

OFROM – corpus Oral de Français de Suisse ROMande : archiver la parole, étudier les usages, enseigner le français

Mathieu Avanzi^{1,*}, Gilles Corminboeuf², François Delafontaine¹, Laure Anne Johnsen¹, Matthieu Monney² et Maguelone Sauzet¹

1. Université de Neuchâtel.
2. Université de Fribourg.

*mathieu.avanzi@unine.ch

Résumé. Le corpus Ofrom est une archive constituée de transcriptions d'enregistrements réalisés auprès de locuteurs bien ancrés en Suisse romande. Créée par une équipe de linguistes soucieux de disposer de données empiriques pour décrire les pratiques linguistiques de cette région de la francophonie, la base contient plus de dix ans après son ouverture au grand public près de 1,3 millions de mots. Cet article expose les fondements méthodologiques du corpus et ses outils, puis présente quelques exploitations possibles pour l'étude scientifique du français et pour son enseignement.

1 Introduction

L'élaboration et la mise à la disposition de corpus informatisés pour la communauté scientifique constituent un enjeu majeur pour la description des usages dans les langues naturelles. Depuis plusieurs décennies, les initiatives se sont multipliées dans l'espace francophone pour documenter le français réellement pratiqué par les usagers (voir les bilans d'étape de Bruxelles *et al.* 2009, Avanzi *et al.* 2016a et Benzitoun & Cappeau à paraître). Côté helvétique, ce n'est qu'avec le projet PFC, lancé dans les années 2000 (Durand *et al.* 2002), que des enregistrements de français de Suisse romande ont enfin été réalisés et inclus dans une base ouverte au grand public. Avec la récolte initiée en 2007, puis l'ouverture du site au grand public quelques années plus tard, la base de données orales OFROM (<https://ofrom.unine.ch/>) a non seulement permis de disposer du premier corpus de français oral de Suisse romande, mais aussi de rattraper, grâce aux 1,3 millions de mots qu'elle contient actuellement, le retard pris par rapport à d'autres pays francophones¹. Jusqu'à présent, ce sont essentiellement les études dont OFROM constitue la base empirique qui ont fait connaître la ressource (*inter alia* Corminboeuf *et al.* 2020). Hormis par l'intermédiaire des contributions d'Avanzi *et al.* (2016b) et de Johnsen (2020, 2022) pour le volet Enseignement, la ressource OFROM n'a que très rarement été présentée de manière circonstanciée. C'est à une telle présentation que nous nous employons dans les pages suivantes. Dans un premier temps, l'accent sera mis sur les outils et les principes méthodologiques qui ont permis d'aboutir à la configuration actuelle du corpus (§2). Dans un second temps, nous illustrerons quelques possibilités d'exploitation du corpus, que cela concerne la description du français (§3) ou son apprentissage (§4).

2 Le corpus OFROM : principes méthodologiques, outils et composition

2.1 Principes méthodologiques

À l'origine, OFROM dans sa version « classique » se compose de deux types d'enregistrements principaux : d'une part, des extraits d'entretiens dirigés à prédominance monologique où l'interviewé, un témoin bien ancré dans la localité de Suisse romande où il est enregistré, répond à des questions posées par l'enquêteur ; d'autre part, des interactions impliquant généralement deux témoins suisses romands qui conversent de manière informelle. Les sujets couverts dans ces enregistrements varient, touchant des aspects tels que la langue, les métiers, les voyages, les loisirs, les relations de voisinage, les projets personnels ou des événements particuliers vécus par les locuteurs. Ils peuvent également aborder des thèmes politiques, la situation linguistique de la Suisse (Diémoz *et al.* 2020), voire les coutumes locales de la région où résident

les locuteurs. Depuis 2022, le sous-corpus « multigenres » a été intégré à la base, dans un but de diversification des types de discours, notamment dans une perspective d'enseignement.

