Description du corpus

Origines et conception du CFPB

Si toute collecte systématique de français parlé à Bruxelles remplit un vide dans les corpus actuellement accessibles, des données comparables à celles déjà disponibles pour d'autres variétés s'avèrent encore plus utiles. Même si les normes du français ont évolué au cours du dernier quart de siècle pour intégrer un pluricentrisme, Paris continue à représenter la variété standard la plus neutre. Une comparaison avec des données parisiennes existantes constitue donc un avantage pour déterminer d'éventuelles spécificités bruxelloises. Par ailleurs, une technique de collecte qui détourne l'attention de la forme linguistique pour se centrer sur le contenu est essentielle pour contourner l'obstacle de la pratique consciemment surveillée dans un contexte d'entrevue. En outre, un contenu centré sur la ville est particulièrement judicieux dans le cadre de Bruxelles.

Par chance, un corpus parisien se focalisant sur la perception de la ville existe déjà: le Corpus de français parlé parisien (CFPP2000) lancé par Sonia Branca-Rosoff et actuellement dirigé par Florence Lefeuvre avec l'aide linguistique de Mat Pires et le soutien informatique de Serge Fleury. L'équipe du CFPP2000 a développé son corpus 'en vue d'une description du français « commun »' qui bénéficierait bien entendu d'une dimension diatopique. Notre corpus de français parlé à Bruxelles (CFPB) converge donc avec le projet parisien dans sa conception et dans ses thèmes. D'une part, la collecte des données se réalise au travers d'entretiens laboviens semi-dirigés par des enquêteurs plus ou moins empathiques. Le corpus bruxellois comprend également des dialogues (plus faciles à transcrire mais potentiellement assez formels vu le contexte de l'interview) et des multilogues (dont la difficulté de transcription peut être compensée par la spontanéité des échanges entre informateurs qui se connaissent et sont susceptibles d'entrer dans des échanges naturels). Finalement, le protocole de collecte élicite un éventail de tâches discursives variées incluant narration, description et argumentation. D'autre part, les deux corpus partagent le thème des relations des habitants à leur quartier, leur commune et leur ville. En plus d'attirer l'attention des informateurs sur le contenu plutôt que la forme, ce thème est accessible à tous, indépendamment de l'âge ou du niveau d'éducation. Toutefois, le CFPB ne se contente pas de reproduire le protocole de la collecte parisienne et le questionnaire a été adapté à la situation bruxelloise pour assurer sa pertinence. En effet, la conception spatiale des deux villes diffère fondamentalement. Alors que Paris s'articule autour d'oppositions binaires comme Paris *vs* banlieue ou rive gauche *vs* rive droite, et se divise en arrondissements, ce découpage n'a pas la moindre pertinence à Bruxelles où s'imposent d'autres concepts comme le quartier (ex. les Marolles, quartier populaire du centre-ville par rapport au Sablon, quartier huppé voisin du premier), la commune (Schaerbeek, commune à forte immigration *vs* Watermael-Boitsfort,

commune résidentielle), la région (vu le statut intermédiaire de Bruxelles, devenue région neuf ans après la Wallonie et la Flandre) ou la communauté linguistique.

Si le contact des langues se produit dans tout centre urbain, les enjeux linguistiques sont plus apparents à Bruxelles étant donné son bilinguisme . Nous avons donc ajouté des questions intéressantes pour la situation étudiée : l'emploi des langues, notamment le néerlandais, le rapport à la Flandre et à la Wallonie, etc. Nous avons aussi ajouté des questions sur la langue (« Connaissez-vous des expressions typiquement bruxelloises ? » ou encore la cuisine régionale.

