Apport des méthodes lexicométriques à l'étude d'un texte: Evolution du vocabulaire, coupures thématiques et stratégie discursive

Mónica Bécue Bertaut
Dpt. Estadística e Inv. Operativa
Universitat Politècnica de Catalunya

1. Introduction

Les outils de la statistique textuelle sont habituellement utilisés pour analyser des corpus textuels de grande ampleur (cf. Lebart et Salem, 1994; Thoiron *et al*., 1988). Dans ce travail, nous voulons montrer que ces outils sont utiles lors de l'étude d'un texte long mais unique. L'étude du vocabulaire et de sa répartition apporte alors des éléments d'information sur la construction du texte.

On cherchera, en particulier, à situer les coupures thématiques et à déterminer quel est le vocabulaire des parties délimitées. On verra que la répartition des mots dans le texte apporte une information intéressante sur le rythme et le contrôle du discours, tout particulièrement quand le texte est oral et élaboré au fur et à mesure de son émission, comme c'est le cas ici.

Le texte analysé est le réquisitoire prononcé à la fin d'un procès pour assassinat à l'Audience de Barcelone. On peut s'attendre à que ce texte qui veut démontrer une thèse, persuader et convaincre les auditeurs, offre des éléments réthoriques particulièrement riches. Il s'agit d'un texte isolé et clos sur lui-même. La stratégie mise en oeuvre par le procureur - choix des mots, argumentation, ordre, rythme de l'exposition, etc.– correspond certainement au style réthorique du procureur mais dépend en grande mesure du cas concret– de ses particularités, de l'histoire juridique du cas, des juges qui composent le tribunal et de leurs habitudes - .

2. Présentation du réquisitoire

2.1 Le procès

Le patron d'un sauna-relax - où, en fait, l'on pratiquait la prostitution - et son associé sont accusés d'avoir planifié et réalisé l'assassinat d'une des femmes qui travaillaient dans leur

local, afin de toucher une très confortable assurance-vie souscrite par la victime et ayant pour bénéficiaires le patron du sauna et ses enfants, et non les propres enfants de la victime. Celle-ci avait été retrouvée morte, égorgée, son cadavre abandonné dans un terrain vague d'une ville de la banlieue de Barcelone.

Le procès a été suivi par l'équipe de Pompeu Casanovas[1]. Le procès dura deux jours, le réquisitoire final a une durée d'une heure quinze minutes. L'absence de preuves matérielles rend particulièrement complexe le rôle de l'accusation publique et donne lieu à un réquisitoire du procureur plutôt riche et relativement long.

2.2 Le réquisitoire

En règle générale, le réquisitoire doit partir des faits et mettre en oeuvre un raisonnement qui leur donne le caractère d'indices, en les reliant de façon logique au délit, au delà de la matérialité des preuves. Le procureur l'élabore au fur et à mesure du déroulement du procès, en fonction des déclarations des accusés, des témoins et, éventuellement, des experts. Il ne peut écrire qu'un simple schéma de son intervention.

Le plan est assez classique: les délits dont le procureur va accuser le(s) prévenu(s) sont énumérés; le cadre juridique applicable est ensuite précisé; puis les faits sont décrits, reliés les uns aux autres afin de prendre place dans ce cadre; finalement les faits sont qualifiés juridiquement, qualification qui clôt ce texte.

Néanmoins, chaque réquisitoire est le fruit d'une stratégie spécifique que le procureur adopte selon ses habitudes et de son style propre, mais aussi en fonction du cas et de ses difficultés ou particularités.

Dans le cas qui nous occupe, le procureur accuse les deux prévenus de trois délits: outre l'assassinat. il considère la commission d'un délit de proxénétisme et d'un délit d'escroquerie à l'assurance.

Lors de son interrogatoire, le patron a affirmé que la femme assassinée était sa compagne sentimentale, ce pourquoi ils avaient souscrit une assurance-vie réciproque. Il a apporté des éléments pour prouver qu'il était absent de Barcelone le jour du crime, en voyage à Burgos, et que, alarmé par la disparition de la victime, il l'avait recherchée activement.

Le procureur cherche à démonter cette version, à montrer ses invraisemblances. Il fait voir que les accusés avaient choisi une femme dépourvue, lui avaient fait croire à une relation sentimentale afin de lui faire souscrire une assurance-vie et avaient attendu le moment propice pour la tuer de façon à avoir un alibi et pouvoir toucher cette assurance. Sur quoi fonder cette accusation? Il n'y a pas de preuves matérielles indiscutables. L'unique témoin direct, une employée du sauna, qui avait entendu une conversation entre

---

[1] Le laboratoire d'Ethnologie et Anthropologie du Droit de l'Université Autonome de Barcelone a suivi et filmé le déroulement intégral du procès (Casanovas et al. 1995) réalisé à l'Audience Provinciale de Barcelone en 1993, dans le cadre d'une recherche sur l'administration judiciaire et son fonctionnement.

les deux accusés commentant leur projet de tuer la victime, s'est rétracté, pris de peur, devant le juge d'instruction d'une première déclaration faite à la police. Le procureur souligne l'absurdité des déclarations des accusés, en particulier en ce qui concerne la relation de la victime et du patron ainsi que le comportement de ce dernier le jour de l'accident, lors de la prétendue recherche de la victime.

2.3. Préparation du texte

Le réquisitoire oral a été transcrit en employant une orthographe correcte, mais en conservant les incorrections syntactiques de certains passages, et sans introduire de signes de ponctuation. Les coupures marquées par le procureur, au moyen d'un silence ou d'un changement d'intonation, ont été utilisées pour segmenter le réquisitoire en 55 “espaces discursifs” (table 1).

On obtient ainsi un corpus dont la longueur totale est de 10417 occurrences: ce corpus est formé à partir de 1806 mots distincts. La longueur moyenne d'un espace est de 189.4 occurrences, mais la variabilité est assez grande: le plus court est constitué de 54 occurrences, le plus long de 463.