Dans le corpus « classique », les fichiers audios proviennent essentiellement d'enquêtes de terrain réalisées par des étudiants qui enregistrent, dans une grande majorité des cas, des personnes de leur entourage. Les enregistrements destinés à intégrer l'interface sont sélectionnés en fonction de leur intérêt linguistique, de l'état du corpus (critères de diversification, d'équilibre, de représentativité) et de leur qualité sonore.

Dans le sous-corpus « OFROM-multigenres », les enregistrements proviennent de sources diverses : enquêtes réalisées par des étudiants, chercheurs ou enseignants universitaires, extraits provenant de médias ou émanant d'institutions publiques. La sélection des extraits dépend d'objectifs pédagogiques généraux, comme la volonté d'illustrer la diversité en termes de communication (types d'interaction, genres de parole, proximité / distance communicative, etc.) ou spécifiques, en lien direct avec des enseignements universitaires particuliers.

Nous utilisons le principe de double transcription pour traiter les enregistrements (une première transcription est révisée une ou deux fois), et suivons les principes méthodologiques proposés par Blanche-Benveniste & Jeanjean (1987), en utilisant l'orthographe standard du français (sans trucages)². La transcription se fait directement dans le logiciel *Praat* (Boersma & Weenink 2023) qui permet un alignement du texte et du son. La durée de chaque extrait du corpus « classique » est comprise entre 10 et 20 minutes, alors que celle du sous-corpus « multigenres » varie fortement selon le type d'interaction. Ainsi, un message vocal de moins d'une minute peut côtoyer l'enregistrement d'une table ronde de près d'une heure.

Pour chaque locuteur, des informations de nature socio-démographique sont consignées dans des métadonnées (origine géographique, âge, sexe, niveau de formation, activité professionnelle). Depuis peu, le degré de proximité et la nature de la relation entre plusieurs témoins est précisée ; il s'agit là d'une des modifications récentes apportées à la base et qui concerne uniquement les enregistrements intégrés depuis 2021. En plus des métadonnées des témoins, sont inscrites les métadonnées de l'enregistrement, qui fournissent la date et le lieu de l'enquête, les dates de transcription et de révision, le type de discours (v. *infra*, §4.2.2) ainsi que la qualité sonore.

2.2 Interface et outils

Une fois les enregistrements recueillis, transcrits, révisés et anonymisés³, ceux-ci sont mis en ligne. Lors du dépôt, la transcription est augmentée d'une annotation morphosyntaxique à l'aide du logiciel DisMo (Christodoulides *et al.* 2014). Cette annotation se charge de la segmentation en mots (*tokenization*) et de leur annotation en lemmes et en catégorie grammaticale – informations utilisées ensuite pour la recherche automatique (*infra*, §3). Le jeu d'étiquettes de DisMo est conçu spécialement pour correspondre à la production orale.

Le développement et la maintenance du site sont assurés par Pierre Ménétreay. Hébergé sur un serveur de l'Université de Neuchâtel, le site fournit deux moyens d'accès aux données orales : l'accès au moyen d'un moteur de recherche et depuis 2021, l'accès via un atlas sonore interactif qui permet de visualiser les points d'enquête ajustés à la zone, chacun d'entre eux fournissant un extrait audio d'une vingtaine de secondes accompagné de sa transcription et de ses métadonnées. Cette nouvelle visualisation cartographique offre donc un moyen parlant d'illustrer la variation diatopique du français parlé en Suisse romande, pour les spécialistes comme pour les non spécialistes.

