Outils d'aide à l'exploitation d'entretiens semi-directifs : étude de l'interaction entre intervieweur et interviewés sur un corpus ethnoécologique

Julien Bonneau

BCL – Nice Sophia Antipolis – Nice – France

Abstract

This paper will present a combination of textual data processing using ALCESTE and LEXICO software in order to solve problems of interactions between an interviewer and interviewees in semi-directive interviews. The corpus is a collection of semi-directive interviews done with the staff of Cévennes National Park, in the context of a doctoral thesis in ethnoecology. The paper will show the data formatting according to TEI and XML rules, which allow both their transformation in specific software formats used for the analysis, and their restructuring to differentiate speakers. Finally, we will present the different corpora's processing accomplished with ALCESTE and LEXICO software to obtain a linguistic, thematic and significant data "cartography" adapted for improving the analysis of semi-directive interviews as well as their interaction between an interviewer and interviewees.

Résumé

Cet article présente une combinaison de traitements, issus des logiciels d'analyse de données textuelles ALCESTE et LEXICO, pour traiter du problème de l'interaction entre intervieweur et interviewés dans des entretiens semi-directifs. Le corpus est une collection d'entretiens semi-directifs réalisés sur la région du Parc national des Cévennes durant un travail de thèse en ethnoécologie. L'article présente la formalisation des données selon les recommandations TEI et XML pour leur transformation au format propriétaires des logiciels d'analyse utilisés et leur restructuration décroisant les tours de parole des locuteurs. Il décrit ensuite les différents traitements réalisés sur ces corpus par les logiciels ALCESTE et LEXICO pour obtenir une « cartographie » d'indices linguistiques, thématiques et énonciatifs, propres à améliorer l'analyse d'entretiens semi-directifs et l'interaction qu'ils mettent en œuvre.

Mots-clés : entretiens semi-directifs, émergence d'une partition à partir d'une hiérarchie, formalisation de données, analyse de contenu, interaction discursive autour de questions ouvertes, linguistique exploratoire.

Introduction[1]

Dans le cadre d'une collaboration avec R. Dumez, chercheur en ethnoécologie au Muséum National d'Histoire Naturelle, nous avons utilisé des méthodes de statistiques textuelles (Lebart, Salem, 1994) pour proposer des dispositifs d'aide à l'analyse d'entretiens semi-directifs, afin de donner aux ethnologues des moyens de critiquer les résultats obtenus par leur direction d'entretiens.

[1] Merci à C. Fabre et L. Tanguy, du laboratoire ERSS à l'Université Toulouse le Mirail, qui m'ont encadré pour ces recherches et à R. Dumez, qui m'a cédé son corpus et sans qui ce travail n'aurait pu être réalisé.

La première partie de notre exposé présentera les particularités du corpus et la formalisation, aux formats TEI et XML et propriétaires, nécessaire à son exploitation par des traitements automatiques.

Nous exposerons ensuite un premier résultat qui posera la question, à l'aide du logiciel ALCESTE, de la pertinence de la partition contrastive du corpus proposée *a priori* par la démarche ethnologique.

La dernière section de cet article proposera une grille d'analyse, articulée autour des logiciels ALCESTE et LEXICO, permettant de mettre en valeur des interactions entre intervieweur et interviewés lors de la réalisation des enquêtes de terrain, le but étant de quantifier et critiquer la posture théorique de l'intervieweur lors de sa conduite d'entretiens et de l'articuler avec les réponses fournies par les interviewés. Cette question est d'autant plus centrale que le manque de rigidité des entretiens semi-directifs interroge sur la comparabilité des résultats ainsi obtenus.

1. Le corpus

Les réalités du terrain, les contraintes temporelles, la recherche d'informations spontanées ou suffisamment élargies, président souvent au choix du type d'enquêtes qui s'offre au chercheur lors d'une étude sur une population donnée (Grangé, Lebart, 1993). Dans le cadre d'une thèse en ethnoscience (Dumez, 2004), R. Dumez, chercheur au Muséum National d'Histoire Naturelle, a retenu l'entretien semi-directif, pour interviewer la population en activité sur le Parc National des Cévennes (PNC). Le but de ses recherches étant de mettre en valeur les diverses représentations de cette population d'un même espace (le PNC), l'entretien semi-directif lui est apparu comme l'outil permettant « un questionnement suffisamment large pour cerner autant que possible l'existence de catégories (lexicales) » (Dumez, 2004). Il permet la production de textes libres à l'interviewé, susceptibles de contenir des informations étendues sur la problématique de départ, des réponses auxquelles le chercheur n'aurait pas pensé *a priori*, faisant apparaître des « catégories » qui sont, selon R. Dumez, autant de marqueurs des représentations des divers groupes d'acteurs en présence.

Nos travaux s'intégrant dans le cadre d'une analyse secondaire, le matériau de base de nos analyses est une collection de 66 entretiens semi-directifs réalisés par R. Dumez auprès de la population du PNC. A partir de ces entretiens, nous avons constitué un corpus au format XML (Harold, Means, 2005) selon les recommandations de la TEI (Burnard et al., 1996), pour permettre sa conversion, à l'aide de scripts Perl, aux formats propriétaires des différents logiciels utilisés pour cette étude, ALCESTE et LEXICO. Dans le corps de texte du corpus sont explicités, à l'aide de balises, les tours de parole des locuteurs, ainsi que les informations nécessaires à la caractérisation des entretiens, les caractères sociaux des divers locuteurs, ajoutés dans un texte liminaire précédent chaque entretien.

