêlant images et sons sur une seule ligne temporelle, n'offre pas d'unité de sens minimal identique aux mots dans un texte, reconnaissables et extractibles, mais que ses unités résulteront d'une interprétation. Le processus d'indexation en place à l'INA, est donc manuel et partiellement fondé sur la langue naturelle, ce qui induit des phénomènes de variabilité et de coût élevé. Notre objectif consiste par conséquent à assister les tâches des documentalistes, comme la recherche dans la base ou la réindexation des documents selon des nouveaux besoins d'utilisateurs. A ces fins, nous avons constitué un corpus de descriptions documentaires, sur le thème de la politique intérieure, pour tester l'hypothèse de repérage de termes et de relations sémantiques utiles à l'indexation susceptibles d'enrichir le thesaurus. Les premiers résultats, à base statistique, ouvrent des pistes à des travaux d'ingénierie linguistique.

Mots-clé : acquisition terminologique, lexicométrie, document audiovisuel, recherche d'information, relations sémantiques

1. L'INA, institut de l'audiovisuel ou institut du textuel ?

L'INA, Institut national de l'audiovisuel, créé en 1974 pour l'archivage des documents audiovisuels publics [1], gère une base de données audiovisuelles de plus de 900 000 heures de programme, qui s'enrichit de 50 000 heures par an [2]. Ces archives ont pour vocation d'être exploitées, c'est-à-dire commercialisées ou diffusées auprès de publics professionnels, d'où la mise en place de méthodes de recherche de documents. Plus précisément, les documents audiovisuels [3] sont l'objet d'un processus d'indexation qui permet d'effectuer des recherches dans la base de données. Nous donnerons ici à *indexation* la définition suivante : c'est une paraphrase structurée d'un document dans une langue naturelle ou contrôlée rendant le document exploitable en vue d'un usage donné.

Les documentalistes, ont à la fois pour tâche d'indexer les émissions et de faire les recherches dans la base de données. Par conséquent, elles [4] visionnent les émissions récupérées auprès des diffuseurs et les *interprètent sous forme de mots*, dans des *notices documentaires*, qui comprennent des données factuelles, normées, univoques (dates...) et textuelles, partiellement normalisées (résumé, mots-clés issus d'un thesaurus). Le thesaurus [5] est remis à jour périodiquement, ce qui, vu son volume, devient une tâche de plus en plus difficile manuellement. L'indexation comme la recherche de documents audiovisuels se fondent donc majoritairement sur des descriptions textuelles.

Le paradoxe n'est qu'apparent, puisqu'à l'usage, cette démarche reste la plus efficace en matière de recherche documentaire. Notre travail se fonde d'ailleurs sur la constatation que la recherche assistée d'information la plus efficace réside dans le couplage de moteurs de recherche en texte intégral avec un outil de sémantique structurée, de type thesaurus.

Par conséquent, notre objectif premier est de proposer des pistes d'acquisition terminologique pour assister l'enrichissement du thesaurus, notamment en contrôlant davantage les relations sémantiques entre mots-clés, l'objectif ultime consistant à améliorer le processus d'indexation pour améliorer celui de la recherche documentaire. Une coopération plus étroite entre le thesaurus documentaire, normalisé, rigide mais qui maximise la recherche, et les résumés, libres, expressifs mais non réguliers, est à définir.

La démarche adoptée, dans un premier temps, est la suivante : mener des analyses statistiques sur une collection de notices, afin de tester la (non) présence de régularités terminologiques et/ou syntaxiques, pour commencer, selon les résultats, des traitements d'ingénierie linguistique (acquisition terminologique, typage de relations sémantiques).

Nous présentons d'abord le cadre de nos travaux et les problèmes posés : comment les notices documentaires sont-elles élaborées, d'une part, et d'autre part, quel moyen d'accès au document audiovisuel le thesaurus et le résumé en langage libre permettent-ils ? Nous rappelons ensuite le statut exploratoire de l'expérimentation lexicométrique présentée. Nous exposons enfin l'expérimentation : le choix d'un logiciel de lexicométrie et les enjeux de la constitution du corpus de test (corpus de politique intérieure, pendant la décennie 1980), les conclusions au stade actuel.