Le moteur de recherche, accessible via l'onglet « Recherche dans le corpus », permet de faire des requêtes simples dans la base. Les requêtes sont réalisées à l'intérieur d'un champ de recherche principal, dans l'onglet « Contexte » (voir Figure 1). La séquence recherchée peut être une chaîne de caractères arbitraire, un mot (« token ») donné, une catégorie grammaticale (« part of speech ») ou un lemme :

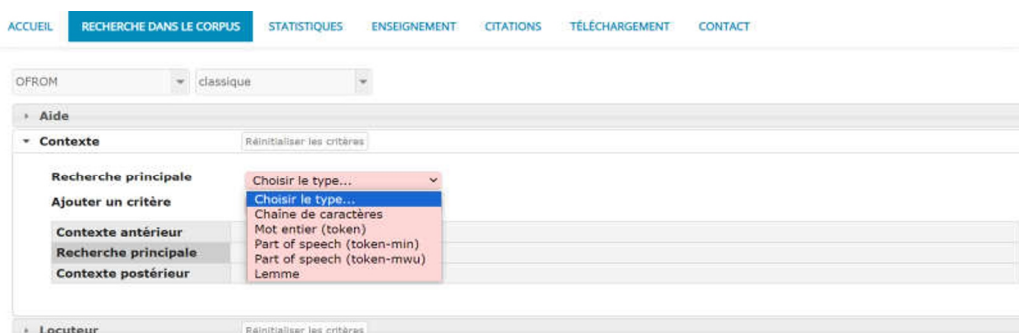


Figure 1. Interface du moteur de recherche.

Une fois la requête réalisée, l'outil fournit une liste de résultats numérotés avec pour chacun l'accès à l'extrait sonore et à une fenêtre de détails affichant les métadonnées. Il est également possible d'élargir à partir de là le contexte à souhait. Si l'extrait élargi contient les interventions de plusieurs interlocuteurs, différentes couleurs distinguent désormais la transcription des segments (Figure 2).

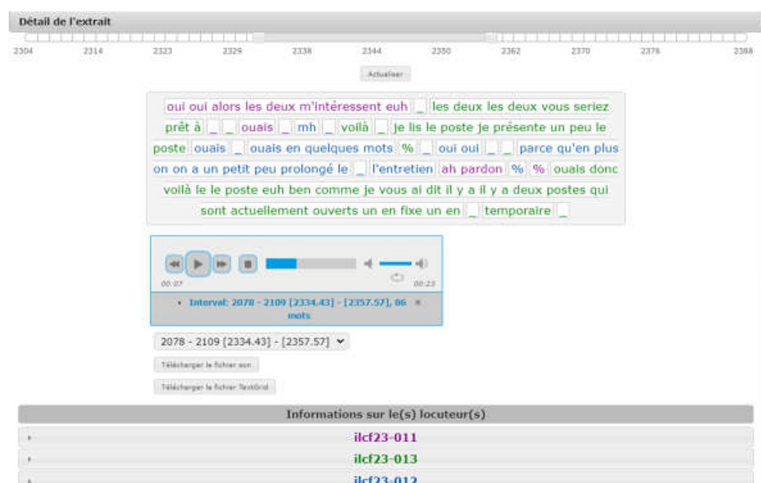


Figure 2. Visualisation d'un extrait élargi à la suite d'une recherche dans le concordancier.

Les résultats de la recherche renvoient aux segments contenant la requête, ce qui explique qu'un segment incluant plusieurs fois la requête ne s'affiche qu'une seule fois. Chaque occurrence comme l'ensemble des résultats peuvent être téléchargés et le système découpe des sons en fonction de l'empan choisi.

Par ailleurs, deux champs correspondant aux cotextes antérieurs et postérieurs permettent de rechercher les unités adjacentes à la séquence recherchée (*infra*, §3.2 et 3.3). La recherche peut être restreinte selon des critères socio-démographiques (filtre « Locuteur ») ou propres à l'enregistrement (filtre « Enregistrement »), par exemple selon le type d'interaction ou le degré de proximité communicative.

2.3 Architecture de la base et composition du corpus

Comme indiqué *supra* (§ 2.1), le corpus OFROM est pour le moment réparti en deux sous-corpus : OFROM-classique qui comprend exclusivement des entretiens et des conversations ; et OFROM-multigenres qui vise à diversifier les types de discours (conférences, entretiens d'embauche, débats, réunions, messages vocaux, contes, enseignements, etc.), dont la liste évolue au fil de l'intégration de nouveaux types d'enregistrement. Pour une recherche comme pour la visualisation de l'atlas, l'utilisateur peut sélectionner un seul des deux corpus, ou travailler sur l'ensemble de la base de données.

