L'âge avancé en perspective longitudinale et ses outils : LangAge, un corpus au pluriel

Eman El Sherbiny Ismail[1], *Annette* Gerstenberg[*1], *Marta* Lupica Spagnolo[1], *Friederike* Schulz[1], *Anne* Vandenbroucke[1]

[1]Université de Potsdam, Département des langues romanes, 14469 Potsdam, Allemagne

Résumé. L'âge avancé en perspective longitudinale et ses outils : LangAge, un corpus au pluriel. En marge des groupes d'âge habituellement représentés dans les échantillons sociolinguistiques, LangAge se positionne comme un recueil d'entretiens et d'enregistrements vocaux consacré à l'âge avancé de la vie. Les participantes et participants sont issues de différents milieux et appartiennent pour la plupart à la tranche d'âge des 70 ans et plus. De plus, le corpus documente jusqu'à dix ans de la vie d'une partie de ces personnes âgées. Il est ainsi possible de suivre l'évolution des mêmes individus à travers plusieurs années et d'éviter, dans la comparaison de différentes couches d'âge, les difficultés habituelles des échantillons en temps réel qui ne peuvent jamais équilibrer les particularités biographiques des individus inclus. Le sous-corpus « couples » regroupe les rencontres avec dix couples durant cette période, ce qui permet d'aborder un domaine rarement étudié. LangAge est conçu, dans l'ensemble, pour contribuer à une image linguistique plus différenciée de la génération la plus âgée. Il en résulte un corpus « au pluriel » dont la plupart des transcriptions alignées et des fichiers son sont disponibles en libre accès. L'outil LaBB-CAT est utilisé pour la publication et consultation en ligne. Nous montrerons comment sa configuration tient compte de l'architecture complexe du corpus et correspond, en même temps, aux principes FAIR tout en respectant les droits de la personne.

Abstract. Longitudinal ageing and analysis tools: LangAge, a plural corpus. LangAge is a collection of interviews and voice recordings dedicated to old age, which is not usually represented in sociolinguistic samples. The participants come from different backgrounds and are mostly in the 70+ age group. For a part of this group, the corpus records follow-up interviews after seven or ten years. It is thus possible to follow the evolution of the same individuals through several years and to avoid, in the comparison of different age strata, the usual difficulties of real-time samples which can never balance the biographical particularities of the individuals included. The "couples" sub-corpus brings together encounters with ten pairs during this period, thus addressing a rarely studied area. LangAge aims at contributing, overall, to a more differentiated linguistic picture of the older generation. The result is a "plural corpus" of which most of the aligned transcripts and sound files are available in open access. The tool LaBB-CAT is used for online publication and browser-based

[*] Corresponding author : gerstenberg@uni-potsdam.de

query. We will show how its configuration takes into account the complex architecture of the corpus and corresponds, at the same time, to the FAIR principles while respecting the rights of the individual.

1 Introduction[i]

« Orléans ça a toujours été une ville test »
(Affaire Classée, INA 1969)

« on dit qu'Orléans est une ville test. Pourquoi ? Il y a à Orléans une proportion d'ouvriers, de paysans, d'employés, de fonctionnaires, de tout le corps des métiers qui est à peu près celle de la France. Il y en résultent fatalement des réactions qui sont – celles de la France. »
(Affaire classée, INA 1969).

On ne présente plus la ville d'Orléans aux publics de corpus français. Il y a de cela 50 ans, la ville d'Orléans devenait un point de référence tout d'abord grâce à L'Étude sociolinguistique sur Orléans (ESLO1) et à sa reprise (ESLO2). La productivité scientifique qui en témoigne s'étend des premiers travaux de l'équipe anglaise (Lonergan, Kay & Ross 1974 ; Biggs & Dalwood 1978), en passant par l'intégration des archives dans la linguistique de corpus moderne (Bergounioux, Baraduc & Dumont 1992 ; Bergounioux 2010), la réalisation de la suite sous la forme d'ESLO2 (Baude & Dugua 2011), jusqu'aux évaluations les plus récentes sous le signe de la microdiachronie (Abouda & Skrovec 2018) et enfin les méthodes du *natural language processing* (Flamein & Eshkol-Taravella 2021). C'est grâce à cette documentation dense et variée de l'usage linguistique qu'il est possible et pertinent d'observer de plus près une seule tranche d'âge. En effet, ce sont en grande partie des personnes de la dernière génération qui ont participé aux entretiens menés pour le corpus LangAge depuis 2005.

Dans ce cadre, la question de la notion d'âge, détaillée au paragraphe 2, est à la fois possible et nécessaire. Le paragraphe 2.1 montre la composition du corpus et le paragraphe 2.2 approfondit cette question par l'illustration du corpus « couples », corpus de micro-diachronie au sein de LangAge, complétée par les questions de l'anonymisation liées à la transcription (paragraphie 2.3).

Le paragraphe 3 porte sur LaBB-CAT, utilisé pour la gestion du corpus ainsi que pour la publication et consultation en ligne. Nous montrerons comment sa configuration permet de profiter de l'architecture longitudinale et composite du corpus.

Le paragraphe 4 pose la question suivante : comment un corpus spécialisé sur un thème étroit peut-il être conçu et mis à disposition conformément aux exigences des principes FAIR (2021) ? Cela dépeint, en somme, l'image de l'ensemble des enregistrements constituant LangAge comme corpus au pluriel.

L'objectif de ce document de synthèse est d'illustrer nos décisions en termes de récolte, préparation et publication des données. Cela est d'autant plus nécessaire que les décisions prises en matière de design et d'architecture du corpus LangAge ne sont pas arbitraires ou neutres, mais reflètent notre approche théorique des phénomènes du vieillissement langagier.

