

Analyse sociolinguistique d'un corpus oral par regroupement hiérarchique

Annette Gerstenberg

Ruhr-Universität Bochum D-44780 Bochum

Résumé

Cette communication propose une approche de statistique multidimensionnelle d'un corpus d'orientation sociolinguistique, composé d'entrevues biographiques transcrites. La formation du corpus porte sur le rapport entre l'âge élevé des participants et leur langage, dans la mesure où les interviewés appartiennent majoritairement à la tranche d'âge de 70+ (corpus LangAge, Gerstenberg, 2009). Le groupe des participants se présente donc de façon assez homogène du point de vue de leur âge mais la question reste ouverte de savoir quelle homogénéité on peut décrire en ce qui concerne la variation linguistique interne et les conditions extralinguistiques (sociales, biographiques) distinctives. Dans ce contexte, une approche multidimensionnelle comme le regroupement hiérarchique offre la possibilité d'éviter toute définition à priori des variables extralinguistiques à étudier, mais elle sépare les niveaux de description des variables linguistiques d'un côté et leur interprétation sociolinguistique de l'autre côté (Deumert, 2004). L'analyse des variables linguistiques mène, par regroupement hiérarchique, à des configurations similaires ou « profils linguistiques » voisins et, dans une deuxième étape, on arrive à la discussion des caractéristiques communes des participants qui présentent des profils linguistiques similaires. Avant de présenter l'analyse statistique et graphique en forme de diagramme en arbre, le choix des variables linguistiques et les modalités d'annotation seront décrits, et aussi le codage qui prépare le traitement statistique (pour lequel le langage de programmation, des fonctions statistiques et graphiques sous le logiciel R ont été utilisés). Pour conclure, les particularités linguistiques et extralinguistiques des branches du diagramme en arbre seront soulignées.

Abstract

This contribution proposes a statistical approach, adopting the method of hierarchical clustering on sociolinguistic data. These data had been collected in an inquiry among mostly elderly French-speaking people, with whom had been made biographically orientated interviews. The transcribed versions of these interviews were organized in a corpus called LangAge, providing the possibility to annotate linguistic variable semi-automatically and to make use of tools for lemmatization and part-of-speech-tagging. The corpus is designed to learn more about the internal structure of an old generation, as it is represented in the group of participants and restricted to the genre of biographically interview which has a considerable part of monographic narratives. On the example of five linguistic variables, this contribution discusses the possibilities to model statistically the correlation of these variables on the one hand, and on the other hand to ask for specific individual and inter-individual configurations of these variables, as they can be observed in an hierarchical cluster. This procedure offers the opportunity to group the participants only considering the linguistic variation observed, while the extra-linguistic evaluation of the clustering will be done in a second step. In the final discussion it will be shown that this approach leads to a new idea of the relation between linguistic variation and sociolinguistic factors which goes beyond the parameters of age, sex, education and socioeconomic «class».

Keywords: corpus linguistics, hierarchical grouping, French, sociolinguistics, variation

1. Introduction

La comparaison entre le langage documenté par les différents groupes d'âge représente un des intérêts principaux de la sociolinguistique d'inspiration labovienne qui s'intéresse non seulement à la variation en synchronie, mais aussi au changement linguistique qui s'y exprime. Par conséquent, c'est dans la perspective historique que l'on distingue entre les groupes d'âge en synchronie, s'agissant d'une projection possible du changement linguistique des décennies passées. Ainsi, les personnes âgées deviennent des témoins du langage plus ou moins archaïque appris lors de leur jeunesse, supposant souvent une conception d'un comportement conservateur des locuteurs âgés (Ager, 1990 : 4).

Une telle conception, même si entre les adhérents d'un groupe d'âge on distingue ultérieurement selon la classe sociale (si difficile à définir) et le sexe des locuteurs, comporte une vue homogénéisante. Par conséquent le terme de *groupe d'âge* devient – très souvent – un synonyme de *génération*.

Dans la contribution suivante, cette équation présente le point de départ de l'analyse. Le corpus sur lequel se base l'analyse permet, par sa composition, de préciser la conception d'une génération linguistique en explorant les structures marquantes à l'intérieur d'un même groupe d'âge.

Ces structures seront décrites premièrement sur la base de la variation linguistique interne à l'exemple de cinq variables. Ce type d'exploration des structures d'un groupe d'âge mène à modéliser les combinaisons des variables linguistiques internes par regroupement hiérarchique, ce qui est complété par la discussion des relations avec les variables extralinguistiques. La valeur heuristique de la méthode du regroupement hiérarchique consiste dans le fait que la discussion des configurations calculées mène à découvrir quelles caractéristiques extralinguistiques lient les participants regroupés dans une même classe. Cette interprétation peut inclure tout type d'information concernant les participants. Cela permet d'aller outre la série conventionnelle des variables de l'âge, du sexe, de l'éducation et de la classe socio-professionnelle, et de considérer aussi les données qualitatives des entrevues. Une telle approche interprétative semble adaptée à la situation de l'âge élevé où on peut assumer une individualité formée pendant toute la vie.