Questions méthodologiques

Echantillonnage

Le projet ambitionne de recueillir à terme un échantillon représentatif des variables sociolinguistiques (diatopie, diachronie, diastratie, diagénie) présentes à Bruxelles. Pour ce qui est de la variation diatopique, nous souhaitons donc rassembler des données pour les 19 communes bruxelloises, éventuellement raffinées par quartier si nos ressources le permettent. La dimension diachronique recouvre un large éventail de classes d'âge (de 19 à 102 ans actuellement), et un critère important a présidé au choix des locuteurs : être Bruxellois natif ou en tout cas être installé dans la capitale depuis de très nombreuses années. La dimension diastratique est assurée par la sélection d'informateurs d'arrière-plan socio-économique ou éducatif varié (du concierge de l'université à l'avocate en passant par le chanteur populaire typiquement bruxellois). Finalement, on veut assurer une représentation équilibrée des genres.

Dimension éthique

Comme toute recherche impliquant des êtres humains, la constitution du CFPB implique une dimension éthique. Le protocole, approuvé par les comités d'éthique des universités d'Aston et de Saint-Louis – Bruxelles, assure aux participants l'anonymat s'ils le désirent, et la liberté d'interrompre l'entrevue à tout moment. Avant toute entrevue, l'enquêteur présente les grandes lignes du projet et sollicite l'accord écrit de l'informateur de participer à la collecte linguistique et de permettre la diffusion en ligne de l'entretien éventuellement anonymisé. L'informateur remplit également une fiche signalétique reprenant ses coordonnées personnelles et ses antécédents scolaires et professionnels. Ces données permettront d'interroger les données en fonction de variables indépendantes telles que le sexe ou la tranche d'âge.

Identification des entrevues et des participants

Chaque entrevue est identifiée selon un code composé des éléments suivants, séparés par des tirets :

- les lettres CFPB pour *Corpus de français parlé à Bruxelles* ;
- le code postal du domicile de l'informateur, par ex. 1070 pour Anderlecht ;
- un chiffre, par exemple 3, qui indique qu'il s'agit de la troisième entrevue réalisée dans cette commune. Dans le cas où l'entrevue a été réalisée en plusieurs parties, ce numéro est suivi d'une lettre a, b, c... indiquant la chronologie des conversations.

Chaque participant est identifié par :
- le code de l'entrevue tel que décrit ci-dessus ;
- ses initiales ;
- la lettre *–i* s'il s'agit de l'intervieweur, éventuellement suivie de la lettre *-m* pour *multiple*, dans le cas où celui-ci conduit plusieurs entrevues. Ce codage permet de séparer les informateurs des enquêteurs (pas nécessairement bruxellois), mais également de repérer ceux qui animent plus d'une conversation.

En outre, chaque informateur se voit attribuer un pseudonyme approprié à son âge et reflétant les origines de son nom. Par exemple, un René Lefèvre âgé d'une soixantaine d'années recevra un prénom adapté à sa génération (Robert, Roger, Raymond...) et un patronyme à consonance française comme Ledaim tandis qu'une Océane Vandenbosche âgée de 19 ans deviendra Oriane plutôt qu'Odette et son patronyme aura des consonances flamandes comme Vancauwenberg. Ce double système d'identification allie l'efficacité d'un code rigoureux pour le classement mais peu mémorable pour les utilisateurs, avec un système mnémotechnique plus intuitif qui offre un profil de chaque informateur.

Le profil des enquêteurs

Le recueil des données s'est fait notamment grâce au concours des étudiants en langues et littératures françaises et romanes de l'Université Saint-Louis – Bruxelles dans le cadre du séminaire de linguistique synchronique du français. C'est la raison pour laquelle le corpus comprend un grand nombre d'intervieweurs différents, tous relativement jeunes.

Le guide d'entretien sous-tend la collecte des données, et la trame en est plus ou moins suivie dans toutes les entrevues. Mais certains intervieweurs, selon leur aisance, leur personnalité ou celle de leur interlocuteur s'en éloignent plus moins pour introduire d'autres sujets de conversation. Certains sujets sont plus personnels, notamment lorsque les interlocuteurs se connaissent bien et partagent une histoire commune. C'est notamment le cas de certaines entrevues réalisées par les étudiants qui ont choisi des locuteurs qu'ils connaissent bien. Dans le corpus, on a en fait deux styles d'enquêteurs : certains sont proches du guide d'entretien, alors que d'autres se sentent plus libres et s'en éloignent davantage.