Voici un exemple du corps de texte :

```
  <p>
<sp who='RICHARD DUMEZ'>Donc euh, donc en fait juste vous êtes nouvelle sur le
poste...</sp>
  </p>
  <p>
<sp who='M. B.'>Oui</sp>
  </p>
  <p>
<sp who='RICHARD DUMEZ'>Donc euh, C'est juste pour savoir quelle formation vous
avez, c´est pour cibler...</sp>
```

Ces aménagements doivent permettre d'accéder, de manière automatique, au discours de chaque locuteur et de le caractériser par les caractères sociaux de ceux-ci. Les caractères retenus sont : la catégorie socioprofessionnelle (éleveur, banquier, grossiste en matériel agricole, etc.), d'éventuelles activités secondaires (maire, responsable d'association, etc.), l'âge, le sexe, la situation familiale, l'appartenance à la population active/passive, la région d'origine et le lieu de domicile de l'individu.

Nous présentons ici, pour l'exemple, la caractérisation de l'intervieweur :

```
<add resp='Julien Bonneau'>
<label><rs type='person'>RICHARD DUMEZ</rs></label><item><list
type='simple'><head>caractères sociaux</head>
   <label>catégorie(s) socioprofessionnelle(s)</label><item><list
type='simple'><item>ethnoscientifique</item><item>étudiant</item></list></item>
   <label>activité(s) secondaire(s)</label><item>non</item>
   <label>sex</label><item>masculin</item>
   <label>âge</label><item>20-30ans</item>
   <label>situation familiale</label><item>célibataire</item>
   <label>population active/passive, etc.</label><item>active</item>
   <label>lieu de domicile</label><item>Paris</item>
   <label>région origine</label><item>Centre</item></list></item></list>
```

A partir de ce corpus de départ, nous avons extrait diverses nouvelles configurations, selon les besoins de nos analyses en termes de textes et de partitions retenus. Nous avons fait contraster les variables et le contenu textuel de ces nouveaux corpus pour articuler entre eux les discours de chaque groupe sociaux, ainsi que le discours de l'intervieweur vis-à-vis de ceux-ci. Nous présentons dans la suite les corpus obtenus.

1.1. Les corpus au format ALCESTE

Nous avons formaté deux corpus pour leur exploitation à l'aide du logiciel ALCESTE :

Le **corpus complet ALCESTE** est la transcription complète du corpus de départ au format propriétaire ALCESTE. Pour chaque entretien, les tours de paroles de chaque individu ont été regroupés sous des étiquettes, ou « mots étoilés », décrivant leurs caractères sociaux. Les différentes étiquettes utilisées sont les suivantes : **CSP* (catégorie socioprofessionnelle dont les valeurs possibles sont E – éleveur, G – gestionnaire, SP – sapeur pompier, ETH – ethnologue, D – autres), **AS* (activités secondaires), **A* (âge, par tranches de 10 ans), **S* (sexe), **SF* (situation familiale), **PAP* (population active/passive), **RO* (région d'origine, Cévennes – C – ou autre – H), **LD* (lieu de domicile), chacune de ces étiquettes recevant les valeurs spécifiques correspondant aux individus.
Pour chaque entretien, nous avons ainsi obtenu le discours de chaque individu caractérisé socialement. Nous avons, par exemple, pour un entretien de l'intervieweur (R. Dumez) dont nous avons présenté les caractères précédemment :

```
*CSP__ETH *CSP__D *AS__N *S__M *A__20_30 *SF__20_30 *PAP__A *LD__P *RO__H
```

(suivi de l'ensemble de la transcription de ses interventions dans cet entretien)

L'ensemble des traitements ainsi réalisés à partir du corpus de départ forme le corpus complet ALCESTE.

Le **corpus Q ALCESTE** (corpus Questions ALCESTE) contient les interventions de l'intervieweur au travers de l'ensemble des entretiens. Il est caractérisé à l'aide des « mots étoilés » correspondant à la catégorie socioprofessionnelle à laquelle l'intervieweur s'adresse

(**CSP*), la partition par catégorie socioprofessionnelle s'étend révélée comme la plus pertinente au cours de notre étude[2]. Les différentes valeurs possibles des « mots étoilés », servent cette fois à indiquer la présence d'un individu d'une profession donnée lors de la réalisation d'un entretien. Ainsi, ce corpus contient l'ensemble du questionnement auquel a été confrontée chaque catégorie socioprofessionnelle au cours des entretiens.

1.2. Les corpus au format LEXICO

Contrairement au logiciel ALCESTE, LEXICO ne permet pas d'étiqueter de plusieurs valeurs les caractères sociaux pour un individu donné. Ainsi, la partition des corpus utilisés par LEXICO diffère de celle d'ALCESTE en cela que les individus dont un caractère social prend plusieurs valeurs (par exemple, pour le caractère socioprofessionnel, éleveur et gestionnaire) doivent être regroupés sous une balise spécifique à cette mixité, autre que les balises décrivant les individus ne possédant qu'une seule de ces valeurs pour ce caractère (d'une part les éleveurs, d'autre part les gestionnaires).

Pour nos analyses nous avons construit cinq corpus au format LEXICO :

Le **corpus Q LEXICO** (corpus Questions LEXICO) contient l'ensemble du questionnement auquel l'intervieweur a confronté les individus des catégories socioprofessionnelles éleveur et gestionnaire au cours des entretiens. La partition est la suivante : *ELV* (questionnement aux éleveurs), *GEST* (questionnement aux gestionnaires du PNC), *ELV_GEST* (questionnement aux éleveurs, gestionnaires du PNC).