2 L'âge avancé et la perspective longitudinale

Le point de départ de la conception de l'échantillon fut l'observation que l'âge avancé n'est que marginalement représenté dans les études sociolinguistiques (cf. dernièrement Hekkel 2021, chapitre 3). Cependant, les personnes âgées devraient avoir leur place dans la

sociolinguistique (Pichler et al. 2018) tout comme dans la société, surtout sous le signe du bien vieillir (Gerstorf et al. 2015) qui ouvre nouvelles pistes de recherche. Au lieu de mettre en parallèle les nombreuses études sur le langage des jeunes avec la construction nécessairement simplifiée d'une variété de l'âge avancé, le corpus explore une tranche d'âge en coupe transversale, générationnelle. Avec la notion de génération, le caractère historiquement singulier du langage de personnes âgées ayant vécu une période précise est pris en compte, ainsi que la structure intérieure de la génération : en ciblant un seul groupe d'âge, il est possible de porter attention aux différences internes, qui sont facilement nivelées dans la comparaison avec des groupes d'âge plus jeunes. De plus, le fait que tous les enregistrements ont été effectués par la même personne favorise une meilleure comparabilité des entrevues.

Des études longitudinales complètent depuis quelque temps le panorama des comparaisons entre groupes d'âge (Buchstaller & Wagner 2017). L'application de cette approche à la génération de l'âge avancé invite à découvrir les dynamiques subtilement articulées du vieillissement langagier.

2.1 Le design du corpus

Au centre, le corpus LangAge comprend des entretiens avec soixante-seize personnes, 40 femmes et 36 hommes (Fig. 1), en partie interrogées deux ou trois fois (2005, 2012, 2015), avec un total d'environ 502 000 tokens. Des modules additionnels tels que des enregistrements à bâtons rompus, des sessions d'ateliers d'écriture (Gerstenberg & Kairet 2021) et, avec un petit nombre de personnes, des reprises au-delà des trois séries composent l'ensemble des données (Gerstenberg & Hamilton 2022).

Lors de la prise de contact, l'étude a été présentée comme enquête d'histoire orale qui invitait des témoins à partager leurs récits personnels. Comme toute invitation à participer à une telle initiative, cette formule impliquait une première orientation de l'échantillon. Les futurs participantes et participants ont communiqué leur intérêt au projet, ainsi que leur volonté de partager leurs propres expériences avec un public allemand appartenant au cadre de l'enseignement et de la recherche.

Afin de créer un échantillon varié et équilibré, et de s'adresser à différents milieux, plusieurs moyens de prise de contact ont été choisis, avec différents points de départ.

Un premier groupe a été contacté dans la rue ou dans les magasins du centre historique (11 individus). Une de ces personnes a établi le contact avec des ateliers d'écriture. À la suite de la visite de ces ateliers, un grand nombre de personnes ont été conviées à participer elles aussi au projet. Ce sont des retraités issus de différents domaines professionnels qui profitent de l'occasion de s'adonner à leurs intérêts littéraires (24 individus). Le deuxième groupe le plus important a été recruté dans une maison de retraite catholique qui a accueilli non seulement des personnes moins privilégiées de la périphérie rurale d'Orléans, mais aussi quelques prêtres (19 individus). Enfin, par le biais d'un contact universitaire, nous avons pu inviter un cercle d'amis et d'autres connaissances personelles (22 individus).

Les chiffres indiqués comprennent également les personnes contactées par les participantes et participants, ce que Milroy & Gordon (2003) – entre autres – appellent « l'effet boule de neige ». Le travail de ces auteurs est dédié aux réseaux sociaux. Une relation étroite au sens d'un réseau social en résulte. Cela n'est pas le cas de nos groupes décrits. Les différentes catégories de personnes contactées ont plutôt la fonction d'assurer une approche sous divers angles de l'espace orléanais.

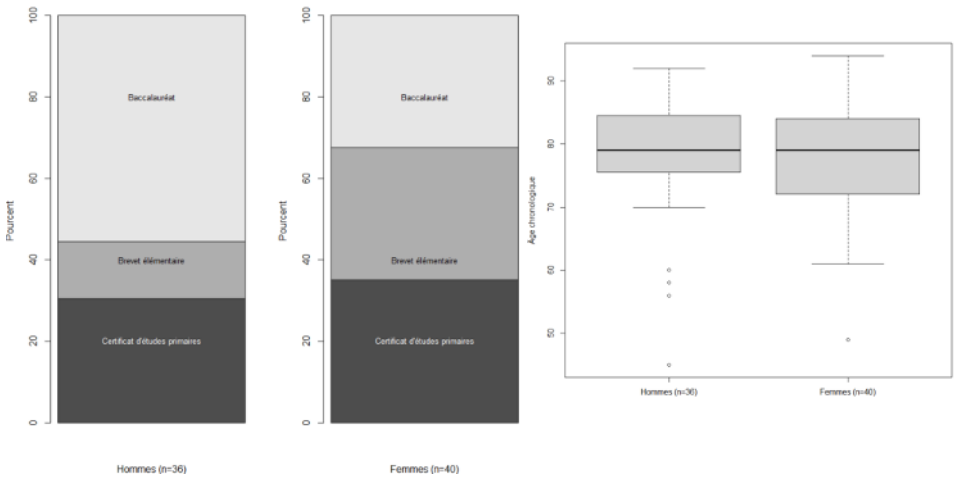

Fig. 1. Hommes et femmes de la première série : niveau d'étude et âge.

Pour la première série d'entretiens de 2005, les questions ont été formulées en reprenant en partie quelques mots clé du questionnaire ESLO1, avec plus de détails portant sur l'enfance, la famille et les expériences vécues en période de guerre et dans beaucoup de cas lors de l'occupation allemande d'Orléans. La deuxième série de 2012 reprenait la même série de questions (Gerstenberg 2015). Une troisième série en 2015 reprenait les points essentiels du récit biographique et ajoutait des aspects concernant la vie après laretraite et les attitudes linguistiques correspondantes. Les métadonnées personnelles étaient complétées par un questionnaire sur les traits personnels suivant le modèle des Big5 (Specht et al. 2014 ; voir aussi paragraphe 3.2).

