CORPUS

Corpus

15 | 2016
Corpus de français parlé et français parlé des corpus

Disfluences et vieillissement langagier. De la base de données VALIBEL aux corpus outillés en français parlé

Disfluencies and language aging. New corpora and tools for exploring spoken French in the VALIBEL database

Catherine T. Bolly, George Christodoulides et Anne Catherine Simon

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/3006
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 octobre 2016
ISSN : 1638-9808

Référence électronique
Catherine T. Bolly, George Christodoulides et Anne Catherine Simon, « Disfluences et vieillissement langagier. De la base de données VALIBEL aux corpus outillés en français parlé », *Corpus* [En ligne], 15 | 2016, mis en ligne le 15 janvier 2017, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/3006

Ce document a été généré automatiquement le 8 septembre 2020.

Disfluences et vieillissement langagier. De la base de données VALIBEL aux corpus outillés en français parlé

Disfluencies and language aging. New corpora and tools for exploring spoken French in the VALIBEL database

Catherine T. Bolly, George Christodoulides et Anne Catherine Simon

1 Dans cet article, nous nous attachons à explorer les possibilités d'investigation qu'offre la base de données textuelles orales VALIBEL, en portant une attention particulière à l'outillage (principalement, le programme *DisMo* pour l'annotation des disfluences) et au corpus Corpage, récemment intégré à la base et dont la population cible concerne des personnes âgées.

1. La base de données VALIBEL

2 La base de données textuelles orales VALIBEL ne constitue pas un corpus mais un regroupement de corpus constitués depuis 1987. Il s'agit donc d'une sorte de « réservoir de corpus » qui est alimenté de manière incrémentale au fur et à mesure des nouveaux projets de recherche nécessitant de collecter des données orales (section 2). Documentées et archivées sous format électronique, ces données peuvent être réexploitées à des fins de recherches variées (section 3), touchant notamment à des questions sociétales cruciales telles que le vieillissement de la population (section 5). La documentation qui les accompagne comprend des métadonnées sur la situation d'interaction et les locuteurs, ainsi que sur la transcription orthographique effectuée. Pour une partie des données, cette transcription est directement alignée sur le signal sonore. Certains corpus font en outre l'objet d'annotations particulières (section 4).

2. Historique du centre VALIBEL

3 Le centre de recherche VALIBEL a été créé en 1989 par Michel Francard. L'acronyme qui a donné son nom au centre (VAriétés LInguistique du français en BELgique) rend compte des objectifs de recherche établis à cette époque : il s'agissait de créer un observatoire des usages du français en Belgique, dans leur variation. L'intérêt s'est d'abord porté sur l'analyse des représentations linguistiques des locuteurs concernant, par exemple, les accents régionaux ou l'insécurité linguistique (Francard, 1993). Des collectes de données orales et de vastes enquêtes par questionnaire sont mises en place pour tester la diffusion et la vitalité des régionalismes lexicaux. Un des résultats est le *Dictionnaire des belgicismes* faisant actuellement référence (Francard, Geron, Wilmet & Wirth, 2015). Entre 1989 et 1999, la majorité des corpus recueillis consistent en interviews sociolinguistiques – comportant le plus souvent une partie de discussion ouverte visant à recueillir des informations sociobiographiques sur le locuteur et à le faire parler librement, et une partie plus contrainte guidée par un questionnaire. D'autres corpus, de taille plus réduite, ont été réalisés ponctuellement pour des études variées (sur la liaison, l'argumentation dans les débats, l'alternance de code français-wallon, etc.).

4 En 2009, le centre s'élargit en accueillant une nouvelle équipe et redéfinit ses objectifs, ce qui se marque par un changement de nom : Valibel - Discours et Variation. La sociolinguistique reste un ancrage théorique important, comme en atteste la participation de Valibel au vaste projet de recueil de données pour l'étude de la Phonologie du français contemporain (PFC – Durand, Laks & Lyche, 2009), qui a permis de renouveler les études sur la prononciation du français en Belgique (Hambye & Simon, 2009 ; Simon, 2012). L'autre axe de recherches concerne l'analyse du discours, en particulier les connecteurs et marqueurs de discours (Bolly, Crible, Degand & Uygur-Distexhe, 2015), les unités de base du discours (Martin, Degand & Simon, 2014), les effets du vieillissement langagier sur la dimension pragmatique (Bolly & Boutet, soumis) ou le traitement de la fluence et de la disfluence à l'oral (projet ARC « Fluency and disfluency markers. A multimodal contrastive perspective », voir Crible, Dumont, Grosman & Notarrigo, 2015). Des chercheurs travaillant sur d'autres langues que le français (en particulier l'espagnol et le néerlandais) se sont également ajoutés à l'équipe (De Cock, 2014 ; Van Goethem & Hiligsmann, 2014), et des études contrastives sont en cours (De Cock & Roginsky, 2015). Depuis une dizaine d'années, un effort particulier a été investi pour recueillir de nouveaux corpus plus diversifiés en termes d'activités communicatives. En guise d'exemple, le corpus « style » présente la particularité d'enregistrer un même locuteur dans deux situations contrastées (par ex. en situation professionnelle et privée) afin de documenter la dimension diaphasique de la variation.