Conventions de transcription

Format des fichiers de transcription orthographique

La transcription orthographique se fait directement sous le logiciel Praat (http://www.fon.hum.uva.nl/praat/) en alignant le texte au son. Les intervalles n'ont aucune valeur théorique. On veille à ce qu'ils ne soient pas trop longs (pas plus de 3 ou 4 secondes) afin de faciliter les traitements ultérieurs. Chaque locuteur possède sa titre propre. Les chevauchements de parole apparaissent dans des intervalles séparés, qui contiennent donc uniquement la parole chevauchée.

Principes

La transcription orthographique des données du corpus *CFPB* suit les quatre grands principes suivants:

1. **Adoption de l'orthographe standard.** On n'utilisera aucun des « trucages orthographiques » qui visent à calquer la prononciation. On écrira *parce que* et non **pasque, ils vont* et non **i vont, ils ont* et non **i-z-ont, je sais pas* et non *chais pas,* etc. De la même manière, on ne notera pas les élisions qui ne sont pas marquées dans le français écrit standard. On transcrira donc *tu deviens* et non *tu d'viens, peut-être* et non **p'têt'*, *gaufre* et non *gauf"*etc. Le principe général de transcription dans *CFPB* est de matérialiser la présence ou l'absence des morphèmes, mais non de s'intéresser à la forme particulière qu'ils peuvent prendre. C'est donc la forme la plus conventionnelle des morphèmes qui est transcrite, et les morphèmes non prononcés ne sont pas transcrits. (cf. le *ne* de négation). On n'élidera donc pas les clitiques, même si l'on rencontre de plus en plus souvent ces graphies dans les écrits non formels, moins surveillés, ainsi que dans la bande dessinée : des deux premières personnes dans les deux cas suivants :

- *je* même devant une consonne : **j'veux* → *je veux*
- *tu* : **t'arrives, *t'veux* → *tu arrives, tu veux*

2. **Absence de ponctuation.** Les transcriptions ne sont pas ponctuées, et cela afin de ne pas orienter les analyses syntaxiques ultérieures. Le seul signe de ponctuation utilisé est le point d'interrogation. Celui-ci sera noté pour indiquer que le locuteur pose une question, que cette question soit marquée par la syntaxe (inversion, tournure en *est-ce-que*) ou par une intonation montante.

Les prises de parole des locuteurs ne commencent pas par une majuscule, le concept de phrase étant abandonné. La majuscule est utilisée uniquement pour les sigles (dans lesquels on ne note pas de point), les acronymes et les noms propres (cf. ci-dessous).

3. **Les pauses silencieuses.** Les pauses silencieuses ne sont pas notées. On laisse des intervalles vides.

4. **Prise en compte de l'oralité des données**, avec notation minutieuse des phénomènes suivants :

- la pause dite pleine : *euh*
- les répétitions de mots (p. ex. : *le le papier*)
- les auto-corrections (p. ex. : *le la carte*)
- les amorces de morphèmes sont notées avec un trait d'union (p. ex. : *on s'ad- on s'adapte*)
- les interjections et onomatopées (cf. liste non exhaustive ci-dessous)
- les chevauchements de parole

Liste des principales conventions

Abréviations graphiques

Aucune abréviation graphique — procédé usuel à l'écrit — n'est employée quand un mot est prononcé dans son intégralité. Tous les termes sont transcrits en entier. On écrira *etcaetera* et non *etc.*

Accents graphiques

Tous les mots transcrits sont accentués, y compris sur les lettres majuscules.

Accords non standards

Les accords non standards audibles ne sont pas corrigés : ils sont transcrits à l'aide des morphèmes usuellement utilisés. On écrira donc *des chevals, je me suis permise, les histoires qu'elle a dits, etc.* Dans le cas des accords du participe passé qui ne sont pas audibles, on considère que l'accord est correctement fait : ainsi, on écrira *les histoires qu'elle a racontées.* Plus généralement, tout ce qui peut apparaitre comme une « faute » de grammaire est transcrit sans correction. On n'ajoutera aucun *(sic)* dans ce cas.