Cette approche reprend la démarche de Biber (2006) qui consiste à distinguer strictement les niveaux de l'analyse interne des données linguistiques et leur configuration « externe » (registres). En sociolinguistique, Deumert (2004) adopte une position similaire, parce que les dix variables (morphosyntaxiques, syntaxiques, morpholexicales) sont d'abord décrites, entre autres par regroupement hiérarchique, dans leurs configurations en variétés, dont les caractéristiques sociolinguistiques seront explorées dans un deuxième temps ¹.

Dans ce qui suit, il sera montré dans quelle mesure une analyse statistique multivariée peut aider à concrétiser l'idée de la variation linguistique à l'interne d'une seule tranche d'âge en décrivant sa structuration. Les données de base dérivent d'un corpus oral, composé par entrevues biographiques. Premièrement, la composition de ce corpus, appelé corpus LangAge, sera brièvement décrite, avant de discuter les variables choisies, aussi en ce qui concerne la corrélation entre elles. Après sera présentée la procédure du regroupement hiérarchique. Les résultats de cette analyse multivariée seront discutés pour caractériser leur caractéristique linguistique, les interprétations possibles du point de vue sociolinguistique et les questions ouvertes ².

¹ Cf. la réception dialectologique par Lenz (2006).

² Un point crucial pour le traitement des données est le choix des logiciels et leur combinaison : Transcriber pour la transcription des données (audio-text), TUSTEP pour la gestion du corpus et de la base des données et R pour les parties statistique et graphique.

2. Corpus LangAge

La construction du corpus LangAge mène à une description qualitative et quantitative du langage de personnes âgées qui représentent la plupart des locuteurs inclus, parce que le langage dans la phase de vie appelée souvent le troisième âge n'est pas aussi bien décrit en sociolinguistique que le langage « des jeunes », qui est aujourd'hui un sujet bien établi de la recherche linguistique.

L'idée d'approcher ce problème à travers la constitution d'un corpus accessible à des analyses quantitatives est née de l'observation que l'âge élevé est en même temps caractérisé par l'individualité des locuteurs, pendant tout le cours de leur vie, et par leur adhérence au collectif de leur génération. Dans toutes les phases de l'analyse, cette double orientation, à l'individu et à l'ensemble des participants, a joué un rôle important.

Les locuteurs inclus ont été priés de participer à des entrevues d'orientation historique et biographique, traitant la propre vie (famille d'origine, enfance, école, formation et vie professionnelle), mais aussi les expériences de la seconde guerre mondiale et les changements sociaux actuels. Les participants inclus ne présentaient pas de graves maladies physiques ou psychiques, ils appartenaient donc à l'âge normal (Brouillet and Syssau, 2000). L'étude présentée dans ce qui suit inclut 46 participants (20 hommes et 26 femmes) d'une seule tranche d'âge, celle de 70 à 94 ans.

La mise en contact directe ou indirecte (boule de neige) impliquait une certaine sélection positive en ce qui concerne le style de vie, c'est-à-dire un certain niveau d'activité sociale et une curiosité personnelle, exprimée dans le simple fait que les participants acceptaient de se faire interviewer par une personne inconnue auparavant. Ce type de sélection positive ne semble pas facile à éliminer, étant présent aussi dans d'autres enquêtes sur l'âge élevé comme la grande étude, représentative, de Berlin (Baltes, 2007) – il semble que surtout les personnes d'un bas niveau d'études aient tendance à refuser la participation à une enquête. Cependant, la structure sociale du groupe des locuteurs de notre corpus ne présente pas les proportions de la population de base, mais elle est plutôt équilibrée et inclut des locuteurs de différents niveaux sociaux (pour les détails, cf. Gerstenberg, 2009). Par conséquent, toutes les classes d'éducation scolaire et une vaste gamme de professions exercées avant la retraite constituent les variables sociolinguistiques relatives (Tab. 2).

L'utilisation d'un questionnaire ouvert mais traitant d'une série de thèmes homogène assurait un grade de comparabilité suffisant entre les participants. En outre, l'orientation thématique et la stratégie de l'interview (questions brèves, écoute active) avaient pour conséquence que les entrevues incluaient de grandes parties narratives et monologiques, dans un style de moins en moins surveillé.

Les données personnelles étaient partiellement rassemblées dans un questionnaire fermé avant l'entrevue (lieu et date de naissance, type d'habitation) et partiellement par l'analyse des entrevues (catégories sociolinguistiques telles formation, profession, famille). Ces données extralinguistiques étaient insérées de façon codée dans une base de données.

Les fichiers sons des entrevues étaient transcrits orthographiquement ; le corpus formé par les 46 entrevues analysées ici compte 309.000 mots graphiques. Le texte transcrit était annoté dans le cours de l'analyse linguistique. En plus, le logiciel Cordial a été employé pour le traitement automatique des données, c'est-à-dire, pour la lemmatisation et la catégorisation des locutions ou multitermes (part-of-speech-tagging).

Les résultats de ces analyses étaient également insérés dans la base de données. La structure de cette base de données a permis la transformation des vecteurs dans un format directement lisible par le logiciel R pour effectuer les analyses statistiques ³ et la production des graphiques.

3. Mesures de la variation dans le corpus LangAge

Les aspects choisis dans ce qui suit pour analyser la variation dans le corpus correspondent à l'orientation du corpus, en tant qu'ils sont liés aux dimensions de l'âge avancé et ses implications possibles pour les usages langagiers correspondants.