1. Ses autres missions sont de gérer le dépôt légal (depuis 1992) et d'assurer la formation, la recherche et la production, dans le domaine de l'audiovisuel.

2. Chiffres de 1998.

3. *Document AV (audiovisuel)* : document qui a une ligne temporelle unique, (à la différence du multimédia qui en a plusieurs) et qui mélange des images et du son ; ici, des documents télévisuels exclusivement.

4. Les services documentaires de l'INA étant majoritairement féminins, «documentaliste» sera ici conjugué au féminin.

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

2. L'indexation des documents AV : du normatif et de l'interprétatif

2.1. Un fonds documentaire de plus de trois millions de documents

Le cadre des recherches décrites est celui de la Direction de la Recherche à l'INA et du Département Droits et archives qui constitue son fonds, puis le commercialise auprès des producteurs de l'audiovisuel (chaînes télévisées, producteur indépendant...).

En 1997, le patrimoine de l'INA comprenait plus de 3 millions de documents télévisés et radiophoniques. En 20 ans, le volume du fonds documentaire a triplé. Nous présentons dans la suite les notices des documents.

2.2. Comment une notice d'un document audiovisuel est-elle rédigée à l'INA ?

Cette analyse[6] est effectuée en langue naturelle. Elle s'efforce de respecter la chronologie du document et s'effectue à deux niveaux, qui sont, en tout ou partie, du ressort de l'interprétation de la documentaliste.

- *le catalogage*, au niveau global du document, comprend des données (titre, date de diffusion, nom du réalisateur, type de production de l'émission...) et une qualification de genre (typologie des émissions audiovisuelles, etc.) ;
- *l'indexation*, qui offre un accès au contenu du document, est effectuée au niveau global du document et à des niveaux plus fins si le document peut être divisé en parties. Elle est réalisée en deux étapes successives : premièrement, l'analyse de l'émission, relevée sur une « fiche d'analyse chronologique », rédigée à la main ; deuxièmement, cette fiche permet la mise au point d'une « notice documentaire», sous format numérique, avec un résumé et l'attribution de mots-clés issus du thesaurus.

2.3. Les enjeux documentaires à l'INA

A la lumière de la pratique documentaire présentée, les principaux enjeux sont liés à :

1. la part d'interprétation dans les notices, son influence sur la recherche documentaire ;
2. le peu de contrôle sur la terminologie... : le résumé est de rédaction entièrement libre ;
3. ...et l'usage d'un vocabulaire normé, ici un thesaurus, dont le volume, à savoir plus de 10 000 noms communs, dépasse les capacités de mémoire humaines ;
4. le volume de documents à traiter, que ce soit au niveau de l'indexation ou lors de la recherche des documents.

La part d'interprétation évoquée justifie l'impossibilité de systématicité ni de cohérence strictes. A titre d'exemple, une enquête de la bibliothèque municipale de Nanterre, en 1993, évalue à 30% les mots-clés communs à deux groupes d'indexeurs, qui décrivent le même document, à l'aide du même thesaurus.

De plus, les fluctuations conceptuelles dans l'utilisation du langage d'indexation sont accentuées par les évolutions des demandes des clients, a fortiori pour les documents d'actualité.

Le texte apparaît à la fois comme la condition d'accès aux documents audiovisuels (les recherches documentaires sont effectuées sur le texte des notices) et leur essence (le flux audiovisuel est défini par un texte).