5 Le développement de la base de données textuelles orales VALIBEL, dans ce contexte, n'est pas une fin en soi, mais constitue la pierre de touche de recherches qui se veulent fondées empiriquement sur corpus. Cela offre également un terrain intéressant d'élaboration méthodologique, concernant les types de données à recueillir, les modes de recueil, de documentation et d'annotation. Le principe qui régit la recherche au centre Valibel reste l'étude de la variation à partir d'usages langagiers attestés et documentés (*i. e.* à partir de corpus), visant à documenter la diversité des pratiques langagières en Belgique francophone, et dans d'autres langues.

3. Description des corpus dans la base de données

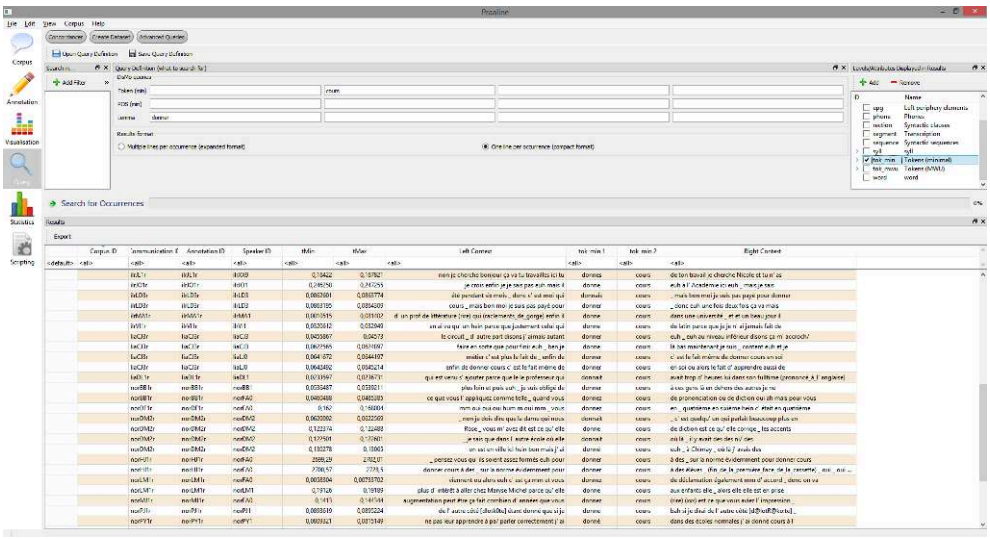
6 Par définition, la base de données est ouverte, et sa croissance est le signe de la vitalité des recherches menées au sein du centre Valibel. Le développement se fait dans deux directions :

– en ajoutant de nouvelles données (enregistrements, transcriptions, métadonnées) (*cf.* ci-dessous et en 5.) ;
– en ajoutant de nouvelles annotations aux données existantes (*cf.* section 4.).

7 Aujourd'hui, on peut dire que la banque de données VALIBEL compte 24 corpus exploitables. Ces corpus ont initialement fait l'objet d'une chaîne de traitement standardisée, décrite dans Dister & Simon (2007), suivie plus récemment d'une phase d'annotation et de traitement des données telle que décrite sous 4.2. Ces corpus représentent actuellement 494 enregistrements sonores, impliquant 568 locuteurs, totalisant 352 heures de parole, accompagnés de métadonnées (informations sur les locuteurs et sur la situation d'interaction) et de transcriptions orthographiques. Ces transcriptions totalisent 3 388 208 tokens[1].