Le comportement linguistique des participants pendant l'entrevue doit être compris en première ligne comme adaptation à la situation communicative. Seulement si on prend en considération les limites du genre textuel qui est l'entrevue biographique, on peut arriver à une interprétation équilibrée de la variation présente dans le corpus : d'un côté, elle permet la configuration d'un profil linguistique ⁴ du participant et de l'autre côté, elle ne couvre qu'une portion limitée de la gamme de styles dont se sert le participant dans les diverses situations communicatives.

Les variables incluses dans l'analyse suivante appartiennent à quatre champs différents : la « richesse lexicale » (relation type-token et classes de fréquence), l'orientation normative (*ne* de la négation, *on* vs. *nous*), style nominal (quotient noms/verbes). Il s'agit de paramètres qui permettent de décrire l'actualisation du genre textuel « entrevue biographique ». Cette actualisation s'explique comme stratégie du locuteur autant que comme expression de ses compétences linguistiques. Ici, il ne s'agit pas de chercher à préciser où se trouve exactement la limite entre « pouvoir » et « vouloir » du locuteur. Ce qu'on a décrit concernant la signification du discours présent dans un corpus en tant que « self-referential system » (Teubert, 2005), nous paraît valable aussi pour la variation linguistique : le corpus reste une représentation incomplète des usages langagiers des locuteurs participants, mais la méthode d'enquête contrôlée garantit la comparabilité des locuteurs participants à l'intérieur du corpus.

Le problème de la richesse lexicale, d'une importance particulière pour le langage des personnes âgées, ⁵ est traité de deux façons qui décrivent (1) le nombre de mots graphiques différents et (2) la quantité des mots graphiques répétés. Pour (1) on a choisi le quotient type-token (TTR) après 2000 mots graphiques. Quant à la répétitivité observable dans les entrevues (2), on a choisi le nombre de mots graphiques présents dans les classes de fréquence égal ou supérieur à 30 (dans ce qui suit, la valeur inverse de cette variable est dénommée F30i). Entre ces deux variables existe une relation, mais elle n'est pas assez prononcée pour renoncer à l'une des deux variables.

Pour discuter l'orientation normative des participants, deux domaines classiques de la linguistique variationnelle ont été choisis, l'emploi du *ne* de négation et l'emploi du pronom *nous* à la première personne du pluriel. La perte du *ne* de la négation est un des phénomènes marquants de la diachronie du vingtième siècle, et aujourd'hui l'emploi normatif du *ne* est considéré généralement comme marqué stylistiquement (Riegel et al., 2004 : 418). L'étiquetage des cas où le *ne* de négation manquait a été fait de façon semi-automatique en TUSTEP. Le pourcentage de *ne* réalisé vs. *ne* non réalisé est dénommé PNE. La deuxième variable morphosyntaxique rend

³ Les commandes de la programmation sous R seront données dans le cours de l'analyse.

⁴ Cf. Dargnat (2008) pour une conception statistique du profil langagier et son interprétation sociolinguistique.

⁵ Cf. pour la discussion de ces variables Baayen (2008 : 252ss) ; Hausser (2000 : 321) ; Tweedie and Baayen (1998) ; Wimmer (2005) et pour les effets de l'âge élevé sur le quotient type-token, Nef and Hupet (1992) et Walker et al. (1988).

manifeste la préférence des participants à utiliser le pronom *nous* vs. *on*. La valeur stylistique de cet emploi résulte du fait que, actuellement, le pronom *on* est presque généralisé (Grevisse, 1993 : §724). Après la discussion des références possibles de *on* et de *nous*, les emplois de *on* dans le cas où ils étaient échangeables avec *nous* ont été étiquetés. Le pourcentage de *nous* vs. *on* dans ces positions est dénommé PNOUS. Ces variables indiquent des phénomènes voisins, mais elles illustrent deux aspects différents, qui ne figurent pas nécessairement en même temps ⁶.

Une prévalence de style nominal a été décrite comme caractéristique pour les registres oraux plus formels (Blanche-Benveniste, 2005 : 53 ; cf. *elaboratedness* et *plannedness* in Atkinson and Biber, 1994). Cette variable a été calculée d'après l'étiquetage morphosyntaxique du corpus à l'aide du logiciel Cordial ⁷, sous la base d'une version tokenisée des transcriptions ⁸.

4. Préparation des données

Les valeurs des cinq variables ont été standardisées (ce qui est indiqué par la Majuscule S : STTR2000, SF30i, etc.). Les données ont été réunies dans une matrice avec 46 lignes (une pour chaque participant) et 5 colonnes, pour les 5 variables incluses. Les participants présentent des « cas » dont le regroupement doit révéler la similarité ou la différence. Pour mieux comprendre la configuration des valeurs mesurées, la corrélation entre les vecteurs des cinq variables a été calculée.

Entre trois paires de variables on peut observer une corrélation significativement positive au-dessus de 0.50 (Tab. 1).