Bien que le rôle de l'intervieweuse qui est également la chercheuse, lors des interviews soit devenu plus naturel au cours des séries, la plupart des souvenirs des prises de contact antérieures n'étaient que très faibles ce qui limite un effet possible de familiarité (Wagner & Tagliamonte 2017). De plus, la configuration des échanges, orientée sur l'histoire orale, établit un rapport presque asymétrique, l'expert étant la personne interviewée. Le style de l'interaction était marqué par l'intérêt personnel de la chercheuse porté à l'échange et par le comportement conversationnel à savoir une écoute active comprenant de nombreuses marques de retour (Slembrouck 2015). Ainsi, les personnes interviewées ont organisé de façon très autonome leurs récits à priori monologiques. Ce cadre ne privilégiait pas a priori l'usage d'un français vernaculaire de l'entretien classique labovien (Labov & Auger 1993). En revanche, un tel langage « ambitieux » (Gerstenberg 2011) pourrait apparaître comme un écho lointain de l'éducation linguistique, strictement orientée par l'écrit, de la Troisième République.

Les critères pour l'inclusion des personnes qui vivent un vieillissement « normal » (Frankenberg et al. 2021) privilégient les compétences pragmatiques (Messer 2015), c'est-à-dire la réalisation de paires d'adjacence au cours de l'échange et le respect des maximes de coopération (Grice 1975). Comme indicateur ultérieur de la condition physique, nous avons établi une durée minimum de l'entretien de 20 minutes ; si ce critère n'était pas réalisé, les personnes étaient considérées comme physiquement très fragiles ou fatiguées, ou un dialogue ne pouvait pas être lancé. Dans les deuxième et troisième séries, quelques cas de troubles cognitifs se sont manifestés, confirmés par les proches, les épouses ou employés des maisons de retraite (Gerstenberg & Hamilton 2022). Pour les études destinées à analyser les profils sociolinguistiques classiques, ces entretiens n'ont pas été inclus dans l'échantillon. Par contre, au sein de LangAge, des échantillons peuvent être constitués pour des études individuelles, permettant de retracer des parcours supra-individuels et

individuels à différents niveaux linguistiques (Kairet 2019, Fuchs et al. 2021). Au lieu de limiter l'analyse du langage de l'âge avancé à une évaluation en termes de maintien ou de diminution des performances cognitives, l'interaction entre les différentes compétences linguistiques et la prise en compte des contextes et usages sont au centre des intérêts. De cette manière, il sera possible de dresser un tableau dynamique et complexe des facettes du vieillissement langagier.

2.2 Le corpus « couples »

Le corpus LangAge dispose d'un potentiel pour aborder diverses questions de recherche concernant le vieillissement langagier d'un point de vue sociolinguistique. C'est pourquoi le sous-corpus « couples » a été créé. Il est composé des entretiens sélectionnés de 10 couples.

La plupart des participantes et participants sont nés dans les années 1920 et 1930, la participante la plus âgée en 1919 et les quatre personnes les plus jeunes après 1940. Trois hommes ont été à l'école jusqu'à 12/13 ans (Certificat d'études), un homme et cinq femmes jusqu'à 15/16 ans (Brevet d'études primaires), les autres jusqu'au baccalauréat. En 2005, les participantes et participants vivent dans leur propre maison ou appartement. En 2015, une participante, veuve, a déménagé dans une maison de retraite (Mme Léger). Tous sont en bonne santé et autonomes dans leurs activités quotidiennes. Au cours de ces dix années d'étude, l'état de santé cognitive de deux individus s'est dégradé au point qu'ils ont dû aller en maison de retraite spécialisée. Deux autres participants sont décédés durant cette période, mais leurs partenaires ont donné des interviews de suivi. Un participant a donné un unique entretien en 2015, il n'était pas disponible pour la première série de 2005. Une participante est présente seulement dans la série de 2015 à cause de problèmes techniques de l'enregistrement de 2005. Chaque individu a été interviewé seul en l'absence du partenaire. Seulement dans un cas, le couple se présentait inséparable dans l'entretien de la série de 2015 ; il y a donc un seul enregistrement pour deux personnes. Comme résumé dans le Tab. 1, le corpus de couples comprend donc un total de 33 entretiens de 34 personnes, d'une durée d'environ 45 minutes chacun, avec un total d'environ 237 500 tokens.

Tab. 1. Le corpus « couples ».

Corpus couples : série	Femmes	Hommes	Moyenne arithmétique de l'âge	Nombre de couples représentés
2005	9	9	73,8	10
2015	10	6	82,1	10

L'avantage de ce sous-corpus est d'une part de permettre un équilibre entre les femmes et les hommes. D'autre part, outre leur âge avancé, tous ont en commun d'avoir passé de nombreuses années de leur vie avec une personne très proche. Cela permet de restituer une grande partie des souvenirs de deux personnes différentes qui ont un statut socio-économique similaire. Ces facteurs contrebalancent donc quelque peu l'hétérogénéité de l'âge dû au parcours de vie individuel. De plus, au cours de la vie à la retraite, on observe, dans notre corpus, une retraite sociale et spatiale progressive. Sous ces conditions, le couple se présente comme cellule de plus en plus importante de la communication quotidienne, avec un possible alignement au cours des années et de possibles changements après la perte de l'autre.

Dans le cadre des études sur la communication des couples par rapport à l'âge, le sous-corpus ajoute la perspective longitudinale à la perspective dominante de coupe transversale (Sillars & Zietlow 1993 ; Luo et al. 2020).

2.3 La transcription et le respect de l'anonymat

Les fichiers audio sont disponibles en format *.wav,[ii] transcrits manuellement dans un format XML aligné, à l'aide du logiciel Transcriber (*.trs). Les principes de la transcription orthographique s'inspirent du guide ESLO (Gerstenberg et al. 2018).

Afin de garantir la confidentialité des données personnelles des individus et de leurs proches, ainsi que pour des raisons éthiques, les enregistrements et les transcriptions des entretiens ont été anonymisés. Cette démarche a été garantie par l'autorisation signée lors de l'entretien. C'est à cette condition que les participants ont signé l'autorisation de publier les enregistrements en ligne. Les versions anonymisées (fichiers son et transcriptions) portent le suffixe _a[nonymisée]. Une deuxième version _o[riginal] a été retenue pour l'usage interne. L'anonymisation est appliquée à tous les noms propres de personne, lieu et institutions (par ex. écoles, entreprises, églises, etc.) qui, dans le contexte où ils sont cités, peuvent révéler l'identité des individus ou des personnes mentionnées. En outre, les dates qui ont une pertinence particulière pour les personnes interviewées (par ex. date de naissance complète) sont anonymisées.