Le **corpus QR LEXICO** (corpus Questions-Réponses LEXICO) contient l'ensemble du questionnement aux éleveurs, aux gestionnaires du PNC et aux éleveurs et gestionnaires, ainsi que l'ensemble de leurs réponses. La partition proposée est *ETHNO_ELV* (questionnement aux éleveurs), *ETHNO_GEST* (questionnement aux gestionnaires), *ETHNO_ELV_GEST* (questionnement aux éleveurs et gestionnaires), *ELV* (réponses des éleveurs), *GEST* (réponses des gestionnaires du PNC), *ELV_GEST* (réponses des éleveurs et gestionnaires).

Le **corpus QR éleveurs LEXICO** contient les questions et réponses des entretiens entre R. Dumez et les individus appartenant au groupe socioprofessionnel éleveur. La partition proposée est *ETHNO_ELV* (questions aux éleveurs), *ELV* (réponses des éleveurs), *ETHNO_ELV_GEST* (questionnement de Dumez aux éleveurs et gestionnaires) et *ELV_GEST* (réponses des éleveurs et gestionnaires du PNC).

Le **corpus QR gestionnaires LEXICO** contient les questions et réponses des entretiens entre R. Dumez et les individus appartenant au groupe socioprofessionnel gestionnaire. La partition est *ETHNO_GEST* (questionnement aux gestionnaires), *GEST* (réponses des gestionnaires du PNC), *ETHNO_ELV_GEST* (questionnement de Dumez aux éleveurs et gestionnaires) et *ELV_GEST* (réponses des éleveurs et gestionnaires du PNC).

Le **corpus R LEXICO** (corpus Réponses LEXICO) contient l'ensemble des réponses des éleveurs, des gestionnaires du PNC et des éleveurs et gestionnaires à l'intervieweur. La partition proposée est *ELV* (réponses des éleveurs), *GEST* (réponses des gestionnaires du PNC), *ELV_GEST* (réponses des éleveurs et gestionnaires du PNC).

[2] Voir partie 2.

2. Validation de la partition ethnologique du corpus à l'aide du logiciel ALCESTE

ALCESTE (Reinert, 2002) est un logiciel d'analyse statistique de données textuelles d'inspiration benzécrienne, développé par Max Reinert, pour le dépouillement et l'analyse de contenu de textes libres. Il propose une méthodologie automatique « clef en main », qui vise à faire émerger l'organisation morpho-lexicale d'un corpus en fonction des caractérisations externes faites de celui-ci.

Nous avons indiqué et caractérisé les unités initiales (ou unités de contexte initiales, u.c.i.), c'est-à-dire les différentes unités qui composent la collection qu'est le corpus à l'aide des « mots étoilés » (voir 1.2.). Les « mots étoilés » sont les variables des unités de contextes et des regroupements qu'en fait ALCESTE.

Dans un premier temps, le logiciel fait subir une opération de troncature à l'ensemble des formes du corpus afin de regrouper les formes morphologiquement apparentées (sur l'argument d'une parenté sémantique) et de limiter leur diversité (ce qui facilite d'autant l'analyse statistique). Il obtient ainsi les formes réduites ou tronquées. Cette opération présente l'inconvénient de faire perdre l'information précise de la forme et du vocabulaire employés, mais garde une information suffisante pour déterminer, d'un point de vue thématique, de quoi « parle » le corpus. ALCESTE exclut également de son analyse un certain nombre de formes, mots outils entre autres, jugées non pertinentes pour une analyse thématique des données afin d'améliorer, lisser les résultats de son traitement.

ALCESTE construit ensuite des unités de contexte élémentaire (u.c.e.), c'est-à-dire des segments de texte de tailles comparables, généralement la phrase, dont la composition interne, le profil en terme de formes tronquées présentes ou absentes et leur association aux caractères externes qui les décrivent, sert d'unité de comparaison (par classification hiérarchique descendante maximisant le critère du χ^2 du tableau des marges, c'est à dire de façon à ce que les deux classes obtenues à chaque étape soient au maximum indépendantes) pour faire émerger la structure thématique du discours du corpus sous forme de classes d'u.c.e. Ainsi, en utilisant la distance du χ^2, ALCESTE permet de déterminer comment les formes tronquées et les « mots étoilés » s'articulent autour des classes obtenues et, plus particulièrement, quelles formes et quels « mots étoilés » sont les plus caractéristiques de chaque classe.

Nous avons confronté le **corpus complet ALCESTE** à cette méthodologie. ALCESTE a ainsi créé 7 625 u.c.e. et six classes. Le nombre de mots « étoilés » (correspondant à des caractères distincts) pris en compte pour décrire les u.c.i. est de 59. Pour les six classes obtenues, quatre d'entre elles sont très fortement corrélées avec les caractères socioprofessionnels des individus. La classe 1 avec les sapeurs pompiers (χ^2 = 371.23, en deuxième position des « mots étoilés » les mieux corrélés pour cette classe), la classe 4 avec l'ethnologue (χ^2 = 357.44, en première position des « mots étoilés » les mieux corrélés), la classe 5 avec les gestionnaires (χ^2 = 715.30, en première position) et la classe 6 avec les éleveurs (χ^2 = 647.03, en première position).

Les caractères socioprofessionnels éleveurs, gestionnaires et sapeurs pompiers correspondent à la description contrastive des différents groupes d'acteurs sur le PNC décrite, dans ses travaux, par R. Dumez. Les résultats obtenus par l'analyse d'ALCESTE confirment la pertinence de ce découpage.