	STTR2000		SF30i		SPNE		SPNOUS	
	r_s	p	r_s	p	r_s	p	r_s	P
SF30i	0.50	<0.01						
SPNE	0.45	0.01	0.36	0.01				
SPNOUS	0.23	0.12	0.18	0.22	0.51	<0.01		
SNV	0.44	<0.01	0.39	<0.01	0.55	<0.01	0.39	<0.01

Tableau 1 : Corrélations (Spearman) entre les variables : r_s et niveau de signification

Ces valeurs soulignent la corrélation entre les deux valeurs de la richesse lexicale, entre la réalisation du *ne* de négation et l'emploi de *nous* vs. *on* dans la première personne du pluriel et une corrélation positive entre la réalisation du *ne* et une tendance au style nominal.

Ces chiffres montrent d'un côté la co-présence de ces phénomènes chez un certain nombre des locuteurs ; mais de l'autre côté, on observe les limites d'une conception d'un locuteur idéal qui emploie de façon régulière tous les moyens à sa disposition en même temps. Dans les réalités du parler, on pourrait conclure face à ce tableau, que les locuteurs utilisent largement la possibilité de combiner librement les éléments « qui font du style » (Larthomas, 2000 : 943) au lieu de suivre une logique abstraite selon laquelle l'idéal stylistique serait réalisé en même temps à tous les niveaux intéressants.

⁶ Si le taux de *ne* réalisés est élevé, on observe également une quantité élevée de *nous* vs. *on*.

⁷ Entre les parties du discours, l'étiquetage des noms et des verbes est très valable, à l'exception des noms propres qui étaient exclus. Également les formes des verbes auxiliaires ont été exclues.

⁸ Le quotient des formes lemmatisées nominales et verbales est dénommé NV.

5. Regroupement hiérarchique

La relation entre les niveaux de l'analyse devient plus complexe si on ne part pas des variables et de leur co-présence, mais des locuteurs et de leur libre choix d'employer les traits linguistiques décrits ici dans les cinq variables. Dans ce qui suit, cette question sera approfondie par le regroupement statistique des participants qui présentent des profils linguistiques similaires : à ce propos, pour découvrir les structures dans le corpus, on a effectué le regroupement hiérarchique en tant que méthode pluridimensionnelle capable de réunir des groupes de locuteurs de façon que leur comportement linguistique, d'après ce qui est mesuré dans l'étude présentée ici, soit le plus homogène possible.

Mais il faut considérer que, si on peut observer des groupes de participants qui présentent des valeurs similaires, il ne s'agit pas de les classer sur une seule échelle de normativité ou grade d'élaboration ; une classe de participants résultant du regroupement hiérarchique peut être caractérisé par des valeurs élevées d'une partie des variables et par des valeurs basses des variables, correspondant au fait que la corrélation entre les variables n'est pas dans tous les cas très prononcée. Le regroupement hiérarchique est fait en quatre étapes ⁹ :

- (1) Formation d'une matrice consistant en 46 lignes pour les 46 participants inclus et cinq colonnes, où les valeurs des variables sont inscrites.
- (2) Calcul des distances entre les dimensions représentées dans la matrice ; la mesure de la distance euclidienne a été choisie (en R : « dist »).
- (3) Choix d'une méthode d'agglomération et emploi de cette méthode (en R : « hclust »). Le regroupement hiérarchique prévoit des méthodes d'agglomération différentes (*linkage rules*) : soit afin de minimiser les distances, soit afin de créer des classes le plus homogènes possible. Après la comparaison des résultats de l'emploi des autres méthodes et suivant les résultats d'autres études d'orientation sociolinguistique (Deumert, 2004 : 111) a été choisie la méthode d'agglomération Ward (en R : method = « ward »), qui minimise la variance entre les distances dans les classes retournées par le regroupement.
- (4) Réalisation graphique du résultat de l'agglomération, en forme de diagramme en arbre.

De ces étapes résulte le diagramme en arbre (Fig. 1) dans lequel chaque participant est nommé avec son code, le sexe et l'âge (p.ex. A16_f_71). En plus, le nombre de quatre classes – qui s'est montré le plus marquant ¹⁰ – a été visualisé par des rectangles, calculés en R avec « rect.hclust », k=4).

6. Résultats du regroupement

Les quatre classes du diagramme en arbre présentent des dimensions très diverses ; dans C1 ne se trouve qu'un seul locuteur, et dans C2 sept. Les classes plus grandes de C3 (13 participants) et C4 (25 participants) sont rapprochées sur le premier niveau de bi-partition (de haut en bas).

Pour valider et décrire les caractéristiques linguistiques de chaque classe et du groupe entier (C1–C4) des rectangles des boîtes à moustaches ont été visualisées (Fig. 2 ; en R : « boxplot », range=1.5). Les rectangles des boîtes à moustaches incluent le second et le troisième quartile (le côté du rectangle correspond donc à l'écart interquartile), la limite entre les deux étant marquée par le médian.

⁹ Cf. pour une description de ces quatre étapes aussi Deumert (2004 : 113).

¹⁰ Une partition en deux classes retourne (C1, C2) et (C3, C4) ; une partition en trois classes retourne (C1, C2), C3 et C4 ; une partition en cinq classes divise C4 en deux branches, entre A14 et A48.