Dans les transcriptions, le sigle NP 'nom propre' remplace l'élément personnel, suivi par une lettre qui spécifie la référence (Tab. 2). Les localités ou villes qui comptent moins de 5000 résidents dans la période de transcription sont systématiquement anonymisées.

Tab. 2. Sigles utilisés pour l'anonymisation.

Sigle	Signification	Exemple
NPp	nom propre, prénom	*Jeanne*
NPf	nom propre, nom de famille	*Moreau*
NPl	nom propre, localité	*Orléans*
NPc	nom propre, pays (country)	*France*
NPe	nom propre, entreprise ou organisation	*Renault*
NPl(A)	dérivé et catégorie	*fond familial fleuryssois*
NPl(N)	dérivé et catégorie	*il était fleuryssois*
NUM	années ou dates personnelles significatives, aptes à identifier une personne ou ses propres	*1 janvier 1968*
anon	motifs divers	*Allô ...* [coup de téléphone]

Le sigle *anon* nécessite une précision. Il est employé pour des mots individuels ou pour des paragraphes entiers dans des cas spécifiques comme des coups de téléphone pendant l'entretien afin de protéger l'anonymat des personnes qui ne participent pas activement à l'interview mais qui interviennent de manière occasionnelle. *Anon* est également utilisé pour d'autres motifs privés ou éthiques qui sont traités au cas par cas, quand une description très détaillée pourrait permettre d'identifier une locutrice ou ses proches, par exemple. L'application des ces règles nécessite donc une bonne connaissance des conditions personnelles et locales.

La dernière phase du procès d'anonymisation consiste à modifier le fichier son avec un script PRAAT (« anonymiser les fichiers audios de longue durée », *anonymise long sound* Hirst 2013 ; PRAAT : Boersma & Wenink 2021) qui manipule les propriétés acoustiques afin de cacher le contenu de ce qui est dit tout en conservant les caractéristiques prosodiques.

Les conversions entre le format de Transcriber (*.trs) et PRAAT (*.textGrid) ainsi que l'adaptation de la transcription pour l'analyse CORDIAL sont réalisées avec des scripts PYTHON (autrice Valerie Hekkel, ancien membre de l'équipe). D'autres étapes de la gestion des transcriptions et des métadonnées sont réalisées par TUSTEP. Le déroulement du travail a été conçu par Julie Kairet (ancien membre de l'équipe). Il commence avec la première transcription qui est suivie par les phases de révision, puis le contrôle de la segmentation et la révision finale pour terminer. Au procès d'anonymisation a été intégré une dernière phase de révision de la transcription afin d'identifier les erreurs de transcription restantes (par ex. fautes de frappes, erreurs d'accord). La révision est effectuée avec l'analyse d'erreurs de l'outil CORDIAL (Synapse 2008) manuellement contrôlées et intégrées dans le document *.trs.

Au-delà de l'anonymisation des transcriptions et des fichiers son, des pseudonymes ont été attribués aux individus, tirés des noms les plus fréquents de leur région et cohorte, selon la statistique de l'INSEE, en évitant les véritables noms des personnes représentées dans l'échantillon. Pour les couples, les mêmes pseudonymes ont été choisis, ce qui correspond à l'usage réel : les couples partagent les mêmes noms de famille.

3 L'exploration longitudinale :LaBB-CAT

3.1 LaBB-CAT pour la gestion de ressources de la langue parlée

Notre objectif principal était de créer une architecture de corpus durable où les données audio, les transcriptions et les annotations de LangAge pourraient être enregistrées afin d'être utilisées de manière flexible dans différents contextes de la recherche sur le vieillissement langagier. Pour ce faire, le choix s'est porté sur le logiciel open source *Language, Brain and Behaviour – Corpus Analysis Tool* (LaBB-CAT, Fromont 2012–2021), développé par Robert Fromont et Jennifer Hay (University of Canterbury, Fromont & Hay 2012) dans le cadre du projet *Origins of New Zealand English.* Il est également appliqué dans le projet des *Carolinas Conversation Collection* qui comprend des échantillons de l'âge avec et sans troubles cognitifs (dementia, Pope & Davis 2011). Ce logiciel est basé sur navigateur et offre une architecture cohérente et consistante pour le stockage, la gestion et la recherche de données de langue parlée. Il consiste en deux sections. Le « backend » fonctionne sur une machine virtuelle, tandis que le « frontend » est accessible par l'interface web (Fig. 2).

LangAge est le premier corpus en langue française géré par LaBB-CAT (voir Fromont 2019 pour d'autres corpus qui utilisent ce logiciel) où il est disponible en libre accès avec une license *Creative Commons* (CC BY-NC-SA 4.0). Une raison importante pour le choix de ce logiciel est qu'il présente une interface qui correspond pleinement aux principes de trouvabilité (*Findability*), d'accessibilité (*Accessibility*), d'interopérabilité *Interoperability*), et de réutilisabilité (*Reusability*) (FAIR, v. entre autres Higman et al. 2019).

En particulier, LaBB-CAT a été spécialement développé pour travailler avec la langue parlée. Il présente donc une interface aisément utilisable qui garantit un accès rapide aux fichiers de transcription et aux fichiers audio (ou video) correspondants. Toutes les séries comprises, LangAge compte, à l'heure actuelle, un total de plus de 740.000 tokens (Fig. 2).