Dans son ensemble, le corpus compte actuellement 1,25 millions de mots pour plus de 500 locuteurs et quelque 82 heures de parole. Si le sous-corpus classique demeure actuellement le plus fourni (Tableau 1), le sous-corpus multigenres est appelé à se développer ces prochaines années.

Tableau 1. Taille et sous-corpus d’OFRM⁴.

	Nb. Enregistrements	Nb. Locuteurs	Nb. Mots	Durée (heures)
OFRM-classique	438	415	1.174.410	76
OFRM-multigenres	55	91	91.907	6
Total	493	506	1.266.317	82

Afin de représenter l’ensemble de la population suisse romande, le corpus est échantillonné par canton, et selon l’âge et le sexe des locuteurs. Dans la pratique, et notamment du fait de la méthode d’enquête reposant sur les réseaux estudiantins, certaines tranches d’âge (les étudiants) et certains cantons sont sur-représentés (Figure 3) :

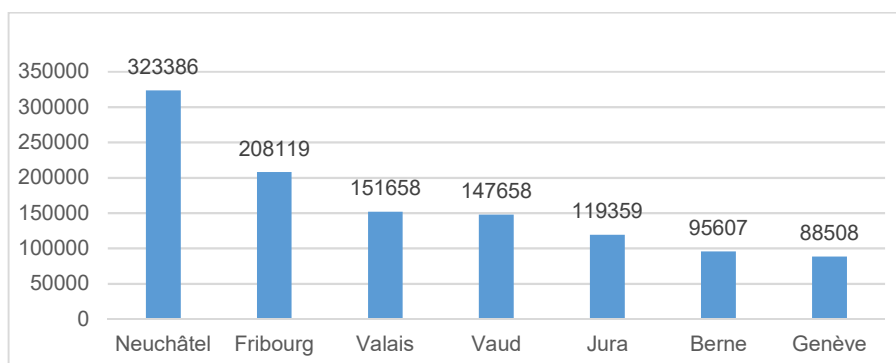


Figure 3. Nombre de mots par cantons (N = 1'266'317)

La population des cantons diffère fortement : en 2022, Vaud comptait par exemple environ 830'000 habitants, Genève environ 514'000, Neuchâtel environ 176'000 et le Jura environ 74'000. En mettant ces chiffres en relation avec ceux de la Figure 3, on peut en conclure qu’actuellement OFROM sur-représente les locuteurs neuchâtelois et jurassiens, mais sous-représente les locuteurs vaudois et genevois. Les tranches d’âge privilégiées des témoins se situent entre 18 et 65 ans. Les locuteurs plus jeunes et plus âgés apparaissent encore sous-représentés. En ce qui concerne le sexe des témoins, 54,73% des témoins sont des femmes et 45,27% sont des hommes.

3 Utilisation de l’interface de recherche d’OFRM

Dans cette section, nous illustrons au moyen d’exemples concrets (*quelque part*, *ça joue* et *autant*) les possibilités d’exploitation du corpus OFROM, et de son interface de recherche.

3.1 Le marqueur (à/en) *quelque part*

Pour une première illustration, notre choix s’est porté sur le marqueur *quelque part*, dont le fonctionnement syntaxico-sémantique pose un certain nombre de problèmes d’analyse.