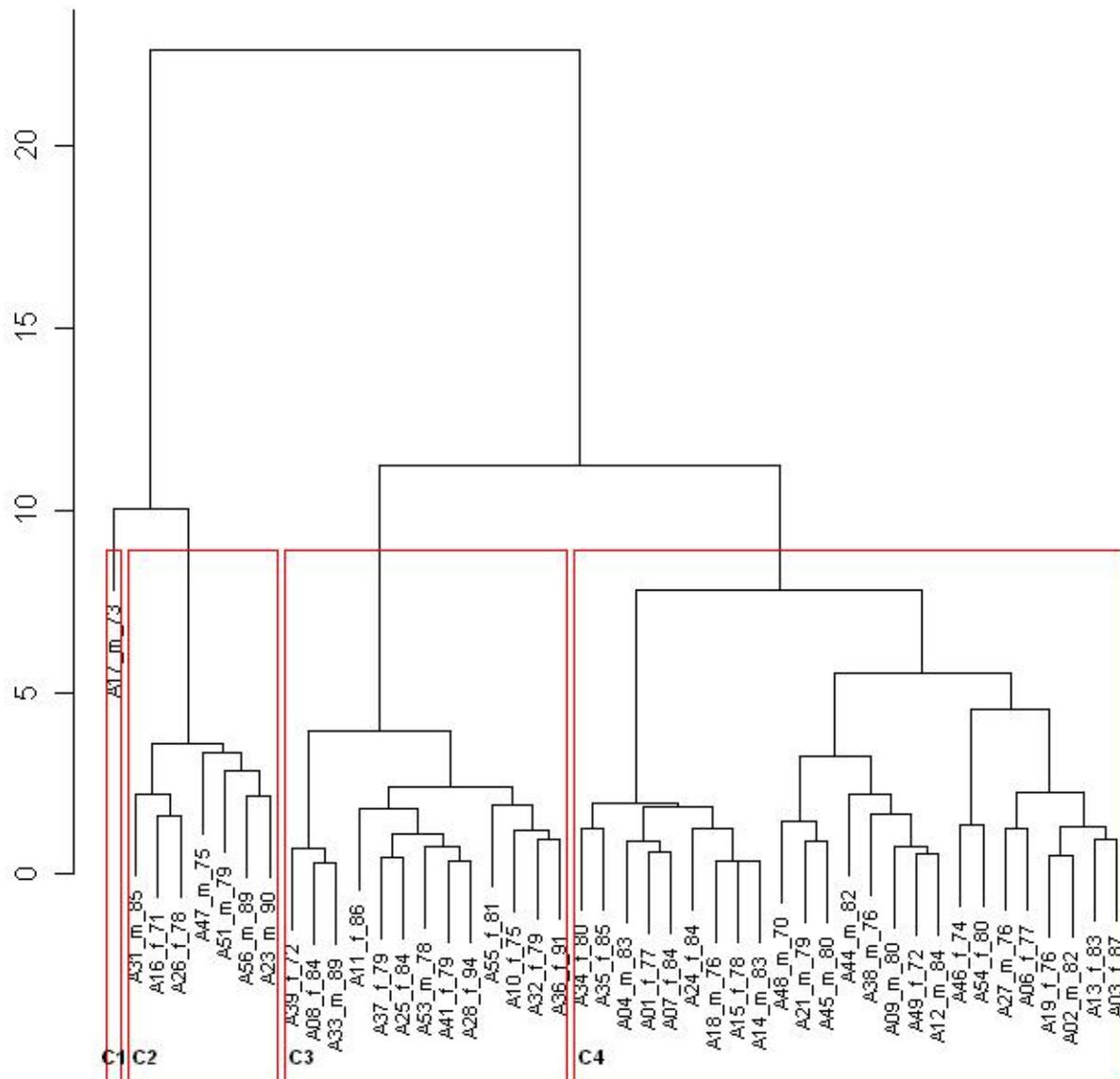


Figure 1 : Diagramme en arbre, méthode d'agglomération : Ward

Le fait que dans C1 soit classé un seul participant correspond clairement aux valeurs élevées de tous les variables de cette classe. L'originalité de ce participant s'exprime surtout dans la différence très marquée entre lui et les autres participants en ce qui concerne F30i et PNOUS.

Les graphiques (Fig. 2) mettent aussi en évidence la position de C2, proche à C1 à l'exception de F30i où C2 est rapprochée à C3 et C4. Par rapport aux classes de C1 et C2, caractérisées par valeurs élevées, la classe de C3 se présente comme « basse » dans toutes les figures.

Même si dans la classe de C2 est regroupé un tiers du nombre de participants regroupés dans C4, l'écart interquartile de C2 est clairement élevé par rapport à C4 dans les cas de PNOUS (1.1) et de F30i (0.9). Les différences entre l'écart interquartile de C3 et de C4 sont très faibles. Sur la base de l'écart interquartile, on pourrait dire que la classe de C2 est plus hétérogène que les classes de C3 et C4.