Alternance de code

L'alternance de code concerne les passages longs, et non les emprunts. On ne considèrera pas comme une alternance de code *c'est un has been*. De même, les passages très courts ne seront pas marqués comme des alternances de code. Cf. cet extrait célèbre de la pièce de théâtre *Bossemans et Coppenolle*: *je ne m'appelle pas Madame Chapeau ça est les crapuleux de ma strotje qui m'ont donné ce surnom parce que je suis trop distinguée pour sortir en cheveux*

Apocope et aphérèse

On considère les apocopes et les aphérèses comme des lexèmes de la langue. Ils ne reçoivent aucune marque particulière.

Chevauchement de parole

En cas de chevauchement de parole, on ne coupe pas les mots. Si seulement une partie d'un mot est prononcée dans le chevauchement, on indiquera la totalité de ce mot comme chevauché. On note les chevauchements dans des intervalles bien séparés.

Chiffres et nombres

Les chiffres et nombres sont transcrits en toutes lettres, avec des traits d'union du début à la fin. Cette dernière convention est adoptée afin d'éviter les erreurs inévitables dans l'utilisation du trait d'union. Noter les nombres en lettre permet de distinguer la prononciation de 70 : *septante* vs *soixante-dix*. On écrira : *mon numéro de téléphone est le zéro zéro trente-deux quatre-cent-septante-quatre cinq-cent-vingt-deux cinq-cent-trente*

Il y a

La tournure *il y a* peut être prononcée de 4 manières différentes: [ilia], [ilja], [ija] ou [ja]. On a en outre un continuum entre ces différentes prononciations qui, souvent, ne peuvent aisément être distinguées. La règle est de transcrire systématiquement *il y a*.

La liaison

On ne note pas le phénomène de la liaison. Seules les liaisons erronées seront notées avec la lettre prononcée entre tirets : *huit-z-euros, des chemins de fer-z-américains, donne-moi-z-en, je souhaite que cette portion du procès-verbal puisse-t-être communiquée rapidement, etc.*

Multitranscriptions

Certaines conventions proposent des multitranscriptions en cas de doute, qu'il s'agisse d'un problème d'écoute ou d'un choix grammatical, au niveau des accords par exemple (cf. : *ses frères et sœurs* vs *ses frère et sœurs* vs *ses frères et sœur* vs *ses frère et sœur*). Dans CFPB, le transcripteur veillera à choisir la solution qui lui parait la plus plausible selon le contexte, sans multiplier les multitranscriptions. En cas d'impossibilité de trancher, il mettre les deux choix entre parenthèses, séparés par une virgule : (choix1 , choix 2).

Le* ne *de négation

L'adverbe de négation *ne* n'est pas audible lorsqu'il est précédé de *on* et suivi d'une voyelle ou d'un *h* muet. Par exemple, quand on entend [o~napAlta~], on ne peut pas dire si le locuteur a construit l'énoncé *on n'a pas le temps* ou bien *on a pas le temps*. Dans ce cas, ce possible *n'* de négation est toujours transcrit entre parenthèses : (n').

Particules de l'oral

Les « particules de l'oral » (interjections, onomatopées, etc.) sont transcrites de manière normalisée. Le tableau ci-dessous recense les graphies de quelques particules courantes. On se reportera au *Dictionnaire des onomatopées* (Enckell et Rézeau, 2005) pour une liste plus complète.

ah aïe bah bé	ben eh euh hein	Hum mais enfin mh moui mouais	oh oh la la ouille pf

Tableau 1: Liste des transcriptions des particules de l'oral les plus fréquentes

Les variantes mineures d'onomatopées ne sont pas distinguées. Par exemple, c'est toujours *pf* qui est transcrit, jamais *pff*, *pfff* , ou *pffff*. La particule *mh* se distingue de *hum* de la manière suivante : *mh* correspond à un acquiescement du locuteur (parfois répété : *mh mh*), *hum* est utilisé dans les autres cas.