Le moteur de recherche livre 79 occurrences, qui se répartissent en deux emplois distincts sur plan syntaxique (exemples 1 vs 2) :

- (1) (a) en Valais à chaque fois que tu vas **quelque part** c’est marqué carnotzet (unine15-064)
- (b) je suis trop bien tombée quoi il va y avoir une tuile **quelque part** (unine08a07m)
- (c) on passait vingt minutes à rechercher pourquoi est-ce que tout d’un coup on était plus ensemble euh sur les partitions mais y avait une petite erreur **à quelque part** (unine08a10m)

- (2) (a) on parle tellement de pollution etcetera et puis euh **quelque part** je culpabiliserais de prendre l'avion pour un oui pour un non (unine11d01m)
 (b) pis ils avaient plus les moyens de de se mettre dans le marché avec les nouveaux produits ce qui est **quelque part** triste pour le village pour les emplois (unine08a11m)
 (c) même si moi je m'intéresse à d'autres musiques j'enseigne quand même euh essentiellement euh la technique classique parce que **à quelque part** j'ai été formée dans ça (unine11b01m)

Pour chacun, la forme est parfois précédée de la préposition *à* (exemples 1c et 2c), ce qui différencie OFROM d'autres corpus. La forme *en quelque part*, attestée une dizaine de fois dans le *Corpus de français parlé au Québec* (<https://applis.flsh.usherbrooke.ca/cfpq/>), n'est en revanche pas (encore ?) attestée dans OFROM.

Dans le premier emploi (extraits 1), le marqueur est étroitement lié au verbe, soit en tant qu'argument valenciel sélectionné par celui-ci (1a), soit en tant que constituant adverbial régi par le syntagme verbal (1b), ce que montre le test du clivage et de l'équivalence avec les proformes *où* et *là*⁵. Dans les deux cas, la forme *quelque part* est en « dépendance interprétative » (Kleiber & Gerhard-Krait 2006 : 151) avec un prédicat localisant et s'intègre au contenu propositionnel, d'où le fait qu'elle est en outre placée dans le champ de la négation (dans 1b, *il ne va pas y avoir une tuile quelque part* se comprend au sens de <il n'y aura de tuile nulle part>). On en conclut que *quelque part* est dans cet emploi une forme régie, ou qui appartient à un syntagme lui-même régi (par exemple dans (*à*) *quelque part par là*). Sémantiquement, le marqueur exprime la localisation indéterminée.

Dans le second type d'emploi (extraits 2), que Kleiber & Gerhard-Krait (2006) qualifient de « modal », la forme *quelque part* présente les propriétés syntaxiques inverses. Elle s'affranchit du noyau verbal, ce que montre l'impossibilité de son extraction par le clivage, sa non-équivalence avec les proformes *là* ou *où* et le fait qu'elle ne peut pas se trouver sous la portée de la négation (dans 2c par exemple, la transformation en *à quelque part je n'ai pas été formé dans ça* montre que le marqueur n'est pas affecté par la négation). De plus, elle ne présente pas les restrictions de sélection susmentionnées : omnicompatible, elle peut s'attacher à des prédicats dépourvus de valence locative, dénotant par exemple des processus psychologiques (Blanche-Benveniste 2001 : 90). Dans cet emploi, *quelque part* forme une clause syntaxiquement autonome (Groupe de Fribourg 2012), qui peut être détachée ou intégrée au plan prosodique. Le corollaire sémantique de ces propriétés formelles est l'expression d'une modalisation approximative, avec pour paraphrases possibles *en quelque sorte* ou *d'une certaine façon*.

En vue d'approfondir l'analyse, nous avons caractérisé les 79 occurrences, réparties dans nos deux emplois, selon les critères suivants : (i) la position : antéposée, médiane, finale (par rapport au noyau verbal) ; (ii) la portée : locale, c'est-à-dire affectant spécifiquement un constituant adjacent, ou globale, c'est-à-dire portant sur l'énonciation dans son ensemble ; (iii) la prosodie : détachée de l'énoncé noyau ou intégrée à celui-ci ou à un constituant adjacent, – en considérant que le détachement est marqué par une pause, un décrochage intonatif et/ou un allongement de la syllabe précédant le marqueur. Le Tableau 2 laisse apparaître certaines tendances :