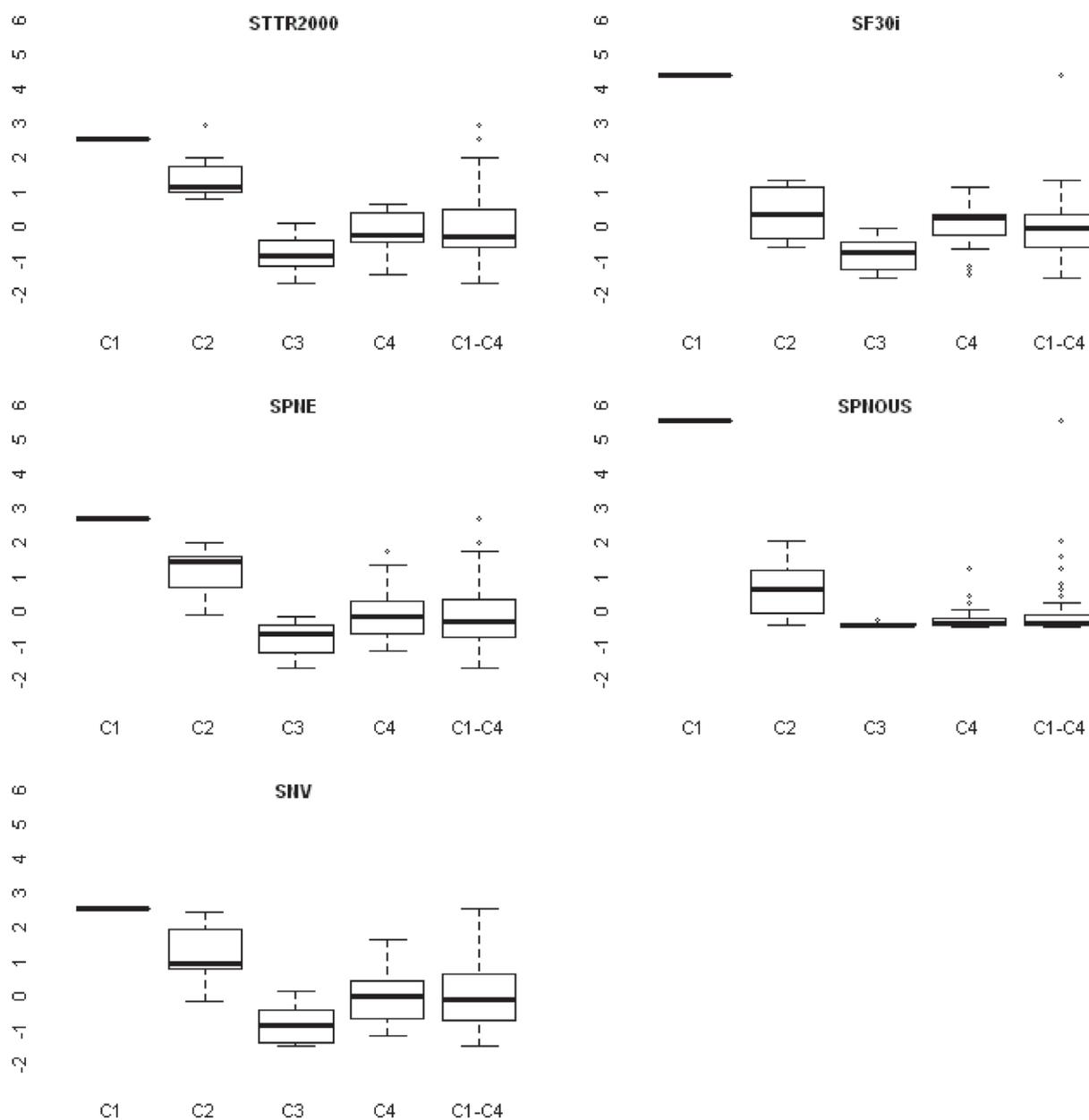


Figure 2 : Boîtes à moustaches des variables linguistiques dans les classes C1–C4

Quant aux traits extralinguistiques caractérisant les quatre classes, C3 présente une majorité féminine (correspondant au niveau d'études) et de l'adhérence aux classes plus basses de l'enseignement et de l'activité professionnelle, et, moins marquant, le contraire dans les classes de C1 et C2 (Tab. 2).

Les proportions des variables extralinguistiques prises en considération sont le plus proches de celle du groupe entier dans le cas de C4.

	sexe		enseignement ¹¹			activité profession av. la retraite ¹²			
	<i>f</i>	<i>m</i>	<i>ens1</i>	<i>ens2</i>	<i>ens3</i>	<i>prf1</i>	<i>prf2</i>	<i>prf3</i>	<i>prf4</i>
C1	0	1	0	0	1	0	0	1	0
C2	2	5	1	1	5	0	0	2	5
C3	11	2	9	2	2	4	7	1	1
C4	13	12	7	10	8	1	9	8	7
C1-C4	26	20	17	13	16	5	16	12	13

Tableau 2 : Variables extralinguistiques

7. Discussion

Le regroupement hiérarchique présente comme structures marquantes d'un côté un groupe composé de huit participants (deux classes) dont l'emploi des variables linguistiques ou marques stylistiques peut être décrit comme « ambitieux » sur tous les niveaux pris en considération et de l'autre côté une grande classe « moyenne ».

Les profils sociolinguistiques des participants regroupés dans les C1 et C2 montrent qu'il s'agit dans plusieurs cas d'anciens instituteurs (A17, A23, A16), professeurs (A26) ou prêtres (A31), mais aussi des cadres (A47, A51, A56). Au vu de ces profils professionnels, on en arrive à la conclusion qu'une profession ou activité liées à un langage modèle a des effets prononcés sur le comportement linguistique, même à partir de la retraite. Mais cela ne vaut pas pour tous les membres de ces groupes professionnels, d'autres professeurs, instituteurs, cadres et prêtres figurant dans d'autres groupes.