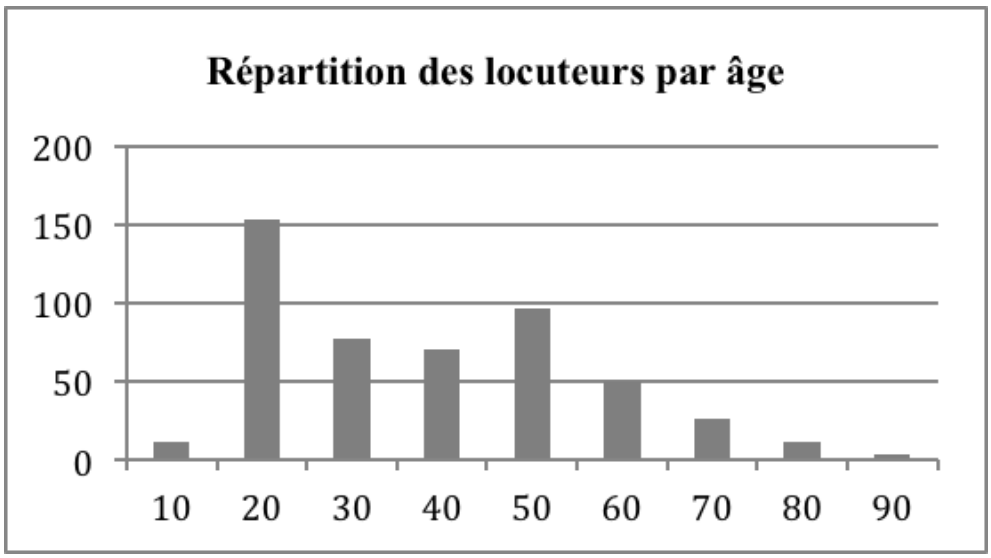
8 Les données sont archivées dans la base de données [moca], qui permet d'interroger à distance les données et de télécharger les fichiers son et les transcriptions (pour plus de détail : Dister, Francard, Hambye & Simon, 2009 ; Simon, Francard & Hambye, 2014). Les métadonnées, qui sont également interrogeables via l'interface [moca], ont été intégrées dans la transcription orthographique sous la forme de TEI Headers pour favoriser l'interopérabilité des corpus. Elles donnent des informations sur les aspects suivants : (i) enregistrement : nombre de locuteurs, relation entre locuteurs, date et lieu d'enregistrement, langue, type d'interaction, durée, nombre de mots, statut de l'enregistrement, etc. ; (ii) corpus : code d'identification, année de constitution, objectif de recherche, nombre d'enregistrements et de locuteurs, nombre de mots, durée ; (iii) locuteur : sexe, âge, localisation géographique, lieu de naissance, degré de scolarité, profession, etc. Le système permet, à l'aide de critères sur les situations d'enregistrement ou sur les locuteurs, de créer des collections de données en vue d'études particulières. Par ailleurs, les données sont désormais analysables et consultables par le biais du logiciel de gestion de corpus *Praaline* (Christodoulides, 2014), qui permet la consultation des transcriptions et leur annotation sous plusieurs couches à l'aide de concordances (Barreca & Christodoulides, 2014 – voir Figure 1), ainsi que l'application d'outils d'annotation automatique (*cf.* section 4.2).

Figure 1. Requête multi-niveaux et présentation des résultats sous forme de concordance dans le logiciel Praaline



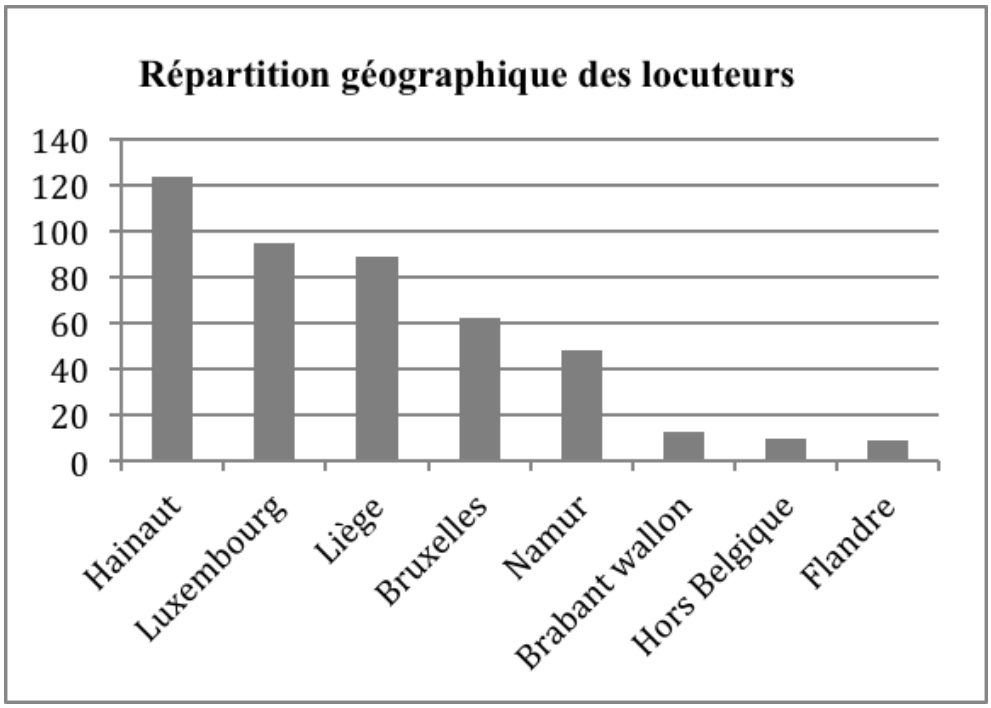
9 La base de données, dans son ensemble, ne présente pas un échantillonnage équilibré de données orales, ni en termes de situations de communication ni en termes de répartition des locuteurs (âge, localisation géographique, etc.). Il s'agit donc de créer, de manière opportuniste, des sous-corpus équilibrés à partir des données disponibles. Ainsi, l'âge moyen des locuteurs est de 30,3 ans, et les locuteurs ayant entre 20 et 30 ans sont les plus représentés (voir Figure 2).

Figure 2. Répartition des locuteurs par âge dans la base VALIBEL



10 En termes de localisation géographique, parmi la grande majorité de locuteurs belges francophones, ce sont les locuteurs du Brabant wallon qui sont les moins nombreux et ceux du Hainaut qui sont majoritaires (voir Figure 3).