Table 1. Premiers espaces discursifs

----espace 1
con la venia por este ministerio fiscal se ha formulado un escrito de conclusiones elevado a la definitiva por varios hechos un delito relativo a la prostitución un delito de asesinato así como un delito de estafa en concurso ideal con el anterior con respecto al primero de ellos es decir al delito relativo a la prostitución no nos queda ningún género de dudas sobre su comisión es evidente de que queda más que suficientemente acreditado que este delito relativo a la prostitución y no delito de intento contra la prostitución como oímos ayer de la defensa es un delito contra la libertad sexual eso ha de quedar bien claro
----espace 2
este delito queda acreditado no sólo por las declaraciones de los testigos que aquí han comparecido así como por todas estas manifestaciones que se han depuesto a lo largo del sumario en la que todos ellos hablan por ejemplo FTA que trabajaba que trabajaba en la referida sauna los servicios que prestaba también por otros elementos de prueba por ejemplo las actas de entrada de registro que si se observa en la misma la entrada de registro de la Avenida Meridiana donde se encontraba la sauna relax había un tablón con los precios de los distintos servicios que se prestaban pero ya no sólo eso sino ya existe la propia confesión de los acusados en todo momento de que se dedicaban a este tipo de acciones

3. Objectif

L’ objectif est de dévoiler comment le procureur utilise et organise l’information dont il

dispose pour structurer son réquisitoire et construire son argumentation. Tout d'abord les mots les plus fréquents seront identifiés. Puis l'étude de la répartition de ce vocabulaire permettra d'extraire les mots stables, utilisés de façon régulière au long du corpus, reflet de la tonalité générale de celui-ci, par opposition aux mots dont l'usage est nettement localisé. Ensuite, le recours à l'analyse de correspondances ainsi comme l'étude de l'accroissement du vocabulaire mettront en évidence la structure chronologique du réquisitoire et faciliteront sa segmentation en périodes significatives. Finalement, la caractérisation de ces périodes par leurs traits lexicaux permettront de donner sens à la structure temporelle.

4. Le vocabulaire du réquisitoire et sa répartition

4.1 Vocabulaire du réquisitoire

Les glossaires des mots et des segments répétés offrent une première information quantitative sur le corpus. Comme il est habituel, les mots les plus employés sont des mots-outils. Le mot plein le plus fréquent est *seguro* (*assurance, 51 citations*) - dans ce qui suit, le numéro mentionné après le mot traduit indique la fréquence d'emploi -. Il s'agit, l'on s'en doute, de l'assurance-vie qui est liée au délit d'escroquerie et qui servira d'argument important pour soutenir le raisonnement sur la commission de l'assassinat. De la lecture des mots et segments prononcés au moins cinq fois ne découle guère de surprise; on y retrouve les particularités du cas. Ainsi on trouve les mots *persona* (personne, 40), *seguro de vida* (assurance vie, 25), *relación* (relation, 26), *MJA* (initiales de la victime, 21), *JCM* (initiales de l'accusé, 18), *hijos* (fils, 21), *FPM* (initiales du complice, 17), *SRT* (initiales de l'épouse de l'accusé, 16), *millones (millions,* 13), *beneficiarios* (bénéficiaires, 11). L'enquête judiciaire et policière induisent certains mots comme *policía* (police, 27), *declaración(es)* (déclaration(s), 43), *delito* (délit, 21), *caso (cas,* 17), *defensa* (défense, 17), *manifestaciones* (manifestations, 16), *prueba* (preuve, 15), *MF (initiales du seul témoin à charge, femme employée du sauna,* 15), *actuaciones (comportement,* 13), *ministerio* (ministère, 13), *acusados (accusés,* 12), *testigo (témoin,* 11).

Il n'y a pas de preuve matérielle irréfutable et, bien que le procureur déclare avec aplomb "*No tenemos una prueba de cargo....lo que no supone tampoco mayor problema*" (nous n'avons pas de preuve à charge...ce qui ne pose pas grand problème), il se préoccupe de montrer qu'il y a de très nombreux faits et données qui, logiquement mis en relation, constituent des indices: *hecho(s)* (fait(s), 50), *otro(s) dato(s)* (autre(s) donnée(s), 32), *indicios* (indices, 11). Il souligne l'importance et le nombre de ces indices*:"un cúmulo de índices*" (un cumul d'indices), *"no estamos hablando de sospechas, no quiero dar esa oportunidad a la defensa, estamos hablando de indicios, estamos hablando de indicios plenamente contrastables"* (nous ne parlons pas de suppositions, je ne veux pas offrir cette opportunité à la défense, nous parlons d'indices, nous parlons d'indices parfaitement contrastables); en effet, selon la jurisprudence citée par le procureur, disposer d'indices solides suffit pour obtenir une condamnation.

Le procureur ne veut laisser aucune place au doute et cherche à donner à son discours un

ton ferme et assuré: *realmente (réellement,* 44), *consta* (il est établi, 31), *perfectamente* (parfaitement, 18), *es evidente* (il est évident, 13) *tenemos* (nous avons, 14), *sabemos* (nous savons, 11). Le procureur souligne les contradictions contenues dans les déclarations des accusés -*si, (si, conditionnel)* est cité 57 fois et *no* (ne...pas) 203 fois- et cherche à renforcer ses propres affirmations -*sí (si, renforcement d'une affirmation,* 18 citations)-.

Ces premiers résultats procurent une information sur les aspects du cas les plus cités dans le réquisitoire, et aussi sur le recours à certains types d'argumentation. Il s'agit, maintenant, de mettre en évidence la répartition temporelle du vocabulaire, et de montrer comment celle-ci constitue aussi la trace des choix stratégiques du procureur.

4.2 Vocabulaire stable et vocabulaire spécialisé

4.2.1 Indice de répartition de Hubert -Labbé

L'indice de répartition des mots que nous utilisons ici a été proposé par P. Hubert et D. Labbé (1990). Cet indice cherche à mesurer la particularité de la répartition de certains mots et peut être utilisé sans qu'aucun découpage du corpus en parties ne soit effectué.

L'indice de répartition est calculé à partir de la longueur des intervalles qui séparent les répétitions successives d'un même mot. Pour un mot de fréquence absolue égale à F dans un corpus de longueur N, cet indice, qui varie entre 0 et 1, peut être considéré comme une approximation à la probabilité d'employer ce mot dans une partie quelconque du corpus de longueur N/F.

Une valeur proche de 1 indique que le mot est employé de façon régulière, une valeur proche de 0 est au contraire la marque d'un usage circonstanciel, localisé du mot. Dans le réquisitoire étudié ici, en ne considérant que les mots utilisés au moins cinq fois, l'indice de répartition varie entre 0.17 - pour le mot *oído* (entendu*)* - et 0.79 pour *partir* (partir*)* de la locution *a partir de* (à partir de*)* -. Comme on le verra plus loin, le mot *oído* est l'un des mots utilisés quand le procureur, après beaucoup d'hésitations, mentionne la déclaration– et la discussion sur sa validité, puisque ce témoin s'est rétracté par la suite –du seul témoin à charge, une femme de ménage du sauna-relax qui avait entendu une conversation entre les deux complices au sujet de l'assassinat.