La structure de l'enquête en forme d'entrevues biographiques permet d'aller au-delà du profil formé par les paramètres de l'éducation et de la formation et activité professionnelles et de prendre en considération les conditions biographiques individuelles. Les participants de C1 et C2 ont des biographies très diverses, ce qui rend plus difficile la proposition d'une explication pour le type de variation que l'on pourrait définir comme stylistiquement marqué. Il faut en outre prendre en compte le type de données analysées, c'est-à-dire le contexte interactif d'une entrevue biographique. Ainsi, on arrive à supposer que la variation linguistique résulte aussi d'un comportement ambitieux provoqué par le fait qu'il s'agissait d'un entretien enregistré. Cette idée de comportement ambitieux semble plutôt adaptée pour expliquer les cas de co-variation exprimée des variables stylistiquement marquées. En plus, le concept d'ambition individuelle offre la possibilité de trouver un lien entre les biographies des participants et leur auto-présentation linguistique durant l'entrevue. Une explication qui se base sur l'ambition individuelle des locuteurs s'adapte au cas où le participant a l'allemand comme langue maternelle (A47). La volonté, exprimée dans l'entrevue, de s'intégrer dans la société française où il a passé sa vie dès l'université, correspond à un emploi marqué du point de vue de la variation linguistique durant l'entrevue. Dans en autre cas aussi, une ambition intellectuelle inachevée dans la jeunesse pourrait correspondre à l'emploi marquant indiqué par le regroupement

¹¹ Les participants sont classés dans quatre groupes d'activité professionnelle exercée avant la retraite. Ces quatre types d'activité professionnelle correspondent à quatre types de formation professionnelle : prf1 = vendeuse, femme au foyer (sans formation), prf2 = sténo-dactylo etc. (p.ex. Certificat d'aptitude professionnelle), prf3 = enseignants (p.ex. École normale des Instituteurs), prf4 = prêtres, professeurs de lycée, cadres supérieurs (diplômes universitaires).

¹² Les participants sont classés dans trois niveaux de formation scolaire : ens1 = Certificat d'Études Primaires, ens2 = Brevet élémentaire, ens3 = Baccalauréat.

hiérarchique : Le participant A51 a dû, pour des raisons familiales, quitter l'école juste après le certificat d'études, mais il a cependant fait sa carrière de cadre supérieur. Une conception d'ambition linguistique peut, dans d'autres cas, être liée à une forte inclination à la lecture. C'est le cas de A23, le plus âgé des participants; dans ce cas, l'ambition serait une ambition littéraire.

Même si avec ces explications on n'a qu'une idée approximative de ce qu'une ambition linguistique peut avoir comme effet sur la variation, il nous semble important de suivre cette approche d'explication individuelle et biographique et de ne pas se contenter d'interpréter la variation linguistique mécaniquement comme résultat prévisible des paramètres comme âge, éducation scolaire et activité professionnelle. Ainsi, par l'analyse statistique on met en relief des cas extraordinaires. Dans une perspective quantitative ces cas singuliers ne sont guère significatifs. S'il nous semble important de les prendre en considération, c'est parce qu'ils semblent jouer un rôle important dans la construction de l'image de la langue des personnes âgées : Le profil linguistique accentué stylistiquement correspond à l'idée très commune d'un langage conservateur et soutenu de la vieille génération. Le diagramme en arbre montre d'un côté que des profils linguistiques de ce type existent et, en même temps, que cette projection ne couvre qu'une partie minoritaire de la communauté des locuteurs âgés.

8. Conclusion

Ce qui caractérise une tranche d'âge en tant que « génération » n'est pas seulement ce qui lie ses adhérents, mais aussi sa diversification interne. L'approche présentée dans cette contribution propose l'emploi des analyses statistiques multivariées pour découvrir – de façon exemplaire – telles structures marquantes dans un groupe d'âge, sur la base d'un choix de variables linguistiques. Le résultat, le diagramme en arbre, présente une grande classe (presque la moitié) face à deux classes (env. un quart) et un cas singulier.

La valeur heuristique de cette approche résulte de la possibilité de séparer l'analyse des traits linguistiques caractérisant les classes du diagramme de la suivante interprétation d'orientation sociolinguistique. L'analyse linguistique menait à la description de deux classes comme « ambitieuses », alors que la classe la plus grande montrait des valeurs moyennes et la dernière des valeurs généralement plus basses. Dans l'analyse sociolinguistique de ces classes, la correspondance entre le niveau de l'enseignement et de l'activité professionnelle était prononcée pour les valeurs linguistiques « hautes » et « basses ». Dans le cas de la classe la plus grande, on observe une corrélation restreinte entre les variables linguistiques et sociolinguistiques. La valeur prédictive des variables extralinguistiques semble restreinte.

Pour mieux comprendre ce trait très marquant de la « configuration générationnelle » qui représente le diagramme en arbre, il nous semble important de formuler les possibles explications davantage sur un niveau qualitatif : Il reste à vérifier dans quelle mesure la structuration décrite résulte d'une ambition et expérience linguistiques individuelles des locuteurs qui peut contribuer, surtout à un âge avancé, à atténuer les effets d'une éducation scolaire ou formation professionnelle inachevées.