Tableau 2. Représentation chiffrée des différents cas de figure⁶

		Type 1 : Emploi régi (Localisation indéterminée)	Type 2 : Emploi non régi (Modalisation approximative)
Forme	<i>quelque part</i>	28	36
	<i>à quelque part</i>	6	8
Position	Antéposée	2	22
	Médiane	0	10
	Finale	32	12
Portée	Locale	34	7
	Globale		37
Prosodie	Détachée	4	19
	Intégrée	30	25

Dans l'emploi 1, *quelque part* est très majoritairement intégré prosodiquement et en position finale, ce qui est en congruence avec sa description syntaxique. Et lorsqu'il est détaché, il est en épexégèse (3a : noter la pause de plus de deux secondes) ou placé dans une liste (3b) :

- (3) (a) c'est exactement la vision que j'avais probablement en Irlande aussi en plus (2.6) **quelque part** par là (unine18a03d)
 (b) puis bon il nous a emmenés dans le bazar
 dans sa
 au-dessus de son échoppe
quelque part puis on s'est retrouvé [...] (unine09a01m)

Dans l'emploi 2, *quelque part* est aussi bien détaché qu'intégré sans démarcation prosodique, ce qui déjoue d'éventuelles attentes provenant de sa description comme un adverbe d'énonciation. Dans le cas où il est détaché, il occupe préférentiellement les positions antéposée et finale. Et lorsqu'il est intégré, *quelque part* fonctionne comme une clause autonome parenthétique, de portée locale. Ainsi, dans (2b), c'est spécifiquement sur l'attribut *triste* que porte la modalisation appliquée par *quelque part*, plutôt que sur l'énonciation complète. On voit que le déplacement de la forme (extraposée ou intégrée) a des répercussions interprétatives notables. De manière générale, on constate un non-recouplement entre syntaxe et prosodie. Cette dernière n'est en effet pas toujours probante pour discriminer les deux emplois étudiés. Dans certains cas ambigus toutefois, comme en :

- (4) je vois beaucoup de gens qui qui + qui cherchent **quelque part** UN ami un ami (unine08a04m ; les petites capitales sur *un* marquent une proéminence accentuelle)

elle sélectionne univoquement l'interprétation modalisante, alors que le verbe *chercher* serait parfaitement compatible avec un régime locatif. On peut voir sur le tracé intonatif stylisé (Figure 4)⁷ que *quelque part* est décroché intonativement du reste de l'énoncé dans lequel il s'insère : il est entouré de frontières prosodiques majeures (marquées par une proéminence accentuelle forte sur *cherchent*, après une pause silencieuse), et prononcé sur un registre plus bas que le reste de l'énoncé, sans modulation :

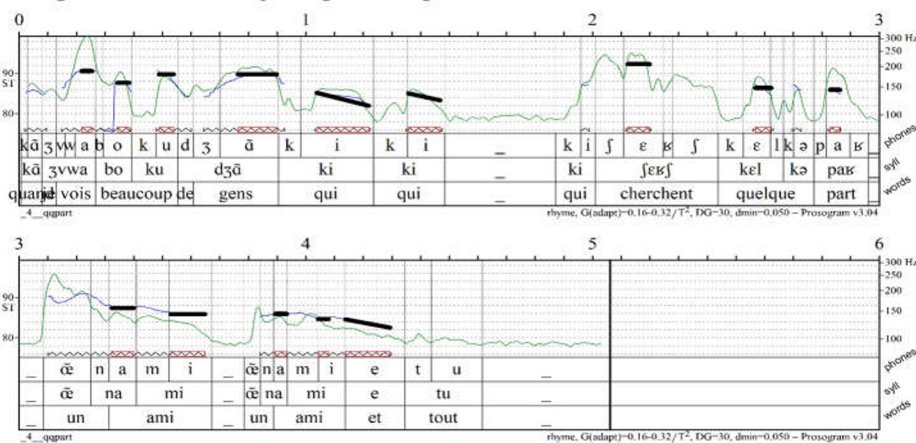


Figure 4. Prosogramme de l'exemple (4)