6. Une notice est présentée en annexe.

3. Le textuel comme mode d'accès au flux audiovisuel : recherche d'une terminologie structurée

3.1. Il n'existe pas d'index du flux audiovisuel

Un texte, contrairement à un document audiovisuel, se découpe en unités [7] qui proposent d'emblée un sens. Le document peut en effet, se découper d'après les espaces typographiques ; ainsi il propose au lecteur des unités qui, par défaut, possèdent une signification dans le système fonctionnel de la langue (celle attestée par les dictionnaires de langue, par exemple) [8]. Pour l'audiovisuel, en revanche, il n'existe pas d'index qui s'impose naturellement, ni de signification a priori. Nul ne peut consulter de dictionnaire d'images par exemple. En d'autres termes, les images n'ont pas de signification identifiée dans un système fonctionnel. Mais dans le cadre de cet article, nous nous contenterons de renvoyer au débat sur l'écriture audiovisuelle, lire notamment J. Mitry (Mitry, 1987) et Ch. Metz (Metz, 1968), et de retenir la solution pratiquée à l'INA pour l'indexation d'une séquence audiovisuelle : une description linguistique.

3.2. La tension entre la langue naturelle, libre, et le vocabulaire documentaire, contrôlé

Nous disposons de points d'accès textuels sous forme électronique, aux documents audiovisuels, que nous avons baptisés *péritextes*. Or, la télévision « parle du monde » ; les « dénominations stabilisées » [9] employées à l'INA relèvent d'une part de multiples domaines de connaissance tout en étant faiblement spécialisées, et d'autre part de phénomènes en évolution constante. Elles sont exprimées sous la forme soit de mots-clés structurés en réseau (thesaurus [10]), normés et partageables, mais rigides et dont la surcharge cognitive a été évoquée, soit de descriptions en langue naturelle, innovatrices mais imprédictibles, irrégulières.

Notre problème consiste à introduire plus de souplesse dans le thesaurus tout en renforçant sa structure, pour la rendre plus manipulable : extraire des mini-réseaux sémantiques injectables dans le thesaurus (acquérir des synonymes, des termes spécifiques, etc.), repérer de nouvelles dénominations (variantes...).

L'analyse exploratoire doit permettre, dans un premier temps, de dégager d'éventuelles régularités terminologiques. Nous avons opté pour un outil statistique, Lexico2.05, présenté en 4.4. La deuxième étape, en cas de régularités exploitables, sera la formalisation des résultats linguistiques sous forme de base de connaissances terminologiques, définie par le groupe TIA (Slodzian, 1995).

4. Analyse statistique exploratoire sur un corpus hétérogène

4.1. Pourquoi des traitements statistiques, sur quel type de textes ?

Le choix d'une analyse statistique, dans la lignée de Lebart et al. (Lebart et al., 1998), s'est imposé d'une part pour sa robustesse : les notices sont bruitées (fautes de frappe, syntaxe parfois raccourcie de « prise de notes») 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

7. *Unité* : tout élément découpable qui forme un tout autonome.

8. Cela ne suppose pas que la langue naturelle repose sur des unités prescriptives, univoques, mais que dans le système fonctionnel, les unités proposent une signification déterminée au niveau global du texte ; sur ces questions, nous renverrons à (Rastier et al., 1994) et (Bachimont, 1999).

9. voir Georges Kleiber. (1999). *Sens et structures*. Presses Universitaires du Septentrion, Villeneuve d'Asq.

10. Un thesaurus est fondé sur des relations hiérachiques et associatives.

ne savons pas encore s'il présente 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, écart-type de 118) elles présentent des éléments répétitifs liés à des thèmes (sport, théâtre...) et des genres (actualités, magazines...).

Le champ des mots-clés, qu'il s'agisse des entités nommées [11] ou des descripteurs (noms communs), est alimenté par le thesaurus. Le thesaurus, élaboré en 1975, comprend environ 160 000 descripteurs (dont 94 % d'entités nommées). C'est un outil évolutif, en cours de refonte.

4.2. Construction d'un corpus de test : les choix à faire

Quel domaine choisir pour favoriser la cohérence thématique ?

C'est le domaine de la politique intérieure qui a été sélectionné a priori, en tant qu'il offre une unité lexicale, voir les travaux présentés par (Habert, 1985), à travers des émissions de type varié : journaux télévisés, magazines, etc. La sélection a été effectuée par mots-clé (branches POLITIQUE INTERIEURE et ELECTIONS du thesaurus) et non par « genre » [12].