4.2.2 Le vocabulaire stable ou la trame du réquisitoire

Le vocabulaire stable, utilisé de manière régulière tout au long du réquisitoire en constitue, d'une certaine façon, la trame. Ces mots servent à donner le ton du réquisitoire et finissent par imposer un message non tant par leur fréquence que par leur régularité. Les causes de cette dernière peuvent être diverses. Elle provient de la situation d'énonciation très particulière que constitue un réquisitoire, des habitudes de langage propres au procureur mais aussi de la stratégie mise en oeuvre par celui-ci pour convaincre le juge et obtenir la condamnation des accusés. Ces mots sont fondamentaux, ils servent à créer un fond qui se maintient constant sur lequel, selon les moments, le procureur cherchera à détacher des faits et des raisonnements plus localisés.

Bien sûr, un grand nombre de ces mots sont des mots-outils, d'emploi indispensable, imposé par la langue. Mais comme le montre la table 2, les mots outils sont loin d'être les seuls mots réguliers. Le mot le plus régulier en est un; il s'agit de *partir*, avec un indice égal à 0.79. En fait, ce mot appartient toujours à la locution *a partir de* (à partir de). Notons que le deuxième mot régulier est le mot plein *testimonio* (témoignage*)* avec un indice égal à 0.78.

En prenant les mots dont l'indice de répartition est au moins égal à 0.60 (table 2), on trouve tout d'abord une série de mots qui se répètent (et, très certainement se substituent les uns aux autres mais cela reste une hypothèse) pour donner un ton assuré, un ton de démonstration sans faille au réquisitoire: *perfectamente, exactamente, totalmente, fundamental, plenamente, evidentemente, lógico* (parfaitement, exactement, totalement, fondamental, pleinement, évidemment, logique). Si l'on ajoute *realmente* et *evidente* (réellement et évident*)*, dont les indices sont seulement légèrement plus faibles, on constate la présence régulière de mots destinés à manifester la conviction du procureur et à créer, par leur apparition régulière, le message de l'évidence de la culpabilité. Il n'y a pas de preuves matérielles, le seul témoin direct qui a entendu une conversation entre les accusés qui commentaient entre eux l'assassinat s'est rétracté. Mais les mots *testigos, testimonio, declaraciones, manifiesta, manifestó, hechos, a partir de, sabemos, decimos* (témoins, témoignage, déclarations, manifeste, manifesta, faits, à partir de, nous savons, nous disons) apparaissent régulièrement pour montrer que le raisonnement du procureur s'établit sur des éléments solides, extrait des conclusions à partir de faits avérés et des déclarations des témoins et des propres accusés. Finalement, on peut noter que certains connecteurs *y, así, sino* (et, ainsi, sinon) sont aussi des mots stables. Il y a bien un discours construit qui relie ou oppose les faits et les déclarations.

La lecture des indices ainsi fournis, fréquence des mots et indice de répartition, peut et doit souvent être complétée par un retour au texte ou, du moins, au contexte des mots pour permettre de lever certaines ambiguïtés et pour mieux capter le sens de leur emploi dans ce texte.

4.2.3 Mots localisés

Les mots localisés indiquent des idées ou des thèmes plus ponctuels, abordés exclusivement dans certaines parties du corpus. Le calcul de l'indice de répartition permet d'en obtenir la liste, mais non de savoir comment ils s'associent entre eux, ni à quel ou quels moments ils interviennent. On peut néanmoins noter que les mots *testigo* (témoin) et *declaración* (déclaration*)* ont, en contraste avec *testigos* (témoins) y *declaraciones* (déclarations), une répartition beaucoup plus localisée. Qu'en conclure? Tout d'abord, qu'une lemmatisation entreprise avant tout traitement ne nous aurait pas permis de détecter ce phénomène. La poursuite de l'analyse et le retour au texte permettront de savoir ce que signifie cette information, et de voir que parmi les témoins il y a "*le témoin*" qui est cette femme qui a entendu la conversation entre les accusés, que parmi les déclarations, il y a *"la déclaration"* qui est celle, justement, de ce témoin.

5. Analyse de correspondance et étude de l'évolution du vocabulaire

Pour effectuer le découpage du corpus, nous utiliserons les résultats offerts par deux

méthodes complémentaires: l'analyse des correspondances comme outil de représentation des séries textuelles chronologiques comme le propose Salem (1991, 1993) et le modèle de partition du vocabulaire présenté par Hubert et Labbé (1988).

Table 2. Mots les plus localisés et mots les plus stables

Mots les plus spécialisés						Mots les plus stables					
	Ind.	Fré	Mot	Ind.	Fré	Mot	Ind.	Fré	Mot	Ind.	Fré
oído	0.17	6	desaparición	0.33	5	las	0.60	100	y	0.64	298
cadaver	0.19	6	hijos	0.34	21	en	0.60	271	ser	0.65	9
devolver	0.20	5	sexual	0.34	8	bueno	0.60	7	hasta	0.65	10
protección	0.20	5	existencia	0.34	9	su	0.60	55	totalmente	0.66	6
cargo	0.23	5	acuerdo	0.34	5	cuando	0.60	28	pero	0.66	50
delito	0.23	21	trtamiento	0.34	6	favor	0.60	6	fundamental	0.66	7
familia	0.23	6	cuatro	0.35	13	ahí	0.60	8	esa	0.66	15
prostitución	0.23	8	suscribió	0.35	7	perfectamente	0.60	18	igual	0.67	9
confidencial	0.23	5	les	0.36	16	con	0.60	105	haya	0.67	6
capital	0.23	7	asesinato	0.36	5	bien	0.60	13	tengo	0.67	7
jubilación	0.23	8	qué	0.36	31	otro	0.61	24	vez	0.68	7
SRT	0.24	16	investigación	0.37	6	al	0.61	39	siempre	0.68	7
fiscal	0.25	9	ayer	0.37	8	aquellas	0.61	6	tema	0.68	5
a–os	0.25	5	muerte	0.38	13	a–o	0.61	12	propias	0.68	7
beneficiaros	0.26	11	sumario	0.38	8	tenido	0.61	7	ocasiones	0.69	7
actos	0.26	6	desde	0.38	10	puede	0.61	7	ese	0.69	14
MF	0.29	15	sostén	0.38	7	declaraciones	0.61	23	así	0.69	22
frente	0.29	7	millones	0.39	13	puesto	0.61	20	creo	0.70	7
tribunal	0.29	8	vica	0.39	26	hechos	0.61	16	primero	0.70	6
relativo	0.30	5	dicho	0.39	11	algo	0.61	5	sería	0.70	7
policia	0.31	27	JN	0.39	7	de	0.61	587	entre	0.70	7
otras	0.31	5	declaración	0.39	20	distintas	0.62	5	plenamente	0.70	6
cincuenta	0.31	7	conocía	0.39	12	FP	0.62	15	eran	0.71	6
mayor	0.32	12	persona	0.40	40	va	0.62	9	unos	0.72	14
testigo	0.33	11	seguro	0.40	51	iba	0.62	6	evidentemente	0.72	5
hace	0.33	10				manifiesta	0.62	7	perdón	0.72	13
						sino	0.62	13	testigos	0.72	5
						manifestó	0.62	5	toda	0.73	6
						cada	0.63	10	entonces	0.74	5
						decir	0.63	36	empezó	0.75	5
						ejemplo	0.63	5	lógico	0.75	9
						todo	0.63	13	contar	0.75	5
						decimos	0.63	13	medio	0.75	5
						mantenía	0.63	5	tenemos	0.76	14
						sabemos	0.64	11	respecto	0.76	6
						mismo	0.64	7	lugar	0.76	8
						delictivo	0.64	5	sobre	0.77	8
						exactamente	0.64	5	tampoco	0.77	5
						hAy	0.64	16	cual	0.77	7
						personas	0.64	15	testimonio	0.78	5
						que	0.64	636	partir	0.79	6
						nuevo	0.64	5			