Phénomènes phonétiques et prosodiques

Les phénomènes phonétiques et prosodiques (prononciations particulières, élisions, allongements vocaliques, liaisons, reprises de souffle, pauses, intonation) ne sont pas transcrits, de même que les prononciations non standards. On peut éventuellement dans ce dernier cas, si cela est pertinent, recourir à la transcription phonétique à côté du lexème. Seules les liaisons erronées seront marquées (cf. *liaisons*). Le tableau ci-dessous synthétise les conventions utilisées :

Phénomène	Marque	Exemple
Acronymes	1re lettre en majuscule	Il travaille à la Nasa
Alternance de code[1]	$< >$	ouais je sais je sais $< hij was 22 jaar oud toen hij dit neer schreef >$
Amorces de morphèmes	~	une piste d'atterrissage d'hélicop~ pour héli~ hélicoptère
Apocopes et aphérèses	aucune marque	Steph de Monac, ricain
Nombres	écrits en toutes lettres, trait d'union du début à la fin	septante-deux, soixante-douze
Noms propres et acronymes	1re lettre en majuscule	George Clooney
Passages inaudibles	xxx	
Sigles	tout en majuscules, sans espace	maintenant il pointe à l'ANPE
Titres (livres, films, etc.)	^ ^	il a regardé ^Autant en emporte le vent^

Tableau 2 : Conventions de représentations des phénomènes phonétiques et prosodiques

Etat d'avancement du corpus (fin juillet 2015)

Le tableau ci-dessous répertorie le corpus tel qu'il se présente à la fin juillet 2015 :

Code	Informateu	Date	nbre	âge	lieu	durée

	r		locuteurs			
CFPB-1000-1	3 F, 1 M	15/04/2014	5	84, 72, 43, 74	Bruxelles Marolles	47 min 56s
CFPB-1000-2	M	15/04/2014	2	85	Bruxelles, Marolles	31 min 11s
CFPB-1000-3	3 F	15/04/2014	4	86, 84, 82	Bruxelles, Marolles	1h 31min 30s
CFPB-1000-4	F	15/04/2014	2	84	Bruxelles, Marolles	22 min 23s
CFPB-1000-5	M	01/05/2015	2	57	Bruxelles, centre	1h 14min 19s
CFPB-1030-1	M	01/05/2015	2	54	Bruxelles, Schaerbeek	1h 4 min 10s
CFPB-1050-1	F	07/01/2014	2	41	Bruxelles, Ixelles	54 min 57s
CFPB-1070-1	M	09/03/2014	2	50	Bruxelles, Anderlecht	1h 6 min 25s
CFPB-1070-2 a,b	F	22,26/03/2014	2	50	Bruxelles, Anderlecht	50 min 21s+ 29min 42s
CFPB-1070-3	F	01/05/2015	2	53	Bruxelles, Anderlecht	1h 9min 20s
CFPB-1082-1	F	26/03/2013	2	56	Bruxelles, Berchem-Sainte-Agathe	1h 9min 55s
CFPB-1082-2	M	01/05/215	2	22	Bruxelles, Berchem-Sainte-Agathe	52min53
CFPB-1083-1	M	25/03/2014	2	34	Bruxelles, Ganshoren	1h15
CFPB-1083-2	F	24/03/2014	2	19	Bruxelles, Ganshoren	1h10
CFPB-1083-3 a, b	M, F	01/05/2015	2/3	66, 51	Bruxelles, Ganshoren	20’39
CFPB-1090-1	F	14/03/2014	2	19	Bruxelles, Jette	1h 39
CFPB-1090-2	M	01/05/2015	2	51	Bruxelles, Jette	1h 7min 2 s
CFPB-1090-3	M, F	8/04/13	3	71, 68	Bruxelles, Jette	1h47
CFPB-1150-1	M	19/04/2013	2	102	Bruxelles, Woluwé-Saint-Pierre	57 min
CFPB-1150-2	F, M	17/04/2013	3	47, 49	Bruxelles, Woluwé-Saint-Pierre	1h04
CFPB-1160-1	M, F	01/05/2015	3	40, 41	Bruxelles, Auderghem	55 min 39s
CFPB-1180-1	M, F	16/03/2014	3	58, 55	Bruxelles, Uccle	1h 18min 36s
CFPB-1180-2	M	19/03/2013	2	60	Bruxelles, Uccle	59 min
CFPB-1190-1	F	21/03/2013	2	70	Bruxelles, Forest	1h08
CFPB-1190-2	F	09/04/2013	2	78	Bruxelles, Forest	1h18
CFPB-1190-3	M	08/04/2013	2	51	Bruxelles, Forest	46min 52s
CFPB-1200-1	F	01/05/2015	2	19	Bruxelles, Woluwé-St-Lambert	53min59s
CFPB1200-2	M	07/01/2014	2	39	Bruxelles, Woluwé-St-Lambert	53min 32s
CFPB-1702	M	16/02/2013	2	76	Bruxelles, Grand-Bigard	73 min