3. Mise en valeur d'interactions entre intervieweur et interviewés à l'aide des logiciels ALCESTE et LEXICO 3

Si les résultats présentés dans la partie précédente confirment la partition proposée par l'analyse ethnologique des données du corpus et sont donc un argument supplémentaire pour la validation de l'expertise de R. Dumez, le choix du recueil des données, les entretiens semi-directifs, amène à se poser la question de la comparabilité des entretiens entre eux. Le découpage en groupe socioprofessionnel étant fixé *a priori* par l'ethnologue, ceci n'a-t-il pas influé lors de la réalisation de l'enquête de terrain, l'intervieweur n'a-t-il pas cherché, même de manière inconsciente, à différencier ces groupes sociaux ? Ces représentations des uns et des autres n'ont-elles pas influé sur la conduite des entretiens ? Dans la suite de cet exposé, loin d'espérer répondre de manière tranchée à ces interrogations, nous allons tâcher de donner des moyens à l'intervieweur ethnologue de critiquer sa conduite d'entretiens par des mesures quantifiées, de lui permettre de confronter sa posture théorique d'intervieweur à la réalité de marques énonciatives et thématiques de l'interaction entre intervieweur et interviewés portée par le corpus.

3.1. Interaction thématique

3.1.1. Un premier indice d'interaction thématique à l'aide d'ALCESTE

En plus des résultats présentés dans la partie II de cet exposé, l'analyse du **corpus complet ALCESTE** par le logiciel ALCESTE fournit aussi les formes tronquées les plus représentatives de chaque classe, présentées dans le tableau 1. Il donne ainsi accès à des marques lexico-thématiques caractéristiques de chacun de ces groupes socioprofessionnels.

Nous avons confronté ces résultats à ceux obtenus par l'analyse à l'aide d'ALCESTE du **corpus Q ALCESTE**, c'est-à-dire au questionnement de l'ethnologue partitionné socioprofessionnellement. ALCESTE a créé 2 237 u.c.e. et quatre classes dont trois sont bien corrélées avec les caractères socioprofessionnels des individus interviewés : la classe 1 représentative des sapeurs pompiers ($\chi^2 = 133$), la classe 3 représentative des éleveurs ($\chi^2 = 151$) et la classe 4 représentative des gestionnaires ($\chi^2 = 112$). Les caractères socioprofessionnels ont été jugés représentatifs d'une classe si le « mot étoilé » correspondant apparaissait dans les cinq éléments (formes réduites et « mots étoilé ») les mieux corrélés à la classe. Les formes réduites caractéristiques de ces trois classes sont présentées dans le tableau 2.

La première information, que confirment ces résultats, est le fait que R. Dumez n'utilise pas le même lexique et ne parle pas des mêmes thèmes avec les uns et les autres des groupes d'acteurs/locuteurs. Il parle plus de feu avec les sapeurs pompiers (*feu+*, *pompier+*, *ecobu+*, *Brul+er*, *allum+er*), d'élevage et de la nourriture du bétail avec les éleveurs (*prairie+*, *bélier+*, *brebis*, *agnelage+* et *mang+er*, *céréale+*, *granule+*, *foin+*, *avoine+*) et de gestion et de la zone géographique du Parc national des Cévennes avec les gestionnaires (*causs+*, *parc+*, *exploit+er*, *gestion<*, *zone+*).

Classe 1 sapeurs pompiers		**Classe 4 ethnologue**		**Classe 5 gestionnaires**		**Classe 6 éleveurs**	
χ^2	Formes réduites	χ^2	Formes réduites	χ^2	Formes réduites	χ^2	Formes réduites
855.09	pompier+	644.87	mot+	468.18	parc+	541.92	mang+er
669.19	allum+er	552.51	fournel+	402.97	agricult<	541.19	genet+
566.65	brul+er	401.02	mont+	299.64	zone+	453.07	herb<
554.07	feu+	323.01	ecobu+	242.98	chambre+	362.07	pouss+er
466.77	vent+	285.45	lozer+	239.04	gestion<	284.18	repouss+er
251.62	ecobu+	269.19	brul+er	203.54	milieu+	181.28	bete+
242.09	eteindre.	262.86	parl+er	170.38	ouvert+	148.37	ronce+
187.53	par+er	247.29	feu+	160.37	act+ion	143.64	fauch+er
171.31	danger+	230.95	tas	155.15	contrat+	138.99	fougere+
165.39	neig+e	222.07	chataigner+	150.03	projet+	133.90	gyrobroi+
158.37	echapp+er	205.56	bogues	143.52	niveau+	129.92	coup+er
139.55	fois	177.09	entendre.	140.56	plan+	126.29	mois
133.75	voisin<	174.43	fournell+	130.72	dossier+	116.86	balai+
133.72	arret+er	157.82	techn+16	117.14	centrale+	111.69	vache+
105.52	gendarm+e	155.71	utilis+er	112.68	agricole+	111.19	plante+

Tableau 1 : corpus complet ALCESTE, formes réduites les plus représentatives des classes 1, 4, 5 et 6

Classe 1 Questions aux pompiers		**Classe 3 Questions aux éleveurs**		**Classe 4 Questions aux gestionnaires**	
χ^2	Formes réduites	χ^2	Formes réduites	χ^2	Formes réduites
332.67	feu+	196.48	mang+er	154.58	causs+
194.77	pompier+	192.10	prairie+	148.20	can+
158.47	ecobu+	154.63	belier+	117.26	parc+
98.35	accord	147.22	lutte+	103.51	prime+
60.73	brul+er	141.85	brebis	92.70	ovin+
41.28	allum+er	135.64	cereale+	92.12	exploit+er
37.26	region+	135.44	achet+er	73.04	chambre+
36.40	surface+	131.48	agnelage+	72.13	gestion<
35.82	apprendre.	124.13	granule+	66.04	niveau+
35.57	lozere	120.12	sortir.	59.62	propriete+
32.91	mot+	116.75	foin+	56.95	lait
31.81	techn+16	77.57	artific<	56.63	aide+
31.38	pyrenees	73.98	printemps	53.41	embroussaille+
30.78	dire+	73.98	avoine+	52.48	zone+
30.60	reglement+er	72.81	naturel+	48.86	viande+