L'analyse de correspondances de corpus temporels, découpés en parties déterminées par la chronologie met à jour leur structure évolutive. Il est fréquent d'observer un premier axe factoriel sur lequel les différentes parties se succèdent de façon ordonnée. En effet, deux parties consécutives sont relativement proches l'une de l'autre parce que les mots apparaissent et disparaissent progressivement. Si cette rénovation est régulière et prédominante, alors les différentes parties se positionnent sur le premier plan factoriel le long d'une courbe de forme approximativement parabolique.

D'autres facteurs peuvent entrer en jeu et altérer ce schéma. En particulier, cette régularité est usuellement rompue lors des moments de particulière signification, quand un événement pénètre avec force dans le corpus et altère la rénovation du vocabulaire.

Ce schéma évolutif a été constaté pour des corpus émis par une même source textuelle, dans des conditions d'énonciation similaires, qui s'étendent sur un temps relativement long et sont souvent influencés par des circonstances extérieures changeantes dont ils sont aussi le reflet.

Il nous a semblé utile d'appliquer l'analyse de correspondances à ce réquisitoire, segmenté en blocs arbitraires de longueurs similaires, puisque la nécessité de démontrer, de présenter un raisonnement logique qui relie les faits en un schéma cohérent, doit entraîner une présentation progressive et ordonnée des arguments.

Pour cette analyse, le réquisitoire est segmenté en dix blocs de longueurs approximativement égales tout en respectant l'espace comme unité discursive. Pour soumettre à l'analyse de correspondances la table lexicale Blocs x Mots, seuls les mots prononcés au moins cinq fois sont conservés.

Une première analyse permet de détecter que le bloc 8 présente un profil lexical fortement atypique; pour cette raison, il détermine le premier axe sur lequel il s'oppose à tous les autres blocs. Une fois mis en évidence ce phénomène local mais qui domine toute l'analyse, et pour obtenir une représentation de la structure globale, qui est bien sûr ce qui est recherché, il est préférable de déclarer ce bloc supplémentaire pour l'analyse de correspondances.

Le plan principal fourni par cette deuxième analyse (figure 1), qui conserve 36% de l'inertie totale, montre que les six premiers blocs présentent une évolution temporelle assez régulière, suivie d'un retour en arrière du bloc 7, et puis d'une progression plus désordonnée des blocs suivants. On ne peut prêter trop d'attention à la position du bloc 8 mal représenté dans ce plan (rappelons qu'il est positionné en tant qu'élément supplémentaire). Le bloc 9 se dévie de la courbe, attiré par les blocs 2 et 3; quant au bloc 10, on peut le considérer comme l'extrême de l'arc de retour commencé au bloc 7. En résumé, si on excepte les blocs 8 et 9, – considérés comme des accidents–, on observe une forme parabolique complétée par un arc de retour vers le début, une structure en quelque sorte cyclique.

Le schéma ainsi observé demeure relativement stable, face à des modifications du seuil de fréquence. Il s'agit d'un texte fortement structuré, à la progression régulière jusqu'au bloc 6; cette progression est interrompue vers la fin du bloc 7 ou le début du bloc 8, lorsqu'un certain phénomène pénètre avec force dans le corpus et donne un caractère très spécifique à ce dernier bloc. Ensuite, on ne retrouve plus cette régularité. On constate donc une division entre une première partie du réquisitoire ordonnée et une deuxième partie, moins contrôlée et moins élaborée. Ces deux parties sont séparées par une courte séquence très atypique.

Le plan principal n'offre qu'une représentation partielle de la correspondances entre mots et blocs. La classification automatique des blocs à partir de leurs coordonnées sur les cinq premiers axes, conservant ainsi 73% de l'inertie, permet de résumer la lecture de ces axes et de compléter les résultats obtenus sur le plan principal. Il a été choisi d'employer une méthode hiérarchique et d'utiliser le critère de Ward généralisé, comme indice de distances entre groupes et entre éléments.

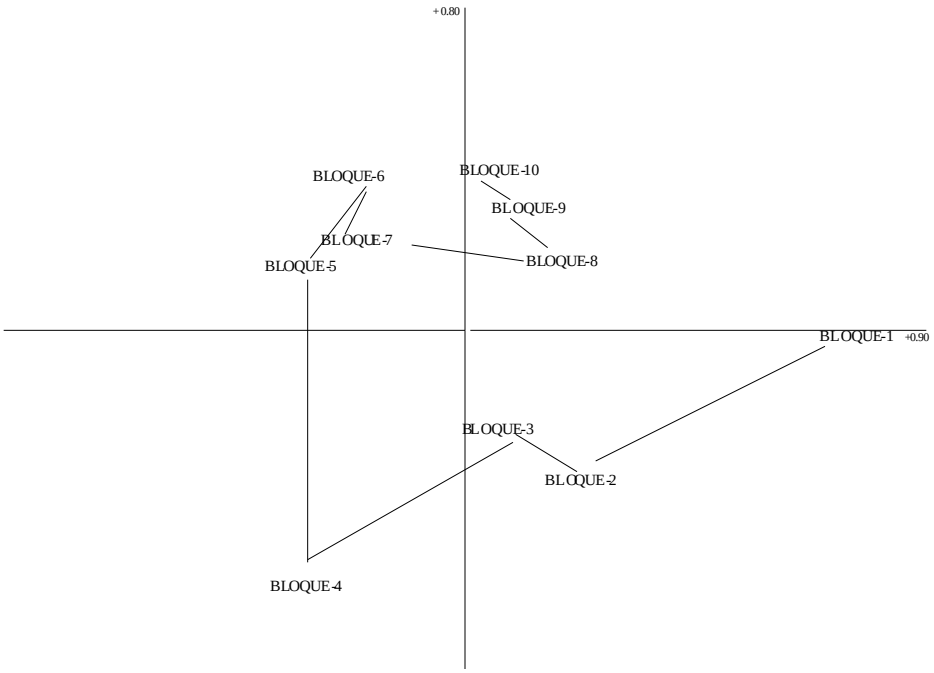

Figure 1. Plan principal de l'analyse de correspondances de la table Blocs_Mots Mots prononcés au moins cinq fois