Les observations consignées ci-dessus permettent surtout d'engager un certain nombre de questions :

- Au plan syntaxique, on peut se demander quelle serait la fréquence, sur un corpus plus vaste, des cas de métanalyse (Béguelin, Corminboeuf & Johnsen 2014), où *quelque part* est analysable à la fois comme un constituant régi et comme une clause autonome. Dans OFROM, une seule occurrence sur 79 nous semble être de ce type :
- (5) enfin oui j'imagine qu'ils ont dû m'influencer quand même euh + **quelque part** mais euh : (unine15z63m)

La prosodie nous semble entériner cette double analyse :

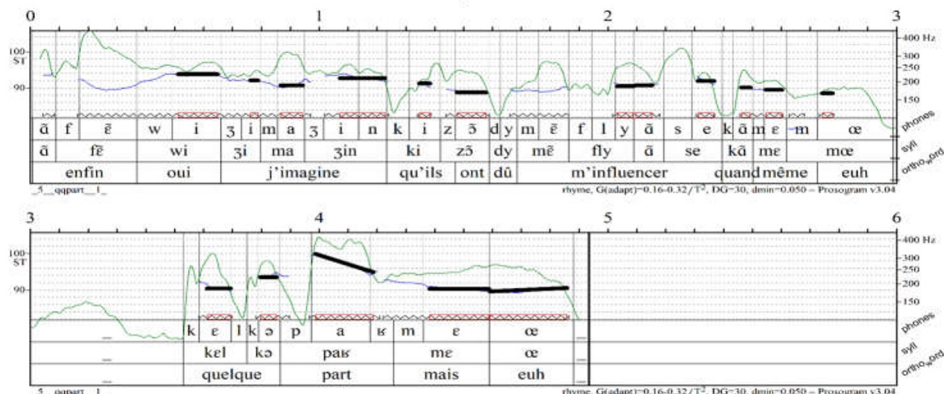


Figure 5. Prosogramme de l'exemple (5)

De telles homonymies de construction pourraient contribuer à expliquer le passage de la localisation indéterminée à l'approximation et donc l'existence de ces deux emplois de *quelque part*.

– Ensuite, il serait intéressant d'étudier plus avant le double fonctionnement du marqueur (2a vs 2b), soit en lui appliquant des analyses distinctes, soit au contraire en cherchant à l'expliquer par une modélisation généralisante (ce que propose Berrendonner 2021 : §3.2.2 pour les adverbes d'énonciation, qui présentent également des emplois intégrés et détachés).

– Au plan sémantique, on peut se demander ce qu'il advient du sème localisant dans l'emploi 2. Le récupère-t-on à un autre niveau de l'analyse linguistique ? Kleiber & Gerhard-Krait estiment que dans l'emploi 2, *quelque part* « localise métaphoriquement la vérité des propositions [affectées par le marqueur] dans un espace de validité », si bien que « la 'localisation' d'une proposition, c'est-à-dire poser qu'elle existe à un endroit, revient à dire qu'elle est vraie à cet endroit » (2006 : 160). Reformulé dans notre perspective, on dira que *quelque part* signifie une modalité qui spécifie dans quelle mesure ou de quel point de vue une énonciation (ou un constituant de celle-ci) est valide (cf. <ceci est vrai *quelque part*>).

– Enfin, une comparaison avec d'autres marqueurs construits sur *quelque* serait instructive, pour déterminer si d'autres pronoms ou lexies indéfinies ont également développé un emploi pragmatique.