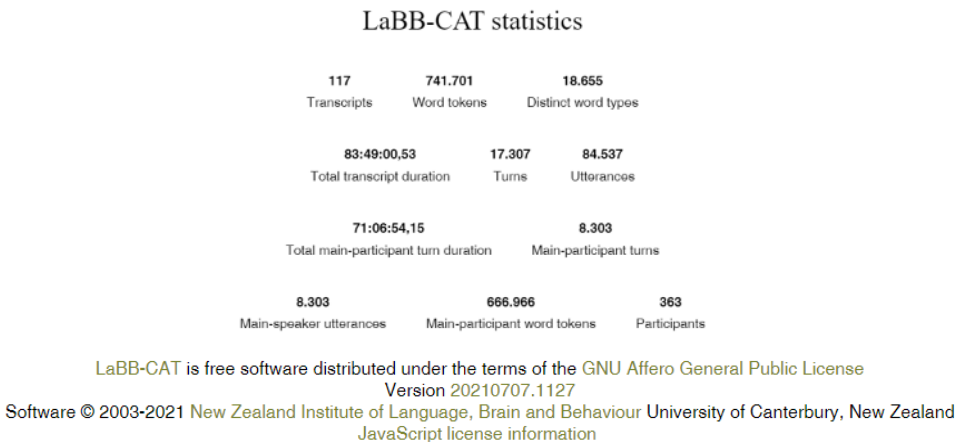

Fig. 2 : Frontend LaBB-CAT, compte administration : statistique.

Les transcriptions enregistrées dans LaBB-CAT peuvent avoir été réalisées sous différents formats, allant de Transcriber (*.trs) pour la transcription simple à ELAN (*.eaf) pour l'annotation syntaxique en passant par PRAAT (*.textGrid) pour l'analyse phonétique. De plus, LaBB-CAT permet de créer un jeu de métadonnées différencié, indispensable à la gestion d'un corpus longitudinal (paragraphe 3.2). Il permet aussi d'intégrer plusieurs niveaux d'annotation manuelles ou automatiques, afin de faciliter la recherche et l'analyse linguistique des données (paragraphe 3.3). Enfin, LaBB-CAT différencie les types de droits d'accès, par exemple, en fonction du type d'autorisation donnée par les participants ou selon le type d'utilisation par un membre de l'équipe, partenaires externes ou personnes sans lien direct à l'équipe. Cette stratification garantit le principe FAIR d'accessibilité du corpus respectant les aspects juridiques et éthiques concernés, cela dans une perspective d'un travail d'équipe de longue durée et basé sur des standards transparents. C'est pour ces caractéristiques que nous avons choisi LaBB-CAT au lieu d'autres systèmes, également utilisés pour la gestion des corpus de langue parlée, tels que EXMARaLDA/COMA (Schmidt & Wörner 2014) ou ANNIS (Krause & Zeldes 2016).

3.2 Le jeu de métadonnées

L'interface de LaBB-CAT permet de configurer un jeu d'étiquettes de métadonnées unique et, pour cette raison, apte à représenter le groupe de témoins de LangAge dans ses aspects les plus pertinents et spécifiques, ainsi que de les organiser en macro-catégories. C'est grâce à cela que le système de métadonnées, mis en place via LaBB-CAT pour le corpus LangAge, combine un degré de standardisation nécessaire pour assurer la comparabilité de nos données avec celles des corpus sociolinguistiques établis (Sankoff 2017) et la flexibilité indispensable pour les adapter aux exigences de LangAge. Les métadonnées ainsi que les contenus sont rédigés dans la langue de l'outil, en anglais.

Nous avons décidé de gérer les métadonnées du corpus LangAge dans deux macro-groupes d'attributs (format d'importation : *.csv), autrement dit, sur les données personnelles (« participant attributes », Fig. 3) et de l'enregistrement (« transcript attributes », Fig. 4). Nous avons décidé de stocker des métadonnées personnelles et de l'enregistrement de façon distincte pour chaque événement de communication. Ce choix permet la mise à jour des informations personnelles, un déménagement en maison de retraite par exemple (voir le tableau ci-dessous). Cela est possible grâce à un identifiant stable des personnes (« participant ID »), composé de trois chiffres combinés, pour chaque entrevue, avec une lettre qui indique la série (Fig. 3). Avec cette différenciation, nous avons

pû établir un schéma d'annotation de métadonnées sensible aux aspects temporels qui est intégré par des éléments textuels tels que les attributs « courte biographie » (Fig. 3) et « activités ». Ces attributs enregistrent les changements entre les entretiens. Ainsi, les différentes valeurs des attributs personnels ne concernent pas « le témoin », mais l'individu au moment de l'entretien. Cela montre la manière dont les facteurs techniques aident à réaliser le but du corpus à savoir de rendre accessible le vieillissement langagier dans une perspective longitudinale et évolutive.

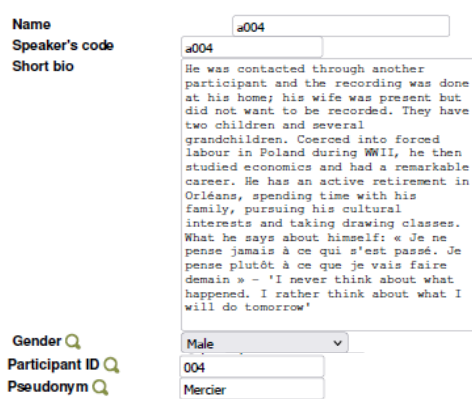

Fig. 3 : Extrait des données personnelles.

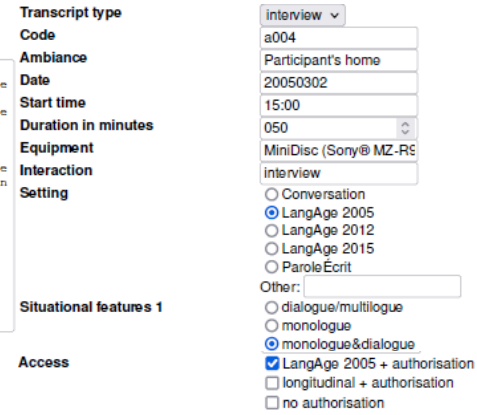

Fig. 4 : Extrait des données sur l'enregistrement.