La coupure de l'arbre, effectuée quand l'indice de distance effectue un saut important, permet d'obtenir quatre classes de blocs actifs et, évidemment, un cinquième classe avec le bloc 8. La première classe contient uniquement le bloc 1, la seconde les blocs 2 à 4 et le bloc 9, la troisième les blocs 5 à 7, la dernière le bloc 10. Il est intéressant de noter que, excepté le bloc 9, la partition respecte l'ordre temporel des blocs. La structure lue sur le premier plan factoriel est ainsi confirmée.

Avant d'étudier la segmentation ainsi obtenue et de rechercher les mots qui caractérisent chaque partie, il est intéressant de confronter ces résultats à ceux offerts par l'étude de la croissance du vocabulaire.

6. Croissance du vocabulaire et ruptures thématiques du réquisitoire

6.1 Modèle de partition du vocabulaire

La méthode d'analyse de la croissance du vocabulaire proposée par Hubert et Labbé (Labbé, 1993), à la suite des travaux de Ch. Muller, offre des critères pour déterminer les ruptures thématiques d'un corpus séquentiel. Ce modèle opère à partir de tous les différents mots prononcés par l'émetteur après avoir effectuer la lemmatisation du corpus. Bien que cette dernière opération ne soit pas réalisée dans ce cas, (en particulier, car nous voulons différencier le singulier et le pluriel de certains mots qui, comme nous l'avons déjà commenté avec *testigo* et *testigos*, peuvent jouer des rôles différents) nous pouvons considérer légitime l'application de ce modèle.

Selon ces auteurs, en tout moment de l'acte de discours, le locuteur dispose d'un vocabulaire "général" polyvalent, réserve de mots utilisés en toutes circonstances et de plusieurs vocabulaires locaux, spécialisés ou thématiques. L'appartenance d'un mot à l'un

ou l'autre des lexiques dépend du locuteur, ainsi que des circonstances de la communication.

La proportion *p* de vocabulaire spécialisé employé dans le texte permet de calculer la croissance marginale théorique du vocabulaire. Selon ce modèle, cette croissance marginale correspond à une exponentielle de multiplicateur inférieur à 1 qui dépend du rythme moyen d'accroissement du vocabulaire propre à celui-ci (Labbé 1993). Ce modèle correspond à une progression régulière; les déviations à la régularité théorique observées indiquent les ruptures thématiques du texte.

Il existe des périodes de croissances supérieures à la moyenne et d'autres de récession. Les premières correspondent à l'apparition d'un nouveau thème, les secondes à l'épuisement du thème en cours, ou bien au retour d'un thème traité antérieurement.

6.2 Estimation de la proportion de vocabulaire spécialisé dans le réquisitoire

En reprenant la proposition de Hubert et Labbé pour ajuster la courbe de croissance on obtient pour ce texte une valeur p=0.20. Cette valeur indique une spécialisation assez élevée de chacune des parties du réquisitoire, cela en tenant compte que ce type de discours obéit à de notables constrictions, et montre qu'il existe une stratégie claire dans la démonstration du procureur.

6.3 Etude de la déviation entre accroissement observé et accroissement espéré

La figure 2 visualise les écarts entre l'accroissement de vocabulaire ajusté et celui observé, ce qui permet de mettre en évidence les ruptures thématiques du corpus. A un point au dessus de l'axe, correspond une richesse de vocabulaire supérieure au niveau espéré à ce moment-là, et à un point au dessous, le contraire. La pente de chaque segment de courbe est aussi intéressante: une pente positive indique un apport de vocabulaire nouveau, une pente négative correspond à de nombreuses répétitions de mots déjà introduits. Les changements d'orientation permettent de segmenter le corpus en périodes: à chaque période correspond une courbe en forme de cloche, qui marque l'apport de vocabulaire propre suivi de son épuisement.

Dans ce réquisitoire, on observe un apport initial de vocabulaire légèrement inférieur à la valeur espérée, fait inhabituel, explicable certainement par la particularité du texte étudié. En effet, comme il est d'usage, le réquisitoire commence par introduire le cadre juridique. On observe à ce moment là un vocabulaire propre à un tribunal, très répétitif: *delito relativo a la prostitución* (délit relatif à la prostitution) *este delito queda acreditado* (ce délit est accrédité). Une arrivée massive de mots se produit tout au long du bloc 2; ce vocabulaire s'épuise progressivement; on peut noter un léger rebondissement autour du bloc 5, mais très rapidement la pente redevient très négative, et cette tendance ne change pas jusqu'au bloc 8, où quelque chose de nouveau se produit. L'unité thématique se rompt, la pente négative se trouve freinée - et si l'on adopte une division plus fine, en 20 blocs, on peut voir qu'à ce bloc correspond une légère croissance du vocabulaire, rapidement épuisée-.

On obtient ainsi deux grandes périodes, la première occupe 75% du réquisitoire et est formée de trois sous-thèmes: le début du réquisitoire (bloc 1), le développement du thème principal (blocs 2 à 6) avec une attention particulière prêtée à un point particulier (autour du bloc 5) et un final prolongé au cours duquel des mots précédemment employés sont repris. A un certain moment, autour du bloc 8, la tendance est rompue, un thème terminal arrive avec force, thème que le procureur ne réussira pas à épuiser.