Tableau 2 : Les formes réduites les plus représentatives des classes 1, 3, 4 du corpus Q ALCESTE

Nous avons souligné et mis en italique, dans les deux tableaux 1 et 2, les formes réduites communes aux classes correspondant à un groupe socioprofessionnel et au questionnement qui lui est associé. Ces premiers résultats montrent des phénomènes de reprises lexico-thématiques entre intervieweurs et interviewés autour de formes réduites. Ces reprises semblent ne pas être utilisées dans des proportions comparables par l'intervieweur d'un groupe à l'autre : on a qu'une seule forme réduite commune aux deux classes se rapportant aux questions aux éleveurs et leurs réponses parmi les quinze premières formes réduites les plus représentatives. Par contre, on voit apparaître cinq formes réduites communes sur quinze dans les classes associées aux questions/réponses aux gestionnaires et aux sapeurs pompiers.

3.1.2. Aller plus loin grâce à LEXICO

Pour tenter d'expliciter et de confirmer les résultats révélés par ALCESTE, nous avons eu recours au logiciel LEXICO (Salem et al., 2003) et plus particulièrement à son traitement des spécificités. LEXICO permet de réaliser des études contrastives entre sous-parties d'un corpus, notamment grâce aux spécificités, indices probabilistes basés sur la loi hypergéométrique, qui rendent compte de la sur- ou sous-représentation d'une forme ou d'un groupe de formes dans les différentes sous-parties du corpus.

Nous avons dans un premier temps constitué les groupes de formes, c'est-à-dire des regroupements de formes qui seront considérées comme une même unité statistique par LEXICO, correspondant aux formes réduites décrites par ALCESTE : les formes réduites soulignées et en italiques des classes 3 et 4 du tableau 2 (les formes réduites utilisées par l'intervieweur et communes aux interviewés) ; et les formes réduites des classes 5 et 6 du tableau 1 (les huit formes réduites repérées par ALCESTE comme les plus caractéristiques des éleveurs et des gestionnaires).

Pour simplifier notre étude, dans la suite de nos expertises, nous nous sommes uniquement intéressés aux groupes éleveurs et gestionnaires. La première analyse réalisée à l'aide de LEXICO a été effectuée sur le **corpus QR LEXICO**, le but étant d'observé si le calcul des spécificités des formes réduites mises en valeur par ALCESTE, c'est-à-dire les formes réduites communes aux classes 5, 6 (gestionnaires et éleveurs) du tableau 1 et 3, 4 (questions aux éleveurs et aux gestionnaires) du tableau 2, était, ou non, confirmé par LEXICO.

Graphique 1 : formes réduites communes aux questionnements aux éleveurs et aux gestionnaires

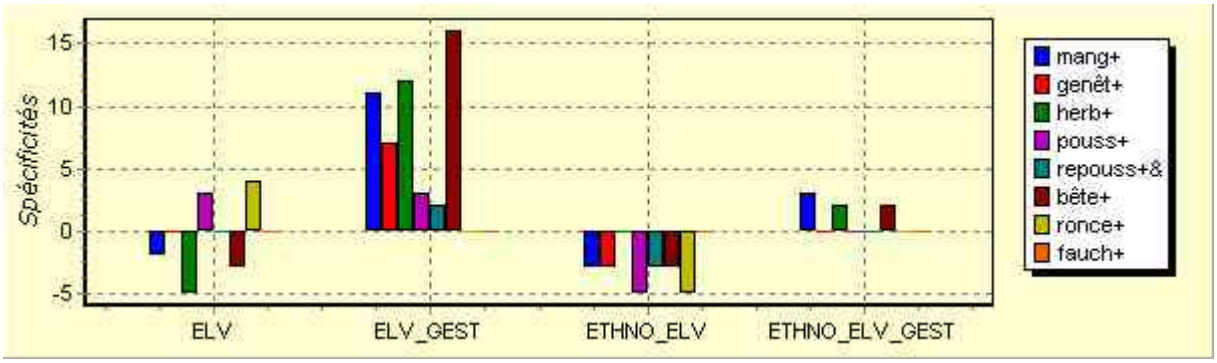
Si le score des spécificités est significatif de la représentation d'une forme réduite dans une partie du corpus par rapport à l'ensemble du corpus, dans le cadre d'une analyse d'entretien, l'interprétation des résultats est articulée autour des questions d'une part et des réponses d'autre part. C'est, alors, l'écart de score des spécificités entre une partie du corpus correspondant aux questions et une autre, correspondant aux réponses, qui est significatif et pas le score en lui-même. Ainsi, d'un point de vue interprétatif, les résultats de l'analyse contrastive fournie par LEXICO doivent être observés de manière différentielle.

Ceci entendu, le graphique 1 complète et confirme pour partie les résultats fournis par ALCESTE. Il confirme les formes réduites associées aux éleveurs (*mang*+) et aux gestionnaires (les autres formes réduites), toutes sur-représentées dans les sous-corpus correspondants. Les formes réduites décrites comme caractéristiques d'un groupe socioprofessionnel par notre expertise à l'aide d'ALCESTE sont toutes confirmées par un indice de spécificité strictement supérieur à 0 dans les parties du corpus correspondantes (ELV et GEST). Le groupe éleveurs-gestionnaires (ELV_GEST) est plus moyennement représenté (spécificités positives ou négatives plus proches de 0) que les deux autres groupes. Mais la composante éleveurs est fortement représentée avec un haut score pour la spécificité *mang*+.