Figure 3. Répartition des locuteurs par localisation géographique dans la base VALIBEL



4. Annotation multiniveau

4.1 Interface syntaxe/prosodie

11 Certains corpus ont fait l'objet d'annotations particulières. Ainsi, le corpus LOCAS (LOuvain Corpus of Annotated Speech) a été annoté manuellement en unités syntaxiques et en unités prosodiques afin d'étudier comment ces deux niveaux d'organisation se combinent pour former des unités discursives (Degand & Simon, 2009). Au niveau syntaxique, par exemple, on a identifié les unités maximales de rection (un élément recteur accompagné de tous les éléments qui en dépendent) et les séquences fonctionnelles ; des éléments non régis (comme les marqueurs de discours ou les associés) ont également été annotés. Du point de vue prosodique, on a perceptivement identifié les frontières prosodiques majeures et intermédiaires, en les assortissant d'un contour intonatif (Christodoulides & Simon, 2015). Les hésitations et les marques d'écoute (back-channels) ont également été annotées. D'une durée de 3 heures 11 pour 36 912 tokens, ce corpus regroupe de manière équilibrée des échantillons représentatifs de 12 situations de parole contrastées entre elles (Martin *et al.*, 2014).

4.2 Annotation morphosyntaxique et détection automatique des disfluences

12 Par le biais du logiciel *DisMo* (Christodoulides, Avanzi & Goldman, 2014), des couches d'annotation supplémentaires ont été appliquées à toutes les transcriptions de la base VALIBEL : une annotation morphosyntaxique (au niveau des tokens isolés et au niveau des unités polylexicales), une lemmatisation et une annotation des disfluences.

L'annotateur automatique *DisMo* prend en compte les phénomènes spécifiques aux conventions de transcription de l'oral (par exemple, l'absence de ponctuation) et est structuré autour de six modules qui s'appliquent en cascade :

1) tokenisation : prétraitement et découpage en unités lexicales ;
2) application de ressources linguistiques : annotation des unités non-ambiguës et établissement de la liste des étiquettes possibles pour les autres cas (à noter que certaines disfluences et unités polylexicales sont reconnues à ce stade, ainsi que les marqueurs de discours et les unités polylexicales potentielles) ;
3) annotation morphosyntaxique préliminaire en parties du discours ;
4) détection des disfluences et de la segmentation ;
5) annotation morphosyntaxique finale, combinée avec la détection des unités polylexicales ;
6) post-traitement des annotations, à l'aide des règles de cohérence.

13 Le codage des disfluences détectées automatiquement par *DisMo* suit le schéma d'annotation présenté de manière synthétique dans la Figure 4 (pour plus de détail, voir Christodoulides & Avanzi, 2015).

Figure 4. Schéma d'annotation des disfluences dans DisMo (dans Christodoulides & Avanzi, 2015)

Niveau 1 : Disfluences **simples** : affectent un seul token		
FIL	Pauses remplies	j' hésite ***euh FIL*** un peu en parler
LEN	Allongement lié à une hésitation	au cercle d'oenologie ***de= LEN*** Bruxelles
FST	Amorce lexicale	comme infirmière ***so/ FST*** sociale
WDP	Pause intra-mot	il m' a dit ça su+ _ ***WDP*** +ffit
Niveau 2 : **Répétitions** où un ou plusieurs tokens sont répétés (exactement)		
REP	Répétition	• les disques ***et REP* et REP_*** lancer les jingles • ***il REP:1 a REP:2*** ***il REP:1 a REP*:2*** ***il REP_ a REP_*** dit que • ***c' REP:1 est REP:2 pas REP*:3*** ***c' REP_ est REP_ pas REP_*** un système génial
Niveau 3 : Disfluences **structurées** (d'édition)		
DEL	Suppression	***c' DEL est DEL vraiment DEL*** ***un DEL*** en tout cas la parole
SUB	Substitution	cette personne ***était SUB**** ***enfin SUB:edt c' SUB_ est SUB_*** un ami de

INS	Insertion	c' est vrai que ***Béthune INS* euh INS+FIL*** ***vivre INS_ à INS_ Béthune INS_*** ça aurait
Niveau 4 : Disfl. **complexes** (combinent plusieurs disfluences structurées)		
COM	Complexe	Leur structure est annotée à l'aide d'un tableau d'empilement

5. Disfluences et vieillissement langagier

14 La problématique du vieillissement de la population et ses retombées socio-économiques dans les pays développés (Berr, Balard, Blain & Robine, 2012) sont au cœur des préoccupations actuelles des chercheurs, toutes disciplines scientifiques confondues. Dans le domaine de la linguistique, en particulier, plusieurs études sur corpus ont été menées durant les cinq dernières années (*cf.* Gerstenberg, 2009, 2011 ; Lee, 2012 ; Bolly & Boutet, soumis) et des réseaux de linguistes se mettent en place à l'international (*cf.* le réseau du CLARe « Corpora for Language and Aging research »). C'est dans ce contexte que le corpus Corpage « A Reference corpus for the elderly's language » a vu le jour (Bolly, Masse & Meire, 2012). Parmi les quelque 212 entretiens récoltés qui constituent le corpus Corpage (106 sujets âgés interrogés ; 2 entretiens par informateur ; environ 144 heures d'enregistrements), 10 entretiens ont été transcrits et révisés selon les normes VALIBEL pour être intégrés à la base de données (8 heures 35 min. ; environ 130 000 tokens). Les entretiens semi-dirigés en face-à-face mettent en scène un étudiant et une personne âgée de plus de 75 ans à son domicile, sur le thème du récit de vie et du rapport à l'âge. Les sujets recrutés ne présentent pas de lésion ni de trouble cognitif majeur. Notons que la constitution de ce corpus est le fruit d'une collaboration interdisciplinaire en sciences humaines (en linguistique, psychologie et psychogériatrie) et suit les normes éthiques recommandées dans le domaine (consentement éclairé oral et écrit, recrutement sur base volontaire, anonymisation des données personnelles, etc.).