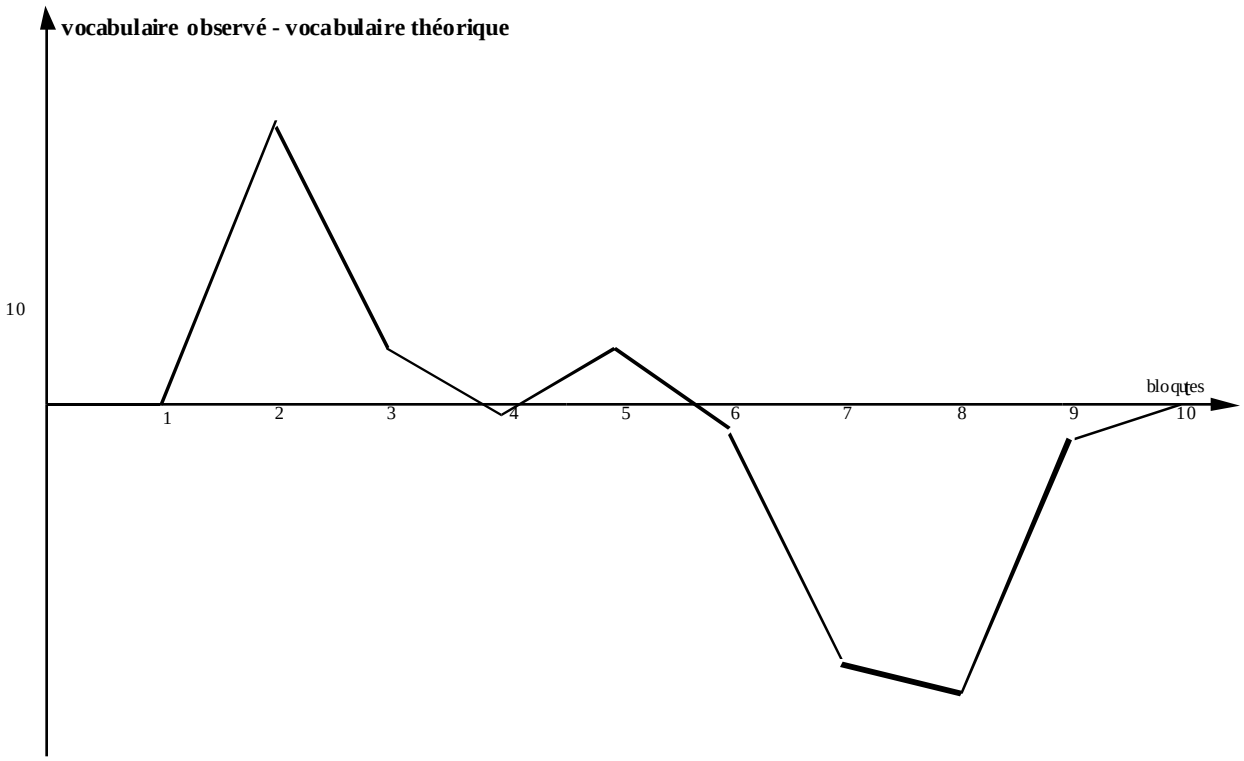

Figure 2. Accroissement marginal du vocabulaire
Déviations au modèle et ruptures thématiques

7. Comparaison des apports des deux méthodes

Les deux méthodes employées offrent une vision distincte du même phénomène: l'évolution du discours. Elles permettent d'étudier la structure temporelle du réquisitoire du procureur et apportent des éléments pour déterminer ses ruptures.

L'analyse de correspondances est très sensible aux phénomènes marqués, ce qui constitue aussi son point faible. C'est ce qui permet de mettre en évidence la très grande particularité du bloc 8, ainsi que la construction beaucoup plus relâchée des blocs qui suivent, en contraste avec la première partie du réquisitoire. Cette méthode permet aussi de suivre l'évolution du corpus, mettant en évidence les retours en arrière, les rapprochements entre parties séparées dans le temps

Le modèle d'accroissement du vocabulaire permet de situer clairement les ruptures thématiques quand celles-ci sont indiquées par un vocabulaire nouveau abondant. Comme

nous le verrons plus loin, la particularité du bloc 8 provient en partie d'un vocabulaire nouveau, mais peu abondant et, par contre, très répété, dans une espèce de bégaiement du procureur, très tendu et nerveux quand il décide de faire usage de cette arme fondamentale tout en ayant peur de la gâcher puisque bien fragile.

Il est très utile de comparer les résultats fournis par ces deux méthodes. Elles produisent des résultats complémentaires qui illustrent des aspects distincts et par là se renforcent mutuellement. Il est intéressant de rappeler que l'analyse des correspondances n'utilise pas les mots peu répétés, tandis que le modèle d'accroissement du vocabulaire opère a partir de l'intégralité des mots différents employés, et de souligner que la convergence des résultats obtenus ne fait que renforcer l'idée que les choix lexicaux sont liés et s'enchaînent.

8. Caractérisation des périodes thématiques

8.1 Retour au texte et délimitation des périodes

Il est possible d'affiner la délimitation des périodes: un retour au texte permet de mieux définir leurs frontières qui ne coïncident pas obligatoirement avec celles des blocs, qui ont été, rappelons-le, déterminées tout à fait arbitrairement, en recherchant uniquement des blocs de longueurs comparables.

Dans ce cas, il a été décidé de découper une première période composée du bloc 1 et du premier espace du bloc 2, période dédiée à argumenter le délit de proxénétisme. La deuxième période commence alors, attaquant la démonstration de l'exécution de l'assassinat par l'accusé; *ya en relación al asesinato* (maintenant relativement à l'assassinat) sont ses premiers mots. Elle englobe les blocs 2, 3 et partie du bloc 4. La frontière correspond à la fin de l'exposition des particularités si surprenantes de l'assurance vie.

Evidemment, il n'existe aucune délimitation inéquivoque entre périodes. Le découpage effectué est une étape nécessaire pour pouvoir rechercher les mots significatifs de chaque période et ainsi rendre compte de la structure du réquisitoire.

Ensuite vient la troisième période, une longue exposition des incohérences du comportement et des déclarations de, principalement, l'accusé inducteur. Cette période s'achève vers le milieu du bloc 7, quand, enfin, le procureur décide d'utiliser son argument le plus important, mais aussi le plus fragile, la déclaration du témoin direct, l'employée du sauna qui a entendu la conversation entre les accusés au cours de laquelle ils commentaient l'assassinat planifié. "*Pasamos a las declaraciones de M.F (...) le había oído decir que se iba a cargar a M.J.*" (Passons aux déclarations de M.F. (...) elle l'avait entendu dire qu'il allait liquider M.J.), voilà les mots qui marquent le début de cette période. Cette quatrième période est donc la partie où le procureur utilise l'unique atout de poids dont il dispose.'