Le graphique 1 montre une forte corrélation entre les suremplois dans le questionnement aux gestionnaires et leurs réponses. Par contre, le graphique ne fait pas apparaître de corrélations particulières entre le *mang*+ des éleveurs et le questionnement auquel ils ont été soumis, l'ethnologue utilisant ce terme envers eux, dans ce que met en valeur cette partition, d'une manière neutre. D'une manière générale, mis à part pour le questionnement des gestionnaires, le discours de l'ethnologue apparaît comme plutôt neutre. On peut donc supposer une influence de l'intervieweur sur les gestionnaires plus forte qu'envers les autres groupes d'acteurs/locuteurs.

Néanmoins, le questionnement au groupe éleveurs-gestionnaires est le plus moyennement représenté (spécificités plus proche de 0, ayant moins d'amplitude que pour les questions aux autres groupes). On peut donner deux explications de ce phénomène : soit les éleveurs-gestionnaires représentent, pour l'intervieweur, un groupe intermédiaire relevant à la fois des éleveurs et des gestionnaires, on retrouverait donc cette double composante dans son questionnement, d'où le phénomène de neutralité qui apparaît par rapport au questionnement de seuls éleveurs ou gestionnaires ; soit, au contraire, les locuteurs interviewés possédant cette double composante amènent l'intervieweur à leur poser des questions plus équilibrées, neutres, que des individus appartenant au seul groupe, éleveurs ou gestionnaires.

Ce graphique permet, de plus, de déterminer à quel questionnement lexico-thématique, quelles forment réduites, chacun des groupes socioprofessionnels ont été confrontés, en comparaison des autres, par l'intervieweur. Ainsi, on voit apparaître des composantes communes dans le questionnement (*parc*+, *niveau*+) et des composantes particulières (*mang*+ pour les éleveurs et les éleveurs-gestionnaires ; *chambre*+, *gestion*+, *zone*+ pour les gestionnaires), l'écart étant jugé significatif s'il est supérieur ou égal à 5.



Graphique 2 : formes réduites caractéristiques des éleveurs, comparaison intervieweur/interviewés

Pour comprendre plus précisément l'articulation des différentes formes réduites, thématiques, utilisées par l'ethnologue avec les interviewés, nous avons utilisé le **corpus QR éleveurs**

LEXICO et le **corpus QR gestionnaires LEXICO**. Nous y avons respectivement recherché les huit formes réduites décrites par ALCESTE comme les plus caractéristiques du discours des éleveurs et des gestionnaires (Tableau 1 classes 6 et 5).

Graphique3 : formes réduites caractéristiques des gestionnaires, comparaison intervieweur/interviewés

Les résultats de ces analyses reportés dans les graphiques 2 et 3 ne montrent pas de corrélation systématique entre l'apparition d'un terme réduit commun entre intervieweur et groupe d'interviewés dans l'analyse d'ALCESTE et une sur-représentation dans le discours correspondant de l'ethnologue (à l'exemple de *mang*+ pour les éleveurs).

Mais l'observation plus précise de l'articulation des formes réduites permet de tirer des enseignements précis sur l'interaction entre intervieweur et gestionnaires. L'avantage, en terme de résultats, d'une analyse contrastive comme celle de LEXICO et de ces spécificités est que, en supposant qu'une interaction neutre, sans orientation thématique de l'intervieweur, implique une représentation neutre ou nulle (en terme de spécificité) des formes réduites dans le discours de l'intervieweur, il apparaît un système de « vases communicants » selon lequel une sur-représentation trouve sa source dans une sous-représentation dans un autre sous corpus et réciproquement.

En croisant ce constat avec les particularités discursives des entretiens semi-directifs, que l'on peut schématiser en un diptyque question/réponse, on peut prévoir et décrire les implications des variations possibles de spécificités d'une sous-partie du corpus, correspondant aux questions, à une autre, correspondant aux réponses, et réciproquement. C'est finalement l'écart entre les spécificités de l'intervieweur et du groupe interviewé qui est significatif (on fixe le seuil d'écart significatif à 5) :

- si l'écart est nul (ou presque), l'influence peut être considérée comme réciproque, l'interaction est neutre (c'est le cas pour *parc*+, *agriculture*+, *gestion*+, *milieu*+, *ouvert*+ et *action*+ entre gestionnaires et intervieweur) ;
- si l'écart est positif : soit l'intervieweur à la plus forte spécificité, auquel cas, plus l'écart est grand, plus il influence l'interviewé (c'est le cas pour *chambre*+ entre gestionnaires et intervieweur) ; soit l'interviewé à la plus forte spécificité, l'intervieweur n'influence pas l'interviewé, plus l'écart est grand et plus l'information thématique qu'il reçoit est libre, spontanée (c'est le cas pour *zone*+ entre gestionnaires et intervieweur).

Ainsi, alors qu'une comparaison globale du graphique 2, avec un ensemble de spécificités négatives pour le questionnement de l'intervieweur aux interviewés, et du graphique 3, plusieurs spécificités positives, semble, au premier abord, indiquer un meilleur rendement dans l'investissement lexico-thématique vis-à-vis des éleveurs que vis-à-vis des gestionnaires,

notre méthode d'interprétation nous pousse à relativiser ce constat : on a pour les éleveurs une interaction neutre pour les formes réduites *mang*+, *genêt*+, *repouss*+, *bête*+ et *fauch*+, une influence de l'intervieweur pour *herb*+ et deux formes réduites spontanées, *pouss*+ et *ronce*+. La comparaison avec les résultats précédents montre un rendement comparable à celui des gestionnaires.