15 Basée sur l'annotation automatique des disfluences avec *DisMo*, une étude exploratoire a été effectuée pour rendre compte de la distribution des disfluences par tranche d'âge, au sein de la base VALIBEL prise dans son intégralité (incluant les données de Corpage). Si l'on en croit la littérature dans le domaine, nous pouvons nous attendre à observer une plus grande fréquence de marques de disfluence avec l'avancée en âge (hésitations, pauses longues, pauses pleines, particules de discours, répétitions de mots, autocorrections, etc.), en même temps qu'un débit de parole ralenti et une articulation moins précise (Searl, Gabel & Fucks, 2002 ; Lee & Barkat-Defradas, 2014 ; Rousier-Vercruyssen, Lacheret & Fossard, 2014). Ces particularités linguistiques sont le plus souvent considérées comme étant la conséquence de changements cognitifs normaux liés à l'âge (Burke & Shafto, 2008), à savoir le ralentissement de la vitesse de traitement de l'information, un accès moins aisé au lexique et des troubles des capacités d'inhibition (Mathey & Postal, 2008). Mais elles peuvent aussi dépendre de besoins physiologiques (par exemple, l'activité respiratoire), d'une volonté de coopération avec l'interlocuteur ou d'un effort de planification cognitive, davantage marqués chez la

personne âgée (Bortfeld, Leon, Bloom, Schober & Brennan, 2001 ; Smith, Noda, Andrews & Jucker, 2005).

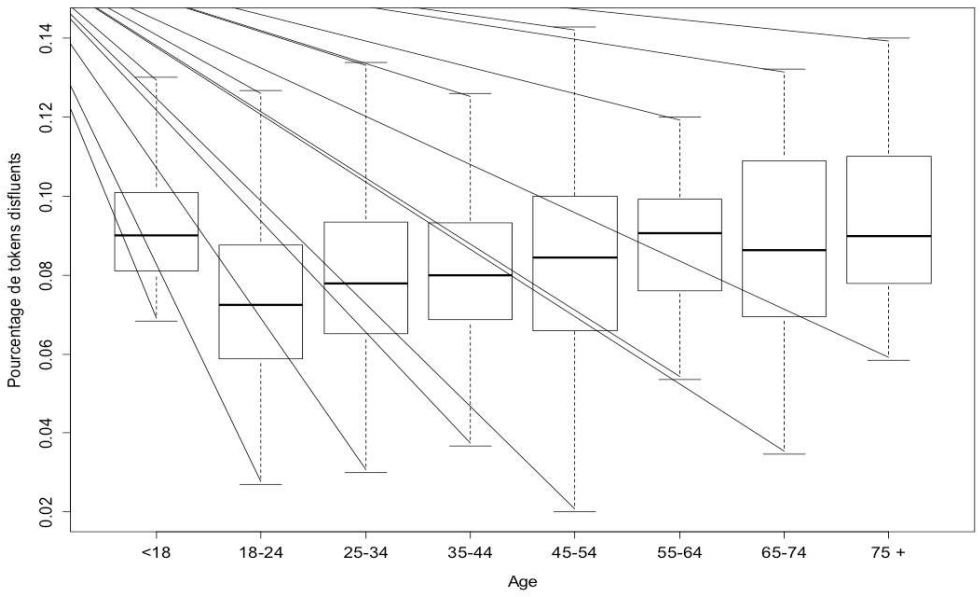
16 Au niveau méthodologique, soulignons que les résultats ne prennent ici en compte que les disfluences annotées aux niveaux 1 et 2 du schéma présenté dans la figure 4. Parmi les disfluences annotées par *DisMo*, nous avons considéré les marques suivantes : les amorces lexicales (FST) (1), les pauses pleines (FIL) (2) et les répétitions (REP) (3).

(1) on avait / euh **q/** un poêle dans la chambre (*Corpage* : ageDM1, 94 ans)
(2) je ne sais pas mais **euh** il y a quelque chose qui ne va pas (*Corpage* : ageDM1, 94 ans)
(3) il n'a plus jamais su avoir **de de de** travail / et / je ne sais pas il avait perdu le nord enfin (*Corpage* : ageDM1, 94 ans)

17 Par ailleurs, la répartition en tranches d'âge par décennies a été adoptée pour faire émerger une éventuelle évolution du nombre de disfluences dans une perspective développementale tout au long de la vie (« lifespan ») (Aldwin, Spiro III, Park & Birren, 2006). Cette étude ne tient donc pas compte des facteurs psychosociaux ni des mécanismes d'adaptation à l'œuvre dans le processus de vieillissement (Freund & Baltes, 2003), mais s'appuie sur une vision purement biologique et chronologique de l'âge (voir entre autres Hamilton, 2001, sur ces questions).