Puis, une fois terminée l'exposition de ce témoignage et des incidents qui l'entourent, le procureur entame la période finale en apportant une information résiduelle, destinée à souligner de nouveau les contradictions et l'incohérence de la version de l'accusé.

Finalement, le réquisitoire se termine par l'évaluation des indices et la qualification qui résume les faits et l'argumentation tout en estimant la participation des accusés. Cette qualification arrive de façon assez brutale, au dernier espace discursif, exprimée par un vocabulaire précis qu'il n'y avait pas lieu d'employer auparavant, comme le propre mot *asesinato* (assassinat), par exemple.

8.2 Caractérisation des périodes du réquisitoire

Une fois la structure chronologique du corpus mise en relief , il est nécessaire de détecter les mots et expressions responsables de son évolution: l'étude des particularités lexicales de chaque période permettra de donner sens au découpage obtenu. En décrivant le découpage effectué, nous avons déjà résumé le contenu des périodes. De fait, ce contenu, et surtout certains aspects qui peuvent demeurer transparents à la lecture classique, se trouvent dévoilés par les caractéristiques lexicométriques des périodes. La table 3 montre, en prenant pour exemple la période 2, la principale information disponible sur le vocabulaire: a) la liste des spécificités positives ou mots sur-représentés de façon significative dans une période; b) les mots de cette liste qui sont des mots localisés (dont l'indice de répartition est inférieur ou égal à 0.40, voir le paragraphe 4.2 ci-dessus) qui sont les mots soulignés; c) finalement, les accroissements spécifiques correspondant à cette liste, c'est à dire les mots dont la fréquence dans la période est significativement supérieure à la fréquence dans le sous-corpus qui contient les périodes antérieures et la période étudiée elle-même, sont indiqués par une flèche ascendante.

Période 1

C'est la seule période pour laquelle le "langage du droit" et celui de l'enquête policière sont importants: *delito(délit), artículo (article), registro (registre), ley (loi).* L'expression *libertad sexual* (liberté sexuelle) appartient aussi à ce langage, en relation avec les délits contre la liberté sexuelle.

Le procureur commence son réquisitoire en exposant très rapidement les trois délits qu'il considère commis par les accusés. Puis il s'attaque au premier d'entre eux, le délit de proxénétisme qu'il nomme délit relatif à la prostitution. Pour démontrer la réalisation de ce premier délit, il n'y a guère de problème car l'existence et les installation du sauna-maison de passe montre quelle est, de toute évidence, son usage habituel. "*Queda acreditado*", (il est accrédité), "*otros elementos de prueba*" (autres éléments de preuve). Pour cette raison, il n'y a aucune raison de rechercher des données ou des indices, et les mots *dato* (donnée) et *indicios* (indices) ne sont pas employés, et *datos* (données) seulement une fois.

Période 2

Maintenent le procureur doit affronter la démonstration de l'assassinat "*ya en relación en relación al asesinato*" (maintenant relativement à l'assassinat) est la phrase qui ouvre cette période. Et les choses se présentent assez mal: "*no hay prueba de cargo*" (il n'y a

pas de preuve à charge), et bien que le procureur balaie ce problème d'un "l*o que no tiene mayor importancia*" (ce qui n'a pas grande importance), il ne doit pas moins s'efforcer de montrer qu'il existe "*un cúmulo de indicios*" (un cumul d'indices), s'appuyer sur la jurisprudence qui reconnait que les délits ne se commettent habituellement pas au grand jour et en public et qu'il est donc suffisant de réunir un ensemble de données, clairement reliées entre elles de telle façon que le délit en soit rationnellement déduit. Selon le procureur il est nécessaire que "*exista un nexo causal y lógico según las reglas del criterio humano (...) para establecer esa relación entre la pluralidad de indicios (...) y el hecho que se va a deducir"* (qu'il existe un lien causal et logique selon les règles du critère humain (...) pour établir cette relation entre la pluralité d'indices (...) et le fait qui va en être déduit). Pour cette raison, le procureur décrit le cadre juridique, les articles de loi sur lesquels il s'appuie. Ensuite, il décrit l'assurance-vie et toutes ses particularités. Tout cela lui permettra ensuite de montrer par le raisonnement appuyé sur des faits avérés que la souscription de l'assurance-vie est suspecte et fait partie d'un montage criminel. En quelque sorte, cette période construit le cadre juridique - il n'est pas nécessaire d'avoir de preuve directe - et le cadre factuel autour duquel il va bâtir son raisonnement qui lui permettra de montrer que les accusés sont coupables.

Table 3
Information lexicale disponible pour chaque période
Mots localisés, mots spécifiques, et accroissements spécifiques
Exemple de la période 2

	Libellé du mot	Accr. spécif.	Fréquence interne	Fréquence globale	Proba.
Période 2					
1	vida	_	18	26	0.000
2	capital		7	7	0.000
3	caso		12	17	0.000
4	millones	_	10	13	0.000
5	beneficiarios	_	9	11	0.000
6	seguro	_	24	51	0.000
7	los		43	113	0.000
8	jubilación		7	8	0.000
9	prueba		10	15	0.000
10	cargo		5	5	0.001
11	devolver		5	5	0.001
12	confidencial		5	5	0.001
13	son	_	12	22	0.001
14	mayor		8	12	0.001
15	tribunal		6	8	0.002
16	hijos	_	11	21	0.003
17	póliza		5	6	0.003
18	suscribió		5	7	0.008
19	dos		8	15	0.009
20	según		7	13	0.013
21	muerte		7	13	0.013
22	de		153	587	0.017
23	tal		5	6	0.017
24	uno		11	26	0.018
25	señor		9	20	0.021

Période 3

L'assurance-vie a été expliquée dans tous ses détails. Le procureur va maintenant apporter des données - pour cette raison *dato* (donnée) est un mot dont la fréquence augmente de nouveau dans cette période- qui montrent l'incohérence du comportement des accusés, et surtout de "l'accusé".

En effet, le patron, inducteur du crime et principal bénéficiaire est placé au centres des débats. Sa conduite, ses déclarations, ce qu'il fait ou ne fait pas, ce qu'il devrait savoir sur la victime, qu'il prétend être sa compagne sentimentale, mais ne sait pas sont autant d'éléments que le procureur cherche à utiliser, à relier logiquement pour montrer l'invraisemblance de la version présentée par la défense.

Les mots spécifiques –par exemple, *sabía* (savait), *no* (ne...pas), *relación* (relation), *conocía* (connaissait), *existencia* (existence) – proviennent de ce raisonnement qui fait allusion à l'étrange comportement du patron. La famille du patron du sauna-relax - avec le mot *familia (famille)* et le nom de sa femme, reproduit ici sous les initiales SRT - est évoquée, pour montrer que là est sa vraie famille, sa vraie relation sentimentale.