Les métadonnées personnelles telles que représentées dans les Fig. 3 couvrent les traits personnels canoniques ainsi que les catégories plus spécifiques du corpus, organisées sous les catégories (1) personnelle (par exemple l'âge et sexe biologique), (2) socio-économique (niveau de scolarisation, classification socioprofessionnelle qui s'inspire de la classification INSEE et du système appliqué dans ESLO1, cf. Mullineaux & Blanc 1982), (3) des conditions de la retraite (état civil, entrée en retraite), (4) biographique (profession des parents, état civil, enfants et descendants, bilinguisme et langue(s) acquise(s) comme langue étrangère), et (5) un bref portrait en texte courant, y compris une citation de l'entrevue qui décrit la personne. Personne n'a indiqué vivre une relation autre qu'hétérosexuelle ou de n'appartenir à aucune des catégories « féminin » ou « masculin ». L'équilibre hommes-femmes est assez bien représenté. En effet, au total, les expériences de quarante femmes et de trente-six hommes ont été recueillies. Les informations personnelles recueillies lors de l'entretien biographique concernant la formation scolaire sont classées en trois groupes et celles concernant les activités socioprofessionnelles sont répertoriées en quatre groupes. Même si ces classifications socio-économiques standard sont plutôt discutables (Ash 2013), elles permettent de situer le corpus dans le tableau des autres ressources du français parlé.

Les métadonnées de l'enregistrement représentées dans la Fig. 4 comprennent l'appartenance à un protocole d'enregistrement (LangAge 2005, LangAge 2012, LangAge 2015 ou les autres modules; voir paragraphe 2.1), l'équipement technique, les informations sur le lieu, la date et la durée de l'enregistrement et, important pour le type d'accès aux données, l'autorisation qui couvre ou non la publication en ligne.

3.3 Recherche et analyse de données

Une fois téléchargé sur LaBB-CAT, le corpus LangAge a été « tokenisé » par LaBB-CAT pour être automatiquement consultable. La recherche dans les transcriptions et annotations

peut inclure toutes les personnes ou un sous-corpus selon les paramètres linguistiques et sociolinguistiques (Fig. 5).

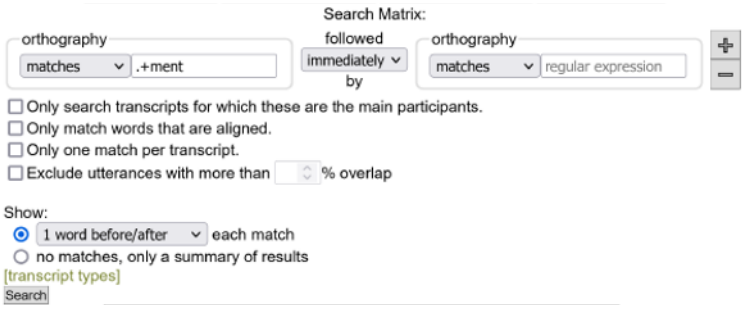

Fig. 5 : Interface de recherche.

Comme le montre la Fig. 5, la recherche peut concerner un mot individuel, une combinaison de plusieurs mots et d'expressions régulières.

Une première version lemmatisée du corpus avec CORDIAL est disponible via le site du corpus. L'intégration d'une version lemmatisée et annotée par marque de parties du discours dans LaBB-CAT est prévue. De même, nous planifions de télécharger sur LaBB-CAT des annotations d'entretiens réalisées manuellement par ELAN et PRAAT et concernant, entre autres, la négation de phrase ou les récits d'expériences personnelles.

Les résultats de recherche sont organisés sous forme de mots-clés en contexte (Keyword-in-Context KWIC, Fig. 6) ce dernier peut être librement défini lors de la recherche. Les occurrences sont regroupées par entrevue, en cliquant sur le mot cible. On accède alors à la transcription de l'entrevue, accompagné du son.

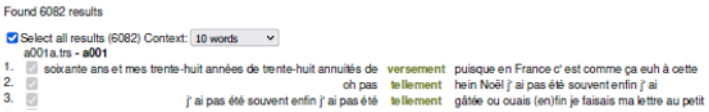

Fig. 6 : Résultats de recherche de l'expression régulière ".+ment".

Les traits FAIR d'interopérabilité et réutilisabilité sont réalisés dans les modes d'exportation de la concordance : sous forme de tablier (*.csv), d'extraits de transcriptions de format PRAAT (*.textGrid) et d'extraits des fichiers sons correspondants. Cela ouvre la porte à des emplois variés et à la recherche multidimensionnelle sur les implications linguistiques du vieillissement (Fuchs et al. 2021).

4 Conclusion : un corpus au pluriel

À quel point un corpus peut-il être FAIR? Le choix de respecter les principes FAIR naît des détails du processus de création du corpus et est limité par la nature des données. Ainsi, le caractère de « corpus au pluriel » et le travail prévu de l'intégration de la base de données prolongent le processus de finalisation. L'enregistrement dans des bases de données spécialisées rendra le corpus plus facile à (re)trouver (*findable*). La composition des métadonnées CMDI et la création d'un identificateur perenne (DOI) sont prévus.

Certaines séries ne peuvent être mises en libre accès (*accessible*) en raison de la nature de leurs données. Les fonctionnalités de LaBB-CAT permettent de paramétrer le type

d'utilisation, sous une licence *Creative Commons*. La grande flexibilité de LaBB-CAT permet l'exportation et l'exploitation des transcriptions avec différents outils et dans différents contextes (interopérable et réutilisable). Au-delà de LaBB-CAT, une partie du corpus a été transférée dans un corpus d'histoire orale (Pagenstecher 2020). Le déplacement d'éléments dans d'autres corpus ou dans d'autres dépôts d'archivage à long terme reste possible.

Les limites de l'adoption des principes FAIR seront repoussées ; différents scénarios de composition du corpus et de sous-corpus ainsi que la possibilité de les intégrer dans différentes bases de données et de dépôts d'archivage à long terme constituent une raison suffisante pour parler ici d'un corpus au pluriel.