Quelle(s) période(s) choisir ?

Une période s'est dessinée en termes de cohérence et de volume : la décennie 1980, qui offre une thématique forte en politique intérieure française, puisque'elle est dite « la décennie Mitterrand ».

Une taille de corpus : 1,3 million de mots

Le corpus est constitué de 10 414 notices (plus de 11 Mo de données textuelles). Les notices les plus courtes comportent une vingtaine de «mots» [13], les plus longues 700.

4.3. Présentation du corpus de test

Le corpus porte en mention de responsabilité le service documentaire : les documentalistes n'apparaissent pas en tant qu'auteur. Il est entièrement rédigé en français.

Il s'agit d'un corpus hétérogène en ce qui concerne premièrement les conditions de production (informatisation de l'outil documentaire en 1985), et deuxièmement, le nombre de producteurs. L'hétérogénéité des notices, dont nous cherchons à tester l'importance à travers l'expérimentation présentée, a trait au contenu même des notices.

Pour vérifier a priori quel degré d'homogénéité thématique le corpus présente, deux procédés ont été mis en place. D'une part, nous avons retrié les notices sur les mots-clés des branches POLITIQUE INTERIEURE ou ELECTIONS. Environ 3% des notices constituaient du «bruit» (notices hors sujet, sur la politique extérieure par exemple). Parmi les 97% restant, une certaine proportion, non évaluée, de notices sont multi-sujets (cas des émissions rétrospectives, notamment) : une partie de la notice est consacrée à la politique, le reste à d'autres événements.

D'autre part, quelques sondages ont révélé un phénomène pressenti mais non quantifiable à l'échelle du corpus, qui est l'usage anormal de mots-clés. Par exemple, ELECTIONS a été une dizaine de fois attribué à des documents traitant de *l'élection* de Miss. Il s'avère néanmoins que la moitié des occurrences POLITIQUE INTERIEURE est suivie de ELECTION, ce qui tend à valider la

11. Nous regroupons sous ce terme les noms propres géographiques, de personnes physiques et morales, etc.

12. Les services documentaires ont créé une typologie de genres, que nous avons jugée trop floue pour servir d'élément discriminant de sélection (c'est un mélange de domaines de connaissances (sciences), et de genres littéraires (science-fiction), selon des niveaux différents : « science » côtoie « sciences naturelles », « médecine santé », etc.)

13. *Mot* : unité typographique définie par les séparateurs.

cohérence thématique dans le choix de sélection des notices.

4.4. L'outil de lexicométrie utilisé : Lexico2.05

L'outil étant déjà disponible sur site, nous avons utilisé Lexico2.05, un ensemble d'outils de statistique textuelle ou lexicométrie développé aujourd'hui par l'UPRES SYLED (Université Paris 3-Sorbonne nouvelle). Lexico2.05 comporte notamment des modules de segmentation et de tris par contexte, un concordancier, un inventaire et un tri des segments répétés par partie [14].

L'intérêt de Lexico2.05 réside notamment dans la possibilité de pouvoir analyser les segments répétés de façon contrastive par partie. Nous avons écarté l'hypothèse d'un découpage selon les dates d'élections françaises, puisque le corpus traite de la politique intérieure à travers le monde, même si la France est surreprésentée, au profit d'une partition chronologique égale : de début 1980 à fin 1983 (soit 2 387 notices), de 1984 à 1987 (plus de 4 800 notices) de 1988 à 1990 (3 500 notices).

5. Des résultats en faveur de régularités repérables

5.1. Premières remarques générales : « des noms ! », ou la forte présence des noms propres

Le nombre de formes [15] est de 77 015, celui des occurrences de 1 364 838, celui des hapax 57 408, soit 75 % des formes. Cette proportion élevée s'explique notamment par la très forte présence d'entités nommées (20% des 100 premières formes pleines, i.e. hors mots-outils).