L'étude présente un échantillon de dimensions limitées, en ce qui concerne le nombre des participants, la tranche d'âge, les profils sociaux et le lieu d'habitation. Cette configuration ne permet donc de généraliser les résultats quantitatifs, c'est-à-dire, les proportions des classes linguistiques et sociolinguistiques. Une conclusion possible pour l'orientation de futures recherches sociolinguistiques, surtout sur le langage de personnes âgées, pourrait être que la combinaison de méthodes quantitatives et qualitatives contribue à remettre en discussion la

valeur prédictive des variables extralinguistiques conventionnelles et à accentuer les facteurs individuels et biographiques.

Références

- Ager D. (1990). *Sociolinguistics and contemporary French*. Cambridge : Cambridge University Press.
- Armstrong N. (2001). *Social and Stylistic Variation in Spoken French. A comparative approach*. Amsterdam: John Benjamins.
- Atkinson D. and Biber D. (1994). Register: A Review of Empirical Research. In: Biber, D. and Finegan, E., editors, *Sociolinguistic perspectives on register*, Oxford University Press: pp. 351-385.
- Baayen R. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Backhaus K., Erichson B., Plinke W. and Weiber R. (¹¹2006). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Berlin-Heidelberg-New York : Springer.
- Baltes P. (2007). Alter(n) als Balanceakt. In Gruss, P., editor, *Die Zukunft des Alterns. Die Antwort der Wissenschaft. Ein Report der Max-Planck-Gesellschaft*, München : Beck, pp. 15-34.
- Biber D. (2006). *University Language. A corpus-based study of spoken and written registers*. Amsterdam : John Benjamins.
- Blanche-Benveniste C. (2005). L'étude grammaticale des corpus de langue parlée en français. In Williams, G., editor, *La linguistique de corpus*. Presses universitaires de Rennes, pp. 47-66.
- Brouillet D. and Syssau A. (editors) (2000). *Le vieillissement cognitif normal. Vers un modèle explicatif du vieillissement*. Bruxelles : De Boeck-Larcier.
- Cordial = Synapse Développement (1994–2008). *Cordial Analyseur Version 14*. Synapse.
- Cresti E. (2005). Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken corpora. In Cresti, E. and Moneglia, M., editors, *C-Oral-ROM: integrated reference corpora for spoken romance languages*. Amsterdam : Benjamins, pp. 209-256.
- Dargnat M. (2008). Profils linguistiques et structuration textuelle. In *JADT2008*, vol. 9/1, pp. 369-379.
- Deumert A. (2004). *Language Standardization and Language Change. The dynamics of Cape Dutch*. Amsterdam : Benjamins.
- Gerstenberg A. (2007). Generation und <Sprachprofil>. Untersuchung zum höheren Lebensalter auf Basis biographischer Interviews. In Hartung, H., Reinmuth, D., Streubel, C. and Uhlmann, A., editors, *Graue Theorie. Die Kategorien Alter und Geschlecht in der Forschung*, Köln-Weimar-Wien : Böhlau-Verlag, pp. 15-34.
- Gerstenberg A. (2009). *Generation und Sprachprofile im höheren Lebensalter. Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews*. Manuscript, Ruhr-Universität Bochum.
- Grevisse M. (¹³1993). *Le bon usage. Grammaire française*. Refondue par André Goosse. Paris-Gembloux : Duculot.
- Hausser R. (2000). *Grundlagen der Computerlinguistik*. Berlin-New York : Springer.
- Larthomas P. (2000). La stylistique. In Antoine, G. and Cerquiglini, B., editors, *Histoire de la langue française 1945-2000*, Paris : CNRS Editions: pp. 937-943.
- Lenz A. (2006). Clustering Linguistic Behaviour on the Basis of Linguistic Variation Methods. In Filppula, M., Klemola, J., Palander, M. and Penttilä, E., editors, *Topics in Dialectal Variation*, Joensuu Yliopisto, pp. 69-98.

- Nef F. and Hupet M. (1992). Les manifestations du vieillissement normal dans le langage spontané oral et écrit. *L'année psychologique*, vol. 92 : 393-419.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing (R 2.7.1)*. R Foundation for Statistical Computing, <http://www.r-project.org> (04.07.2008).
- Riegel M., Pellat, J.-Ch. and Rioul R. (2004). *Grammaire méthodique du français*. Paris : PUF.
- Teubert W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, vol. 10/1 : 1-13.
- Transcriber = DGA and CEP (1998-2008). *Transcriber – a Tool for Segmenting, Labeling and Transcribing Speech.*: <http://trans.sourceforge.net> (19.12.2005).
- TUSTEP. Zentrum für Datenverarbeitung (2004). <http://www.zdv.uni-tuebingen.de/tustep/tdv.html> (19.12.2005).
- Tweedie F.J. and Baayen R.H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, vol. 32/5 : 323-352.
- Walker, V. G., Roberts, P.M. and Hedrick D.L. (1988). Linguistic Analyses of the Discourse Narratives of Young and Aged Women. *Folia Phoniatrica*, vol. 40 : 58-64.
- Wimmer G. (2005). The Type-Token relation. In: Köhler, R., Altmann, G. and Piotrowski, R., editors, *Quantitative Linguistics. An International Handbook (HSK 27)*, Berlin-New York : de Gruyter : pp. 361-368.