D'un point de vue interprétatif, les unités lexico-thématiques que nous décrivons comme libres, spontanées sont certainement des réponses à d'autres thématiques du questionnement de l'intervieweur. De même, l'influence de l'intervieweur peut être interprétée comme une non réponse de l'interviewé ou l'utilisation d'une autre forme lexico-thématique de la part de celui-ci. Ainsi, l'articulation de ces résultats avec des connaissances de type terminologique sur le domaine étudié ou les connaissances de terrain de l'intervieweur, peut ouvrir des pistes sur les influences thématiques globales de l'intervieweur et leurs relations avec des marqueurs lexico-thématiques propres aux interviewés (si on pousse jusqu'au bout le principe des vases communicants, les formes libres des interviewés peuvent être des candidats réponses aux formes neutres ou aux formes influencées de l'intervieweur).

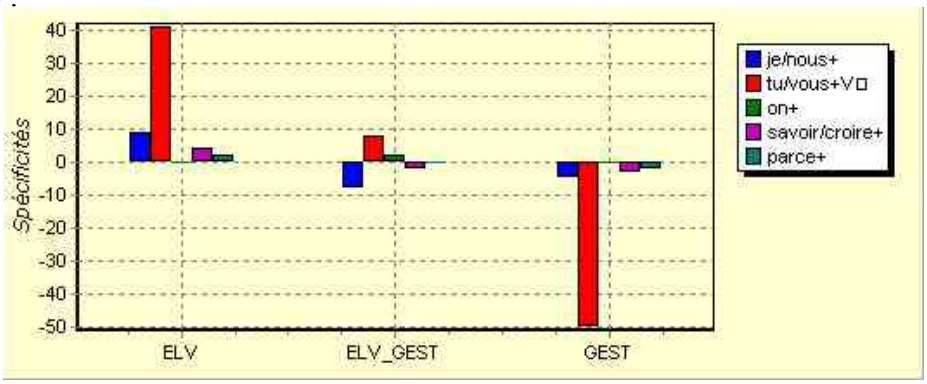
3.2. Influence situationnelle et fonctionnelle

3.2.1. Typologie de Biber

Pour pousser plus loin l'analyse fonctionnelle du discours de R. Dumez, nous allons maintenant nous appuyer sur les hypothèses de D. Biber, (Biber, 1988). Il oppose discours impliqué pour le locuteur et discours informatif. Quelques-uns des principaux traits linguistiques qui caractérisent le discours impliqué sont, pour l'anglais, les verbes privés (*supposer*, *croire*, *savoir*, *douter*), les pronoms des premières et deuxièmes personnes, « *on* », ainsi que les connecteurs (*parce que*, *bien que*). Nous avons réalisé, à l'aide de LEXICO, les groupes de formes correspondant à chacun de ces traits et nous avons recherché dans le **corpus Q LEXICO** (correspondant au discours de l'ethnologue aux éleveurs et aux gestionnaires au format LEXICO), grâce au calcul des spécificités, les éventuels sur- ou sous-emplois de ces groupes, jugeant qu'ils sont aussi valides pour le français, afin de déterminer le niveau d'implication de R. Dumez dans son discours vis à vis des différents groupes socioprofessionnels. Pour simplifier l'analyse nous ne nous sommes intéressés qu'aux groupes éleveurs, gestionnaires et éleveurs-gestionnaires.

3.2.2. Analyse de l'implication de l'intervieweur à l'aide de LEXICO

En confrontant le **corpus Q LEXICO** au calcul des spécificités des groupes de formes correspondant aux marqueurs d'implication décrits par Biber, nous avons obtenu le graphique 4 :



Graphique 4 : marqueurs d'implication, discours de l'intervieweur aux éleveurs et aux gestionnaires

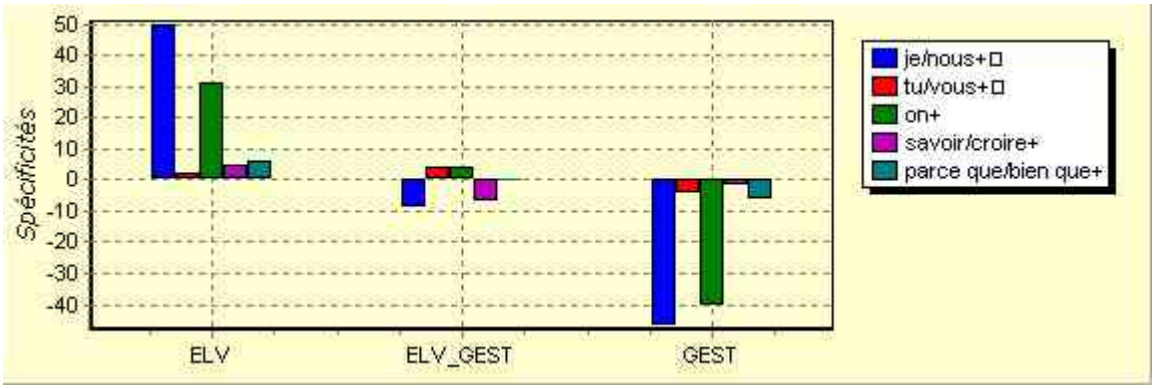
Ce graphique montre un grand écart dans l'implication de l'intervieweur, principalement entre le groupe éleveurs et le groupe gestionnaires du PNC, autour des pronoms de la deuxième personne. Pour simplifier l'analyse nous ne nous sommes intéressés qu'à ces deux groupes.