Bibliographie

accès au corpus: http://www.langage-corpora.org

LaBB-CAT = Fromont, R. (2012–2021). *LaBB-CAT: Language, Brain & Behaviour Corpus Analysis Tool.* University of Canterbury (NZ): New Zealand Institute of Language, Brain and Behaviour. https://labbcat.canterbury.ac.nz/system/ (01.12.2021)

Abouda, L. & M. Skrovec. (2018). Pour une micro-diachronie de l'oral : le corpus ESLO-MD. *Congrès Mondial de Linguistique Française 6*. art. 11004 [1-11]. https://doi.org/10.1051/shsconf/20184611004.

Ash, S. (2013). Social Class. In J. K. Chambers & N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 350–367.

Baude, O. & C. Dugua. (2011). (Re)faire le corpus d'Orléans quarante ans après : 'quoi de neuf, linguiste ?'. *Corpus 10*. 99–118. https://doi.org/10.4000/corpus.203610.4000/corpus.2036.

Bergounioux, G. (2010). Mai 68 vu d'Orléans : la geste et la parole. *Congrès Mondial de Linguistique Française 2*. 1859–1875. https://doi.org/10.1051/cmlf/2010104.

Bergounioux, G., J. Baraduc & C. Dumont. (1992). L'Etude Socio-Linguistique sur Orléans (1966–1991) : 25 ans d'histoire d'un corpus. *Langue française 1*. 74–93. https://doi.org/10.3406/lfr.1992.5812.

Biggs, P. & M. Dalwood. (1978). *Les Orléanais ont la parole.* München: Langenscheidt-Hachette.

Boersma, P. & D. Weenink. [2021]. Praat: Doing Phonetics by Computer. PRAAT. http://www.praat.org.

Buchstaller, I. & S. E. Wagner. (2017). Introduction: Using Panel Data in the Sociolinguistic Study of Variation and Change. In S. E. Wagner & I. Buchstaller (eds.), *Panel Studies of Variation and Change*, 1–18. New York: Routledge.

ESLO = Laboratoire Ligérien de Linguistique. (2014). *Enquêtes SocioLinguistiques à Orléans ; ESLO1 : Corpus linguistique 1968–1974; ESLO2 : 2008–2014*. Université d'Orléans: LLL. http://eslo.huma-num.fr/

FAIR = GO FAIR International Support and Coordination Office. (2021). *FAIR Principles.* Leiden (NL), Paris (F), Hamburg (D): GO FAIR. https://www.go-fair.org/fair-principles/ (01.12.2021)

Flamein, H. & I. Eshkol-Taravella. (2021). Exploitation du corpus Enquêtes sociolinguistiques à Orléans (ESLO) par les outils du traitement automatique des langues et de la géomatique. *Humanités numériques 3*. https://doi.org/10.4000/revuehn.1911.

Frankenberg, C., J. Weiner, M. Knebel, A. Abulimiti, P. Toro, C. J. Herold, T. Schultz & J. Schröder. (2021). Verbal fluency in normal aging and cognitive decline: Results of a longitudinal study. *Computer Speech & Language 68*. 101195. https://doi.org/10.1016/j.csl.2021.101195.

Fromont, R. (2019). Forced Alignment of Different Language Varieties using LaBB-CAT. *International Congress of Phonetic Sciences 2019*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1376.pdf

Fromont, R. & J. Hay. (2012). LaBB-CAT: an Annotation Store. In P. Cook & S. Nowson (eds.), *Proceedings of Australasian Language Technology Association Workshop 2012*, 113–117. Dunedin (NZ): Otago University. https://aclanthology.org/U12-1015.pdf (01.12.2021).

Fuchs, S., L. L. Koenig & A. Gerstenberg. (2021). A Longitudinal Study of Speech Acoustics in Older French Females: Analysis of the Filler Particle *euh* across Utterance Positions. *Languages 6*(4). 211. https://doi.org/10.3390/languages6040211.

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

Gerstenberg, A. (2015). A Sociolinguistic Perspective on Vocabulary Richness in a Seven-Year Comparison of Older Adults. In A. Gerstenberg & A. Voeste (eds.), *Language Development: the Lifespan Perspective*, 109–127. Amsterdam: Benjamins. https://doi.org/10.1075/impact.37.06ger.

Gerstenberg, Annette, Valerie Hekkel & Julie Kairet. 2018. *Corpus LangAge: Transcription Guide.* University of Potsdam: Department of Romance Studies. doi.org/10.5281/zenodo.6444538.

Gerstenberg, A. & H. E. Hamilton. (2022). Older adults' conversations and the emergence of 'narrative crystals': A new approach to frequently told stories. *Narrative Inquiry*. https://doi.org/10.1075/ni.21075.ger. 1–34.

Gerstenberg, A. & J. Kairet. (2021). Prosodic Features and Situational Settings: Doing Reading Aloud in a French Writing Class. In A. Teixeira Kalkhoff, M. Selig & C. Mooshammer (eds.), *Prosody and Conceptional Variation*, 141–156. Frankfurt am Main: Lang.

Gerstorf, D., G. Hülür, J. Drewelies, P. Eibich, S. Duezel, I. Demuth, P. Ghisletta, E. Steinhagen-Thiessen, G. G. Wagner & U. Lindenberger. (2015). Secular changes in late-life cognition and well-being: Towards a long bright future with a short brisk ending? *Psychology and Aging 30*(2). 301–310. https://doi.org/10.1037/pag0000016.

Grice, H. P. 1975. Logic and conversation. In: P. Cole & J. L. Morgan (eds.), *Syntax and semantics. Vol. 3: Speech acts*, 41–58. New York: Academic Press.

Hekkel, V. (2021). *Eine soziolinguistische Betrachtung von parce que-Strukturen in Synchronie und Diachronie.* Universität Potsdam: publishUP. https://doi.org/10.25932/publishup-51396.

Higman, R., D. Bangert & S. Jones. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights 32*(18). 1–9. https://doi.org/10.1629/uksg.468.