Parmi les 50 premières occurrences de formes pleines, apparaissent :« OFF (signale un commentaire sonore ou la mention d'une personne qui n'apparaît pas à l'écran), EMISSION (qui figure en intitulé de chaque notice), PARIS (lieu de production) , JOURNAL (9896 occurrences), TELEVISE (9558 occurrences), POLITIQUE, TF1, TELEVISION, A2, DP (valeur de plan pour « divers plans »), IT1(appellation de 1975 à 1981 des journaux de TF1), 20h, FRANCAISE, LOI, ASSEMBLEE, JEAN, NATIONALE, FRANCE, INTERIEURE, JA2 (journal Antenne 2), PE (plan étendu), PIERRE, JACQUES, GP (gros plan), PROJET, FRANCOIS, MITTERRAND, CHIRAC, MICHEL, ELECTION, 13h, PS, GOUVERNEMENT, MINISTRE, PARLEMENT, CONSEIL, DEBAT, RPR, MIDI, ALAIN, DISCOURS, SON (adjectif possessif, ou bruit?), gouvernement, ministre, K7 (cassette), MINISTRES, ROCARD, politique, PARLEMENTAIRE, ITW (interview), UDF ».

La description de l'image occupe, comme il était prévisible, une place importante (environ 1/8è des occurrences citées), par rapport au vocabulaire proprement politique.

Le nombre d'occurrences des segments répétés chute rapidement: si OFF, EMISSION, PARIS apparaissent au moins une fois par notice, JOURNAL n'apparaît plus que dans 95% des notices (les documents sources sont majoritairement des journaux télévisés). Les formes suivantes, qui caractérisent le contenu, apparaissent dans au plus 62% du corpus.

Le tri par segments répétés fait apparaître peu d'expressions complexes (plus de 2 formes) en langue naturelle. Les expressions les plus longues sont les suivantes : « affaire du Carrefour du développement », « nième tour des élections (présidentielles, régionales, etc.) », « autorisation administrative de licenciement», « créer des emplois ». Le degré de spécialisation paraît faible.

La politique française est surreprésentée. Ainsi parmi les personnages, le premier étranger, GOR-

14. *Segment répété* : ensemble de formes consécutives dont la fréquence est égale ou supérieure à 2 dans le corpus.

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

BATCHEV, qui apparaisse, est au rang 486, après JEAN, PIERRE, JACQUES, FRANCOIS, MITTERRAND, CHIRAC, MICHEL, RPR, ALAIN, ROCARD, UDF, etc.

5.2. Une typosémiotique, pour assister l'extraction de variantes ?

L'usage de la typographie justifie d'une analyse spécifique. Les consignes de rédaction des notices sont en effet les suivantes depuis 1985, à savoir : saisie des différents champs en majuscules, à l'exception du résumé, où seules les entités nommées, par souci de lisibilité, sont en majuscules. Un effet de brouillage est néanmoins induit par la pratique antérieure, où les notices étaient saisies par des opérateurs, à partir de bordereaux écrits à la main, entièrement en majuscules, par les documentalistes. Aujourd'hui encore, certaines documentalistes continuent à ne rédiger les notices qu'en majuscules. Donc, une occurrence en minuscules est issue du résumé (langage libre), et peut figurer en doublon avec la même occurrence, en majuscules, issue d'autres champs ; en revanche, une occurrence en majuscules a pour origine vraisemblable les champs contrôlés, mais peut provenir du résumé.

Nous avons alors testé l'hypothèse de repérer les variantes issues de rédactions en langue naturelle, d'un mot-clé, sur les intitulés de ministres et de ministères. Il apparaît qu'en langue naturelle, il est toujours question des fonctions (ministres) et non des institutions (ministères). En langue naturelle, des expressions précises comme « ministre délégué auprès du ministre » ou «ministre délégué chargé » sont comptabilisées plus d'une dizaine de fois chacune, sans équivalent dans les mots-clés du thesaurus, ce qui constitue l'amorce d'acquisition de variantes et de leurs relations sémantiques.