Les traits correspondants à une forte implication du locuteur sont presque systématiquement sur-représentés dans les parties du corpus où R. Dumez s'adresse aux éleveurs et sous-représentés dans celles où il s'adresse aux gestionnaires, à l'exception de *on*+ qui est moyennement représenté dans ces deux sous-parties du corpus et de « supposer » et « douter » qui sont absents du corpus.

Nous déduisons de cette analyse que R. Dumez adopte une posture énonciative plus impliquée lorsqu'il s'adresse aux éleveurs que lorsqu'il s'adresse à des gestionnaires et réciproquement plus informative vis à vis des gestionnaires que des éleveurs. Qu'elle soit contrôlée ou non par le locuteur, due à ses visées énonciatives ou le fruit de l'interaction entre locuteurs, cette différence de posture est, d'après nous, la marque d'une non équivalence de discours dans le corpus, d'une influence de l'intervieweur, ce qui doit être pris en compte par l'analyste lors de l'observation quantitative de ces données qui ne sont donc pas le résultat d'un questionnement équivalent.

Ce type de phénomène paraît difficile à éviter lors de la réalisation d'entretiens semi-directifs car, par exemple, comme l'évoque R. Dumez, pour obtenir rapidement un grand nombre de données langagières, il faut choisir un positionnement énonciatif qui convient à ses interlocuteurs et ainsi le faire éventuellement varier de l'un à l'autre. La réalité des enquêtes de terrain justifie certainement l'utilisation d'entretiens semi-directifs, mais il nous semble intéressant pour le chercheur d'avoir une description scientifique de ce type de phénomènes lors de l'exploitation des entretiens.

Pour compléter cette analyse nous avons eu recours au **corpus R LEXICO**, qui nous a permis d'observer la répartition des marqueurs décrient par Biber dans le discours des interviewés (graphique 5) :



Graphique 5 : marqueurs d'implication des éleveurs, des gestionnaires et des éleveurs-gestionnaires

Là encore les traits correspondants à une forte implication sont systématiquement sur-représentés dans les parties du corpus correspondant aux éleveurs et sous-représentés dans celles correspondant aux gestionnaires. Il existe donc une beaucoup plus forte implication des éleveurs que des gestionnaires.

Les pronoms de la première personne et le *on*+ sont les formes les plus sur-représentées pour les éleveurs et les plus sous-représentées pour les gestionnaires.

Au *je*+, *nous*+ et *on*+ des éleveurs répond donc le *tu*+ et le *vous*+ dans le questionnement de l'intervieweur. A l'inverse, à l'absence de *je*+, *nous*+ et *on*+ des gestionnaires répond

l'absence de pronoms de la deuxième personne dans le questionnement de l'intervieweur. R. Dumez nous dit dans sa thèse : « Etablir un dialogue avec son interlocuteur est une nécessité pour mener à bien une démarche ethnoscientifique justement basée sur les données tirées d'entretiens. (…) Ce qui importe, c'est de ne pas être identifié à un groupe précis d'acteurs. (…) Acquérir une position de neutralité permet de se placer en dehors des conflits ou des simples oppositions entre acteurs. » (Dumez, 2004). C'est certainement cette recherche de neutralité qui justifie et explique, qu'elle soit ou non volontaire, ce mimétisme de l'intervieweur face à l'implication de l'interviewé et, par là, les écarts d'implication dans le questionnement de R. Dumez.

4. Conclusion

Nous avons exposé dans cet article un travail de formalisation des données et d'articulation de logiciel issus des statistiques textuelles pour mettre en valeur des marques, qui nous sont apparues comme pertinentes, de l'organisation textuelle d'entretiens semi-directifs et de l'interaction entre intervieweur et interviewés qu'ils mettent en jeu. Le résultat de cette étude n'a pas pour prétention de révéler le niveau de comparabilité des enquêtes réalisées selon la technique d'entretiens semi-directifs, mais de révéler des contrastes linguistiques, qui, de notre point de vue, s'avèrent pertinents pour l'analyste désireux d'exploiter et critiquer ses résultats. En aucun cas ce travail ne peut remplacer le nécessaire retour au texte et aux formes brutes pour l'exploitation de ces entretiens, mais il propose une cartographie d'indices thématiques et énonciatifs propres à guider le chercheur dans l'exploitation et la validation de ses résultats, afin, nous l'espérons, d'enrichir et valider d'autant ses conclusions et d'améliorer leur comparabilité.

Références

Burnard L., Sperberg-Mc Queen C. M. (1996). La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange. *Cahier GUTenberg*, 24. http://www.univ-rennes1.fr/pub/GUTenberg/publicationsPS/24-teilite.ps.gz

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press. Cambridge.

Dumez R. (2004). *L'herbe et le feu dans le Parc national des Cévennes, Pratiques de gestion et modes de catégorisation des éleveurs et des gestionnaires*, Thèse d'ethnoécologie. Muséum national d'histoire naturelle. Paris.

Grangé D.,Lebart L. (1993). *Traitements statistiques des enquêtes*. Dunod. Paris.

Harold E. R., Means W. S. (2005). *XML en concentré*. Editions O'Reilly. Paris. 3e édn.

Lebart L., Salem A. (1994). *Les statistiques textuelles*. Dunod. Paris.

Reinert M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en-Yvelines. CNRS.

Salem A., Lamaille C., Martinez W., Fleury S. (2003), *Manuel Lexico 3*. version 3.41, http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/team.htm