Chapitre 9

Description du corpus de parole

Un total de 304 locuteurs ont pu être enregistrés pendant la collecte de données, totalisant 26 h de parole spontanée. Parmi eux, 260 ont été enregistrés lors de sessions d'examen CLES sur les campus de Grenoble et Valence de l'université Grenoble Alpes ; et 44 locuteurs ont été enregistrés lors de « fausses sessions CLES » dans les universités de Waseda et Dōshisha, au Japon.

Dans ce chapitre, nous présentons les trois corpus de données qui ont vu le jour à partir de ces enregistrements : un corpus CLES-FR de locuteurs francophones, un corpus CLES-JP de locuteurs japonophones, et un corpus CLES-EN de locuteurs anglophones natifs. Nous présenterons ensuite les deux dépôt de données publics associés, et les annotations manuelles qui ont été réalisées sur une partie des données pour confectionner un échantillon *gold standard*.

9.1 Corpus CLES-FR

Nous commencerons par décrire l'ensemble des enregistrements collectés lors des sessions du CLES, puis nous présenterons la sélection des données effectuée pour les analyses de notre travail de recherche.

Les 260 locuteurs enregistrés pendant des sessions CLES sont répartis en 232 binômes, 15 trinômes et 13 monômes. Les 13 monômes sont issus de session CLES B1 et présentent une situation de parole différente des autres groupes. Les 247 autres locuteurs ont été enregistrés lors de session CLES B2. La distribution homme/femme est équilibrée (130 femmes, 130 hommes). Une grande majorité des locuteurs a été certifiée B2 lors de l'examen (n=151, 58%), contre 75 B1 (29%) et 34 non-validés (13%). Quarante-cinq locuteurs (17%) n'ont pas déclarer le français comme langue maternelle, parmi eux

22 locuteurs n'ont déclaré aucune langue, 6 locuteurs ont déclarer une langue arabe, 3 une langue chinoise et on compte encore 11 autres langues déclarées par une ou deux personnes à la fois.

Nous souhaitons observer les patterns de pauses et d'accentuation chez les locuteurs francophones B1 ou B2, aussi nous n'avons conservé dans le corpus final que les candidats ayant déclarer le français comme langue maternelle, ayant validé l'un des deux niveaux à l'examen, et nous avons également mis de côté les 13 monômes du CLES B1 car ils présentent une situation de parole trop différente des autres locuteurs. Le corpus final obtenu compte ainsi 170 locuteurs. Nous ferons dorénavant référence à ce corpus francophone B1/B2 par le nom de « CLES-FR ».

Le corpus CLES-FR comprend 99 locuteurs de niveau global B2 (58%) et 71 de niveau global B1 (42%). Le niveau obtenu spécifiquement pour la compétence d'interaction orale est B2 pour 118 locuteurs (69%) et B1 pour 52 locuteurs (31%). La distribution homme/femme reste relativement équilibrée avec 89 femmes (52%) et 81 hommes (48%). Parmi les 170 locuteurs, 11 sont enregistrés en trinômes (6%).

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.2 Corpus CLES-JP

Vingt-neuf étudiants de langue maternelle japonaise ont été enregistrés dans une situation de parole similaire à celle du corpus CLES-FR. Il y a 17 femmes (59%) et 12 hommes (41%). Leur niveau de compétence en anglais est estimé à partir de résultats obtenus à différentes certifications (TOEFL, IELTS, ou Eiken principalement) : 5 d'entre eux sont de niveau équivalent B1 (17%), 15 de niveau équivalent B2 (52%), et 9 de niveau équivalent C1 (31%). Les participants ont été répartis en binômes en faisant en sorte que chaque participant ait un niveau comparable à celui de son interlocuteur. L'un des 15 binômes enregistrés est constitué d'un étudiant de niveau C1 et d'un enseignant d'anglais de langue maternelle japonaise qui a dû participer suite à l'absence d'un candidat. Les tours de parole de l'enseignant ne font pas partie du corpus CLES-JP.

La principale différence avec les locuteurs du CLES-FR, hormis la langue maternelle, est le fait que les participants sont volontaires, rémunérés, et n'ont pas d'enjeu spécifique comme l'obtention d'un diplôme.

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.3 Corpus CLES-EN

Le corpus complémentaire de locuteurs anglophones natifs est quant à lui constitué de 14 locuteurs, tous originaires des États-Unis, et inscrits en licence dans une université américaine. Ils ont entre 20 et 22 ans ($M = 20,5$), 9 d'entre eux sont des femmes et 5 sont des hommes.

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.4 Publication des données

Les enregistrements des sessions CLES ont pu être organisées dans le cadre de l'examen et utilisées intégralement pour notre recherche. Toutefois, seuls les enregistrements pour lesquels l'ensemble des participants ont donné leur accord pour la publication des données a pu être mis à disposition de la communauté. Parmi les 260 locuteurs enregistrés, 162 ont donné leur accord (62%), et 138 d'entre eux font partie d'un enregistrement où tous les locuteurs ont donné leur accord (et donc diffusable en l'état).

Un corpus public d'une partie des enregistrements CLES a ainsi été mis à disposition sur la plateforme Ortolang[1]. Il réuni 62 enregistrements de 128 locuteurs, totalisant 10 h de parole. Parmi les locuteurs, 119 ont déclaré avoir le français pour langue maternelle (93%). On compte 61 femmes (48%) pour 67 hommes (52%). La durée moyenne des enregistrements est de 9 min35 s (min. 5 min12 s, max 14 min30 s). Le résultat obtenu au CLES est B2 pour 62 d'entre eux (48%), 50 ont validé le niveau B1 (39%) et 16 n'ont rien validé (13%).

Les corpus CLES-JP et CLES-EN ont quant à eux pu être entièrement mis à disposition sur la même plateforme **(Lien pérenne)**.

9.5 Annotations *gold standard*

Un échantillon du corpus CLES-FR a été transcrit semi-automatiquement puis corrigé pour constituer un sous-corpus de référence. Vingt enregistrements ont été sélectionnés aléatoirement avec pour contrainte de contenir 20 candidats certifiés B2 et 20 certifiés B1, et un équilibre homme/femme. Le travail d'annotation a été effectué

[1] https://hdl.handle.net/11403/cles-spontaneous-english

par Nathanaël Berthet, stagiaire d'excellence de licence d'informatique à l'université Grenoble Alpes. Une transcription automatique des enregistrements à d'abord été effectuée l'aide du logiciel Whisper (Radford et al., 2022, modèle `base` multilingue), puis manuellement corrigé et segmenté en locuteurs.

Du côté des corpus CLES-JP et CLES-EN, les corpus ont d'abord été automatiquement segmentés en locuteurs avec la pipeline de diarisation `pyannote.audio` (Bredin, 2023), puis vérifiés manuellement dans leur intégralité.