18 Il ressort des résultats une corrélation positive et significative (Spearman r = 0,164 ; d.l. = 857, p < 0.001) entre l'âge du locuteur et la fréquence des disfluences au sein de la base (voir Figure 5). Afin de normaliser les données, cette fréquence a été calculée sous forme de ratio, en divisant le nombre total de tokens « non fluents » (c'est-à-dire ceux qui se trouvent entre le début d'une disfluence et son point d'interruption) par le nombre total de tokens produits par chaque locuteur.

Figure 5. Pourcentage des tokens non fluents en fonction de l'âge du locuteur



19 Les résultats obtenus, qui tendent à confirmer que plus on avance en âge, plus on tend à produire des discours disfluents, doivent néanmoins être nuancés à plusieurs égards. En effet, si des tendances émergent, il ne faut pas oublier qu'il existe des profils idiosyncrasiques de fluence (Shriberg, 1994, 2001), un locuteur pouvant recourir à des pauses pleines (par ex. : *euh*) alors qu'un autre aura tendance à paraphraser dans une

situation similaire (par ex. en réaction au manque de mots). En outre, la catégorie des répétitions annotées inclut dans cette étude les répétitions lexicales perçues comme étant nettement disfluentes (*cf. de de de* dans l'exemple (3) plus haut), mais également des répétitions qui semblent jouer un rôle à un autre niveau dans la production langagière (Rossi, Dominicy & Kolinsky, 2014). Par exemple, la répétition *oui oui* en (4) est une répétition emphatique, qui vient renforcer la valeur d'acquiescement en réaction au propos de l'interviewer. De la même manière, la fonction de la répétition *ça ça* en (5) est ambigüe, puisqu'elle peut être interprétée comme une marque d'hésitation ou comme le résultat d'un procédé syntaxique de topicalisation avec mise en relief du pronom détaché à gauche.

(4) ageMC0 vous vous vous mettez à l'évidence que vous avez bien |- quatre-vingt-deux ans ageBG1 oui hein **oui** -| **oui** bè oui hein / il n'y a pas d'avance (rires) (*Corpage* : ageBG1, 82 ans)[2]

(5) alors un autre c'é/ il était surveillant à D il était professeur ailleurs j'ai je ne l'ai plus jamais vu et tous les autres à part moi je mets peux mettre des croix / **ça ça** m'a fait un |- choc j'ai <ageQL0> mm -| montré la photo à Jacqueline |- et <ageQL0> mm -| on a essayé de retrouver tous les noms des professeurs (*Corpage* : ageJD1, 85 ans)

20 Une étude plus approfondie du rôle cognitif et pragmatique de ces répétitions – en tant que marques potentielles de fluence ou de disfluence – serait donc nécessaire pour déterminer leur rôle dans la planification et dans la coconstruction de l'interaction communicative. Enfin, il ne faudrait pas négliger l'importance des facteurs psychosociaux, tels que le genre (homme/femme) ou la situation communicative, qui jouent un rôle prépondérant dans la production de disfluences par rapport au facteur âge (Bortfeld *et al.*, 2001).

21 Cette première approche exploratoire donne à voir comment, à partir de l'outillage de corpus, des pistes de recherche peuvent émerger pour répondre à des problématiques sociétales fortes. Quelques-unes de ces pistes sont formulées ici sous forme de questions interrogeant l'impact possible des marques de disfluence sur le discours au grand âge (en production et en réception) :

- Quel est le rôle joué par les facteurs environnementaux et psychosociaux liés à la situation de parole (situation de soin, annonce de diagnostic, conversation avec un proche, etc.) dans la production de discours plus ou moins (dis)fluents chez la personne âgée ?
- À partir de quand peut-on considérer qu'un discours disfluent devient problématique et constitue un obstacle au bien-vieillir, tenant compte des mécanismes d'optimisation et d'adaptation (Freund & Baltes, 2003) dont dispose le sujet vieillissant ?
- À l'instar de Davis & Maclagan (2010), ne devrait-on pas considérer le recours à certaines marques de disfluence (pauses pleines, interjections, particules discursives et unités phraséologiques) comme des stratégies adoptées par les plus âgés pour rester impliqués dans l'interaction ?

22 Visant à refléter au plus près l'usage langagier des locuteurs au sein d'une communauté linguistique, voire entre plusieurs communautés, il paraît évident que les approches sur corpus présentent des avantages indéniables pour pouvoir répondre, au moins en partie, à de telles questions.

6. Conclusion

23 Nous avons vu que la base de données VALIBEL, constamment enrichie par de nouveaux corpus et projets de recherches, permettait de faire le lien entre l'outillage des données langagières et leur exploitation dans une visée de recherche fondamentale ou appliquée. C'est ainsi que l'utilisation de programmes d'annotation automatique (p. ex. : *DisMo*), l'élaboration de protocoles d'annotation extrêmement bien documentés (par exemple, le corpus LOCAS ou le projet MDMA) et la possibilité d'interroger les données (et les métadonnées) via une interface fouillée ([moca] ou *Praaline*), permettent d'apporter un éclairage nouveau sur l'usage des locuteurs, tenant compte de variables psychosociales (âge, sexe, niveau d'éducation, etc.) et extralinguistiques (situations de parole, origine géographique, etc.) jouant un rôle important dans la communication langagière.