Il est intéressant de souligner que *familia* (famille) et *hijos* (enfants) ont des sens bien distincts. Le premier terme est toujours une allusion à la famille de l'accusé, le second, le plus fréquemment, une allusion aux enfants de la victime.

Période 4

Le procureur introduit le témoignage de l'employée, non ratifié lors du procès mais si vraisemblable dans sa version initiale. Le procureur emploie alors des mots nouveaux, peu nombreux mais très répétés, mais doit s'appuyer sur certains mots du thème antérieur - de la *declaración(es) (*déclaration(s)*)* des accusés à la *declaracion(es)* (déclaration(s)) de ce témoin- pour rendre plus facile la transition à ce nouveau thème, inquiétant d'une certaine façon. En effet, l'enjeu est crucial, il s'agit de la pièce-clé de son accusation. La tension s'élève, les phrases ne se terminent pas, la dislocation du discours est importante, les répétitions excessives et concernent souvent des mots introduits au cours de ce bloc même. Le ton du procureur, de fait, change. Il adopte une position plus solennelle, il doit marquer qu'il est le *Ministerio Fiscal* (Ministère Public) et se nomme lui-même ainsi plusieurs fois.

Les mots les plus localisés (table 2) sont des mots liés à la déclaration du témoin qui s'est rétracté: *oido (entendu), protección (protection), policía (police), frente* (face), *testigo (témoin)*. En effet, ce témoin, lors de sa première déclaration, celle qui impliquait les accusés en déclarant avoir entendu la conversation citée, avait demandé la protection de la police. Ensuite, la défense avait cherché à détourner cette demande, en une demande de protection contre la police qui, disait-elle, l'avait forcée à dénoncer ses patrons. Le procureur essaie de montrer l'absurdité de *pedir protección a la policía contra la policía* (demander protection à la police contre la police), l'un des arguments employés au cours de la réfutation. Ce moment si important a rompu le développement régulier du discours. Dans ce qui suit, cette régularité ne se récupère pas.

Période 5

Au bloc 9, la pente positive (figure 2) indique l'apparition d'un nouveau vocabulaire relativement important, mais l'analyse de correspondances montre un retour en arrière, c'est à dire l'emploi de mots qui appartiennent aux thèmes initiaux. De fait, le procureur revient sur des points déjà traités, mais pour apporter une nouvelle information et, par exemple, *otro dato más (*une autre donnée*)* constitue un refrain de cette partie; le procureur parle à nouveau de la victime, mais pour préciser que c'était une malade mentale, nouvel argument en faveur de l'invraisemblance de la relation entre celle-ci et le patron du sauna.

Tout au long de ce bloc et d'une partie du suivant, de fait jusqu'à la qualification finale, le procureur apporte de nouveaux détails mais, de fait, mineurs, qui soulignent son argumentation. Il le fait dans un certain désordre, sans conserver un clair contrôle de son discours.

9. Conclusion

Pour conclure, disons en premier lieu que le juge a suivi la thèse du procureur et condamné les accusés à de lourdes peines de prison; cette condamnation a été confirmée en deuxième instance par le tribunal suprême. Précisons aussi que le procureur avait très peu d'espoir d'obtenir cette condamnation et que ses collègues du parquet lui avaient prédit qu'il perdrait le cas. Il avait contre lui le manque de preuves, la qualité de prostituée de la victime et, en contraste, le pouvoir financier de l'assassin qui avait contracté l'un des meilleurs avocats dans le domaine pénal.

La conclusion du procès constitue certainement la reconnaissance de l'adéquation de la stratégie du procureur à l'objectif poursuivi. Malgré une certaine désorganisation de la dernière partie, le réquisitoire constitue un texte construit et structuré et l'information est apportée de manière progressive et organisée. On peut aussi noter un notable contrôle des choix lexicaux ; on peut indiquer, pour exemples, la répartition stable des mots qui indiquent la conviction ou par la claire différentiation entre les termes *familia* (famille) pour faire allusion à la famille de l'accusé et *hijos* (fils), pour mentionner les enfants de la victime.

L'emploi de méthodes statistiques a pour but de mettre en valeur de nouveaux aspects d'un corpus, d'en multiplier les lectures. L'analyse d'un texte unique offre des caractéristiques particulières. Les comparaisons sont internes, elles s'effectuent entre les parties du texte pour, en particulier, rechercher les moments de particulière signification et obtenir des éléments révélateurs de son élaboration et de sa logique interne.

La segmentation en périodes permet non seulement de rendre compte des divers thèmes abordés, et de l'extension qui leur est consacrée, mais aussi de retrouver le rythme du discours et de détecter, par exemple, les thèmes que le locuteur a du mal à aborder ou, au contraire, les moments pendant lesquels le discours est plus fluide.

Il serait évidemment très utile et intéressant de pouvoir comparer les résultats avec ceux qui seraient obtenus par l'analyse d'autres réquisitoires de même ampleur.

Malheureusement, le coût du recueil de ce type de données et la nécessité d'obtenir l'accord des différentes parties impliquées dans un procès ne facilitent pas cette possibilité.

Références

Bécue M. (1992), *Análisis de Datos Textuales: Métodos y Algoritmos*. Paris: Cisia.

Casanovas P., Ardévol E., Cachón M., Riba C. (1995). *Videos als Tribunals de Justícia.* Barcelona: Publicacions ICE, UAB.

Hubert P., Labbé D. (1990), La répartition des mots dans le vocabulaire présidentiel, *Mots*, 22, pp. 80-92

Labbé D. (1993), Un modèle d'analyse du vocabulaires, in: *Actes des secondes journées internationales d'analyse statistique de données textuelles*, Montpellier..

Lebart L., Morineau A., Bécue M., Hauesler L. (1992).*SPAD.T, Système Portable pour l'Analyse des Données Textuelle*s, Saint Mandé: Cisia.

Lebart L., Salem A. (1994), *Statistique Textuelle*, Paris, Dunod.

Muller Ch. (1977), *Principes et méthodes de statistique lexicale*, Paris, Hachette.

Salem A. (1991), La lexicométrie chronologique in: *4º colloque de Lexicologie Politique.* Paris.

Salem A. (1993), Méthodes de la statistique textuelle. Thèse d'Etat, Université Sorbonne-Nouvelle.

Labbé D., Serant D., Thoiron Ph., (1988), *Etudes sur la richesse et la structure lexicales.* Champion-Slatkine, Paris-Genève.