Tableau 3 : Liste des entrevues présentes dans le corpus en juillet 2015

A l'heure actuelle, le corpus comprend donc :

- 30 entrevues sociolinguistiques (réparties en 32 enregistrements[2]) impliquant de 1 à 4 informateurs (26 dialogues, 4 trilogues, 1 quadrilogue, 1 quinqualogue) ;
    - 5 totalement transcrites selon l'orthographe standard, avec la transcription alignée au son (sous Praat) ;
    - 25 partiellement transcrites.

- 21 intervieweurs différents et 39 témoins (18 hommes et 21 femmes) qui représentent les groupes d'âge suivants:

Tranche d'âge	M	F
< 20		3
20-29	2	
30-39	2	
40-49	2	4
50-59	6	5
60-69	2	2
70-79	2	2
>80	2	5

Tableau 4: Distribution des informateurs par âge et par sexe

- Des entrevues représentatives de 12 communes bruxelloises :

	Communes	**Code postal**	**Nb interviews**
1	Anderlecht	1070	3
2	Auderghem	1160	1
3	Berchem-Sainte-Agathe	1082	2
4	Bruxelles-Ville*	1000, 1020, 1040, 1050, 1120, 1130	5
5	Etterbeek	1040	

[2] Quatre enregistrements peuvent en fait se regrouper deux par deux. En effet, il s'agit de deux entrevues réalisées à des moments distincts avec les mêmes locuteurs. Une des entrevues est réalisée au domicile de la locutrice, la suite se poursuivant un autre jour à son cabinet d'avocate (CFPB-1070-2a et 2b). Une autre entrevue a été réalisée à deux reprises à cause de problème techniques : le locuteur, interviewé par sa petite-nièce, a demandé la présence de la mère de celle-ci, ne l'ayant plus vue depuis longtemps. On a donc ici un dialogue puis un trilogue pour la seconde partie de l'entretien (CFPB-1083-3a et b).

	Communes	Code postal	Nb interviews
6	Evere	1140	
7	Forest	1190	3
8	Ganshoren	1083	4
9	Ixelles	1050	1
10	Jette	1090	3
11	Koekelberg	1081	
12	Molenbeek-Saint-Jean	1080	
13	Saint-Gilles	1060	
14	Saint-Josse-ten-Noode	1210	
15	Schaerbeek	1030	1
16	Uccle	1180	2
17	Watermael-Boitsfort	1170	
18	Woluwe-Saint-Lambert	1200	2
19	Woluwe-Saint-Pierre	1150	2

Tableau 4 : Distribution des entrevues par commune

Développements futurs

D'ici le printemps 2016, nous comptons :

- transcrire la totalité des données
- intégrer celles-ci dans l'outil d'interrogation développé pour le CFPP.

Figure 2 : Carte de Bruxelles -19 communes

Notons que Grand-Bigard n'est pas une commune bruxelloise mais qu'elle est située à l'ouest de Bruxelles, dans le brabant flamand. Si cette entrevue figure dans notre corpus, c'est que le locuteur est emblématique du parler bruxellois puisqu'il s'agit d'un chanteur populaire belge, identifié comme typiquement bruxellois. Il a d'ailleurs vécu longtemps dans le quartier populaire des Marolles.