BIBLIOGRAPHIE

Aldwin C. M., Spiro III A., Park C. L. & Birren J. E. (2006). « Health, behavior, and optimal aging : A life span developmental perspective », *Handbook of the Psychology of Aging 6* : 85-104.

Barreca G. & Christodoulides G. (2014). « Un concordancier multiniveau pour des corpus oraux », *Actes de la 21e Conférence Traitement Automatique du Langage Naturel (TALN)*, Marseille, France, 1er-4 juillet 2014.

Berr C., Balard F., Blain H. & Robine J.-M. (2012). « Vieillissement, l'émergence d'une nouvelle population », *Médecine-Sciences* 28, 3 : 281-287.

Boersma P. & Weenink D. (2015). *Praat : Doing Phonetics by Computer (ver. 5.3.63)*. www.praat.org.

Bolly, C. T. & Boutet D. (soumis). « The multimodal CorpAGEst corpus : Keeping an eye on pragmatic competence in later life ».

Bolly C. T., Crible L., Degand L. & Uygur-Distexhe D. (2015). « MDMA. Un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en contexte », *Discours* 16. http://discours.revues.org/9009 ; DOI : 10.4000/ discours.9009.

Bolly C. T., Masse M. & Meire Ph. (2012). *Corpage. A Reference Corpus for the Elderly's Language*. Louvain-la-Neuve : Université catholique de Louvain (Valibel - Discours et variation & Psychological Sciences Research Institute).

Bortfeld H., Leon S., Bloom J., Schober M. & Brennan S. (2001). « Disfluency rates in conversation : Effects of age, relationship, topic, role, and gender », *Language and Speech* 44 : 123-149.

Burke D. M. & Shafto M. A. (2008). « Language and aging », *The Handbook of Aging and Cognition* 3 : 373-443.

Christodoulides G. (2014). « Praaline : Integrating tools for speech corpus research », *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014 : 31-34.

Christodoulides G., Avanzi M. & Goldman J.-Ph. (2014). « DisMo : A morphosyntactic, disfluency and multi-word unit annotator : An evaluation on a corpus of French spontaneous and read speech », *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014 : 3902-3907.
www.corpusannotation.org/dismo.

Christodoulides G. & Avanzi M. (2015). « Automatic detection and annotation of disfluencies in spoken French corpora », *Proceedings of Interspeech*, 1849-1853.

Christodoulides G. & Simon A. C. (2015). « Exploring acoustic and syntactic cues to prosodic boundaries in French. A multi-genre corpus study », *Proceedings of the 18th International Congress of Phonetic Sciences*, non-paginé.

Crible L., Dumont A., Grosman I. & Notarrigo I. (2015). *Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs. Version 1.0.* Working paper. Université catholique de Louvain et Université de Namur.

Davis B. H. & Maclagan M. (2010). « Pauses, fillers, placeholders and formulaicity in Alzheimer's discourse », *in* N. Amiridze, B. H. Davis & M. Maclagan (éd.) *Fillers, pauses and placeholders* (Typological Studies in Language 93). Amsterdam, Philadelphia : John Benjamins, 189-215.

De Cock B. (2014). *Profiling Discourse Participants. Forms and Functions in Spanish Conversation and Debates* (Pragmatics & Beyond New Series 246). Amsterdam : John Benjamins.

De Cock B. & Roginsky S. (2015). « Identités discursives sur Twitter : Construction de l'identité de député européen en période pré-électorale. Comparaison entre la France, l'Espagne et le Royaume-Uni », *in* F. Liénard & S. Zlitni (éd.) Communication électronique : enjeux, stratégies et opportunités. Limoges : Lambert-Lucas, 137-148.

Degand L. & Simon A. C. (2009). « On identifying basic discourse units in speech : Theoretical and empirical issues », *Discours* 4, http://discours.revues.org/5852.

Dister A. & Simon A. C. (2007). « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica* 1, 1 : 54-79.

Dister A., Francard M., Hambye Ph. & Simon A. C. (2009 [2007]). « Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009) », *Cahiers de l'Institut de linguistique de Louvain* (*CILL*) 33, 2 : 113-129.

Durand J., Laks B. & Lyche C. (éd.) (2009). *Phonologie, variation et accents du français*. Paris : Hermès.

Francard M. (1993). « Trop proches pour ne pas être différents. Profils de l'insécurité linguistique dans la communauté française de Belgique », *Cahiers de l'Institut de linguistique de Louvain* 19 : 61-70.

Francard M., Geron G., Wilmet R. & Wirth A. (2015). *Dictionnaire des belgicismes*. De Boeck : Bruxelles.

Freund A. & Baltes P. B. (2003). « Pour un développement et un vieillissement réussis : sélection, optimisation et compensation », *Revue québécoise de psychologie* 24, 3 : 27-50.