Hirst, D. (2013). Anonymising Long Sounds for Prosodic Research. In B. Bigi & D. Hirst (eds.), *Tools and Resources for the Analysis of Speech Prosody*. TRASP 2013, 36–37. Aix-en-Provence: Laboratoire Parole et Langage. http://www2.lpl-aix.fr/~trasp/Proceedings/19906-trasp2013.pdf (01.12.2021)

INA (ed.). 1969. *Affaire classée ? – Un reportage de Jacques et Maurice Gugowson ; Pierre André* [Documentation vidéo de 1969, publié le 29.03.2019, mis à jour le 08.07.2021]. Paris: Institut national de l'audiovisuel. https://www.ina.fr/contenus-editoriaux/articles-editoriaux/1969-la-rumeur-d-orleans-un-retour-au-moyen-age/ (01.12.2021).

Krause, T. & A. Zeldes. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities 31*(1). 118–139.

Kairet, J. (2019). Silent Pause Duration and Distribution in Older Womens' Speech: a Case Study with 4 Within-speaker Comparison. In O. Niebuhr, J. Neitsch, S. Berger, K. Fischer, J. Michalsky, S. Eisenberger & M. Jelínek (eds.), *Proceedings of the 1st International Seminar on the Foundations of Speech – Pausing, Breathing and Voice. University of Southern Denmark,*

Sønderborg, Denmark, December 1–3, 2019, vol. 1, 61–63. Sønderborg: University of Southern Denmark.

Labov, W. & J. Auger. 1993. The Effect of Normal Aging on Discourse: A Sociolinguistic Approach. In H. H. Brownell & Y. Joanette (eds.) *Narrative Discourse in Neurologically Impaired and Normal Aging Adults*, 115–133. San Diego (CA): Singular.

Lonergan, J., J. Kay & J. Ross. (1974). *Étude sociolinguistique sur Orléans. Catalogue d'enregistrements.* Colchester: Typoscript.

Luo, M., M. Neysari, G. Schneider, M. Martin & B. Demiray. (2020). Linear and Non-Linear Age Trajectories of Language Use: A Laboratory Observation Study of Couples' Conflict Conversations. *The journals of gerontology. Series B, Psychological sciences and social sciences*. https://doi.org/10.1093/geronb/gbaa041.

Messer, R. H. 2015. Pragmatic Language Changes During Normal Aging: Implications for Health Care. *Healthy Aging & Clinical Care in the Elderly* 7. 1–7. https://doi.org/10.4137/HaCCe.S22981.

Milroy, L. & M. J. Gordon. (2003). *Sociolinguistics. Method and interpretation*. Oxford: Blackwell.

Mullineaux, A. & M. Blanc. (1982). The Problem of Classifying the Population Sample in the Sociolinguistic Survey of Orleans (1969) in Terms of Socio-Economic, Social and Educational Categories. *ITL – International Journal of Applied Linguistics* (55). 3–37. https://doi.org/10.1075/itl.55.01mul.

Pagenstecher, C. (2020). *Oral-History.Digital – Interviewsammlungen als Forschungsdaten.* Berlin: Freie Universität Berlin.

Pichler, H., S. E. Wagner & A. Hesson. (2018). Old-age Language Variation and Change: Confronting Variationist Ageism. *Language and Linguistics Compass 12*(6). 1–21. 10.1111/lnc3.12281.

Pope, C. & B. H. Davis. (2011). Finding a Balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory 7*(1). 546. https://doi.org/10.1515/cllt.2011.007.

Sankoff, G. 2017. Before There Were Corpora. In S. Wagner & I. Buchstaller (eds.), *Panel Studies of Variation and Change*, 21–51. New York: Routledge.

Schmidt, T. & K. Wörner. (2014). EXMARaLDA. In J. Durand, U. Gut & G. Kristoffersen (eds.), *The Oxford handbook of corpus phonology* (Oxford handbooks in linguistics), 402–419. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.030.

Sillars, A. L. & P. H. Zietlow. (1993). Investigations of Marital Communication and Lifespan Development. In N. Coupland & J. Nussbaum (eds.), *Discourse and Lifespan Identity*, 237–261. Newbury Park (CA): SAGE.

Slembrouck, S. (2015). The Role of the Researcher in Interview Narrative. In: A. De Fina & A. Georgakopoulou (eds.), *The Handbook of Narrative Analysis*, 237–254. Oxford: Blackwell.

Specht, J., W. Bleidorn, J. J. Denissen, M. Hennecke, R. Huttemann, C. Kandler, M. Luhmann, U. Orth, A. K. Reitz, J. Zimmermann, J. J. A. Denissen & R. Hutteman. (2014). What drives adult personality development? A comparison of theoretical perspectives and empirical evidence. *European Journal of Personality 28*(3). 216–230. https://doi.org/10.1002/per.1966.

Synapse. (2008). *Cordial Analyseur version 14.0.* Paris: Synapse Développement.

TUSTEP. (2021). *Tübinger System von Textverarbeitungs-Programmen.* Tübingen: Universität Tübingen; Zentrum für Datenverarbeitung.

Wagner, S. E. & S. A. Tagliamonte. (2017). What Makes a Panel Study Work? Researcher and Participant in Real Time. In: S. Wagner & I. Buchstaller (eds.), *Panel Studies of Variation and Change*, 213–232. New York: Routledge.

[i] La section corpus « couples » a été conçue et réalisée par Friederike Schulz, la section sur la pseudonymisation par Eman El Sherbiny Ismail, la section sur l'architecture de la base de

données LaBB-CAT par Marta Lupica Spagnolo ainsi que la rédaction générale réalisée avec Annette Gerstenberg. Nous tenions à souligner l'apport considérable des membres de l'équipe, anciens et actuels, notamment, par ordre alphabétique, Hélène-Véronique Essomba-Etama, Valerie Hekkel, Freya Hewett, Julie Kairet, Adélie Soumier-Vendé, Erick Velázquez Godínez.

[ii] Dans la première série, l'enregistrement a été réalisé avec un MiniDisc (Sony® MZ-R91) et le condenser microphone Philips® (SBC ME570). A partir de 2012, le MiniDisc a été remplacé par un Olympus Linear PCM Recorder (LS-5 with 44.1kHz).