5.3. Les possibilités de repérage de l'émergence de nouveaux thèmes

Certains phénomènes apparaissent massivement : par exemple, dans la 3è partie chronologique, les événements omniprésents sont les manifestations de la place Tian An Men, à Pékin, en 1989. Une hypothèse de travail consiste à détecter à un niveau antérieur, donc plus fin, l'apparition de ce qui sera une thématique forte et l'émergence d'une molécule sémique.

Par exemple, la partie 2 est centrée sur « DEVAQUET », « ENSEIGNEMENT » et « COHABITATION ». En partie 1, ces formes sont sous-représentées : ENSEIGNEMENT LOI, fréquence : 217, sous-emploi : -E15 ; DEVAQUET ALAIN, 210, -E15 ; cohabitation, 236, -E16 ; COHABITATION, 714, -E48.

6. Conclusion et perspectives

Le discours des et sur les actualités apparaît comme un discours vulgarisateur (faible part des expressions complexes, peu de complexité en termes de patrons syntaxiques). Toutefois, le constat devrait être complétée par l'analyse des émissions télévisuelles spécialisées en politique. Les entités nommées y occupent une place prépondérante ; l'acquisition de variantes peut s'appuyer sur des spécificités du corpus (effet de typosémiotique). Un protocole d'évaluation est en cours de définition avec une documentaliste.

Nos pistes de recherche sont maintenant les suivantes : d'une part, nous devons travailler dans le sens de l'homogénéisation thématique du corpus, par exemple, construire un corpus sur la politique intérieure française. Les éléments discriminants sont à trouver (FRANCE n'est pas systématiquement en mot-clé). D'autre part, trois ouvertures apparaissent en faveur de travaux ultérieurs :

– développer les méthodes de repérage des entités nommées (typosémiotique, outils spécifiques), dans la perspective de pouvoir l'associer à la reconnaissance automatique de vi-

sages ;
- profiter de la masse volumique du corpus textuel potentiel pour extraire les variantes de segments répétés et les liens sémantiques, à des fins de structuration du thesaurus, selon les travaux de Ch. Jacquemin (Jacquemin, 1997) ;
- envisager des explorations par patrons lexicaux plutôt que par dénominations.

Références

Bachimont B. (1999). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In Charlet J., Zacklad M., and Kassel G. editors, *Ingénierie des connaissances*. Eyrolles, Paris.

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., and Berry L. (1998). *Exploring textual data*. Kluwer Academic Publishers, Dordrecht.

Metz C. (1968). *Essais sur la signification au cinéma*. Klincksieck, Paris.

Mitry J. (1987). *La sémiologie en question, langage et cinéma*. Les Editions du Cerf, Paris.

Rastier F., co-auteurs : ., Cavazza M., and Abeillé A. (1994). *Sémantique pour l'analyse*. PUF, Paris.

Slodzian M. (1995). Epistémologie de la terminologie. *La banque des mots*, 7:pp11–18.

Annexe : Une notice du corpus Politique 1980-1990

EMISSION record 943

MOTCLE1(1)= MERMAZ LOUIS

MOTCLE1(2)= ASSEMBLEE NATIONALE

MOTCLE3(1)= MASURE BRUNO

RDATCRE(1)= 19811020

RDATDIF(1)= 19810702

RDUR(1)= 000145

RFORM(1)=JOURNAL TELEVISE

RPRDLIB(1)=TELEVISION FRANCAISE 1

RPRDLIEU(1)=PARIS

RPRDSIG(1)=TF1

RRES IMAGES DE MERMAZ,LE NOUVEAU PDT DE L'ASSEMBLEE NATIONALE: PDT UNE FETE,LA NUIT,IL MANGE DES MERGUEZ DVT UN FEU DE BOIS.IL FAIT SA CAMPAGNE POUR LES LEGISLATIVES A VIENNE(ISERE)-G.P.DE SA FEMME1 ENFANT-MERMAZ AU TRAVAIL DS SON BUREAU,FETANT LE 10 MAI,ARRIVANT A PIED A L'A.N. RD DS IT1 NUIT.

RTI(1)=PORTRAIT MERMAZ

RTICOL(1)=IT1 20H