Gerstenberg A. (2009). « The multifaceted category of 'Generation' : Elderly French men and women talking about May 68 », *International Journal of the Sociology of Language* 200 : 153-170.

Gerstenberg A. (2011). *Generation und Sprachprofile im höheren Lebensalter. Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews* (Analecta Romanica 76). Frankfurt am Main : Klostermann.

Hambye Ph. & Simon A. C. (2009). « La prononciation du français en Belgique », *in* J. Durand, B. Laks & Ch. Lyche (éd.) *Phonologie, variation et accents du français*. Paris : Hermès, 95-130.

Hamilton H. E. (2001). « Discourse and aging », *in* D. Schiffrin, D. Tannen & H. E. Hamilton (éd.) *The Handbook of Discourse Analysis*. Malden, Oxford : Blackwell, 568-589.

Lee H. (2012). *Langage et Maladie d'Alzheimer : Analyse multidimensionnelle d'un discours pathologique*. Thèse de doctorat (non publiée). Montpellier : Université Paul Valéry - Montpellier III.

Lee H. & Barkat-Defradas M. (2014). « Complexité phonétique et disfluence dans le vieillissement normal et dans la maladie d'Alzheimer », *SHS Web of Conferences* 8. EDP Sciences : 1315-1327.

Martin L., Degand L. & Simon A. C. (2014). « Forme et fonction de la périphérie gauche dans un corpus oral multigenre annoté », *Corpus* 13 : 243-265.

Mathey S. & Postal V. (2008). « Le langage », *in* K. Dujardin & P. Lemaire (éd.) *Neuropsychologie du vieillissement normal et pathologique*. Issy-les-Moulineaux : Elsevier Masson, 79-102.

Rossi D., Dominicy M. & Kolinsky R. (2014). « The inference of affective meanings : An experimental study », *Language and Cognition*, 7/3 : 351-370.

Rousier-Vercruyssen L., Lacheret A. & Fossard M. (2014). « Pauses silencieuses, planification discursive et vieillissement langagier », *Nouveaux Cahiers de linguistique française* 31 : 197-203.

Searl J. P., Gabel R. M. & Fulks J. S. (2002). « Speech disfluency in centenarians », *Journal of Communication Disorders* 35, 5 : 383-392.

Shriberg E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat. University of California at Berkeley.

Shriberg, E. (2001). « To 'errrr' is human : Ecology and acoustics of speech disfluencies », *Journal of the International Phonetic Association* 31, 1 : 153-169.

Simon, A. C. (éd.) (2012). *La variation prosodique régionale en français.* Bruxelles : De Boeck/Duculot.

Simon A. C., Francard M. & Hambye Ph. (2014). « The VALIBEL Speech Database », *in* J. Durand, U. Gut & G. Kristoffersen (éd.) *The Oxford Handbook of Corpus Phonology*. Oxford : Oxford University Press, 552-561.

Smith S. W., Noda H. P., Andrews S. & Jucker A. H. (2005). « Setting the stage : How speakers prepare listeners for the introduction of referents in dialogues and monologues », *Journal of Pragmatics* 37 : 1865-1895.

Van Goethem K. & Hiligsmann Ph. (2014). « When two paths converge : Debonding and clipping of Dutch *reuze* 'lit. giant ; great' », *Journal of Germanic Linguistics* 26, 1 : 31-64.

NOTES

1. La banque de données compte aussi une grande quantité d'enregistrements en cours de traitement : 379 entrées de métadonnées encodées dans le système sans transcriptions, et 520 fichiers son sans transcription correspondante.
2. Dans les conventions de transcription VALIBEL, les symboles |- et -| indiquent le début et la fin d'un passage de parole en chevauchement.

RÉSUMÉS

Après avoir fait l'état des lieux de la base de données VALIBEL en la situant dans son contexte institutionnel, nous mettons en exergue dans cet article quelques possibilités d'investigation qu'offre la base en regard de ses évolutions récentes. Une attention particulière est portée à l'outillage des corpus en termes de disfluences (avec le programme *DisMo*) et à l'étude du vieillissement langagier (liée au corpus Corpage). Nous concluons en montrant en quoi l'enrichissement constant de la base (en outillage et en corpus) permet d'ouvrir de nouvelles pistes de recherches dans des domaines encore peu explorés en linguistique, eu égard à des problématiques sociétales majeures.

This paper aims at giving an overview of the VALIBEL database as it stands today. In addition, it opens up new perspectives with respect to more recent advances regarding (semi-automatic) annotation, as well as with regard to new corpora created to address societal issues (*cf.* the Corpage corpus). Particular attention is paid here to the automatic detection of disfluencies in the corpus data (using the *DisMo* program), with a developmental view on language and aging.

INDEX

Mots-clés : corpus, annotation, français, disfluences, vieillissement
Keywords : corpus, annotation, French, disfluencies, aging

AUTEURS

CATHERINE T. BOLLY

Universität zu Köln, Université catholique de Louvain

GEORGE CHRISTODOULIDES

Université catholique de Louvain

ANNE CATHERINE SIMON

Université catholique de Louvain