3.2 Recherche de l'expression *ça joue*

En Suisse romande, l'expression *ça joue* jouit d'une grande popularité et d'un caractère presque emblématique (elle est reprise par exemple en 2021 comme titre d'une émission télévisée de la *Radio Télévision Suisse*, tirant parti de la polysémie du verbe *jouer*). L'emploi intransitif de *jouer* au sens figuré, en particulier dans une tournure impersonnelle avec le pronom démonstratif neutre *ça*, serait toutefois une création assez récente (DSR 2004 : 478). Lorsqu'elle prend les sens de « fonctionner correctement », « tomber juste » (en parlant de comptes ou d'opérations chiffrées) ou bien encore « marcher, aller, convenir », d'après la définition du DSR, cette construction dispose d'un large spectre d'applications possibles dans les interactions orales.

Afin d'explorer la vitalité de cette expression dans le corpus OFROM, il convient de bien préparer la requête à réaliser dans le moteur de recherche. En effet, pour un verbe aussi fréquent et polysémique, le risque est grand d'obtenir beaucoup de bruit ou alors, au contraire, d'exclure des variantes intéressantes de cette construction avec une recherche trop restrictive. Pour rechercher la construction *ça joue* et ses différentes réalisations possibles dans OFROM, il faudra donc commencer par inscrire comme objet principal dans le volet « Contexte » du formulaire de recherche soit l'étiquette « lemme » en saisissant directement l'infinitif *jouer* (Figure 6), soit dans la liste des parties du discours (« part of speech ») l'étiquette « verbe » en précisant *jouer*.

The screenshot shows a search engine interface with the following elements:

- Top left: 'OFRON' dropdown and 'Tous les sous corpus' dropdown.
- Section 'Aide' with a sub-section 'Contexte' and a 'Réinitialiser les critères' button.
- 'Recherche principale': A dropdown menu set to 'Lemme' and a text input field containing 'jouer'.
- 'Ajouter un critère': A dropdown menu set to 'Choisir le type...'.
- 'Contexte antérieur': A section with 'Recherche principale' set to 'Lemme' and a text input field containing 'jouer'.
- 'Contexte postérieur': An empty text input field.

Figure 6. Objet principal de la recherche *ça joue*

Le formulaire du moteur de recherche est conçu de manière à permettre d'affiner les critères de recherche, en ajoutant à un critère de « Recherche principale » deux champs correspondant aux cotextes antérieurs et postérieurs qui permettent de rechercher les unités adjacentes à la séquence ciblée. Ces cotextes peuvent être formés de plusieurs séquences qui sont chacune conditionnées par un opérateur : elles peuvent être nécessaires (ET), en alternative (OU) ou interdites (SAUF), et cela pour une position donnée (en nombre de mots) ou dans un intervalle de temps (en nombre de secondes). Au paradigme entier de *jouer*, il faut donc ajouter un critère supplémentaire, dans le champ dédié : on choisit sous l'étiquette « mot entier » la forme *ça*, avec l'opérateur booléen « ET » pour croiser les deux ensembles recherchés. Il faut ensuite préciser que ce mot doit précéder le verbe *jouer*, dans un empan d'une seconde.

The screenshot shows the search engine interface with the following elements:

- Top left: 'OFRON' dropdown and 'Tous les sous corpus' dropdown.
- Section 'Aide' with a sub-section 'Contexte' and a 'Réinitialiser les critères' button.
- 'Recherche principale': A dropdown menu set to 'Lemme' and a text input field containing 'jouer'.
- 'Ajouter un critère': A dropdown menu set to 'Mot entier (token)'.
- 'Texte': A text input field containing 'ça'.
- 'Opérateur': A dropdown menu set to 'ET'.
- 'Dans l'intervalle de': A text input field containing '1' and a dropdown menu set to 'Seconde(s)'.
- 'Contexte temporel': A dropdown menu set to 'Antérieur'.

Figure 7. Ajout d'un critère à la recherche principale

Le bouton « Ajouter » permet de valider la prise en compte de ce critère et la requête prend forme dans un encadré situé juste sous ce bouton. Dans le même formulaire, pour limiter encore le bruit dans les résultats, il est possible d'ajouter des critères afin d'exclure, à l'aide de l'opérateur booléen « SAUF », certaines constructions qui ne relèvent pas de notre étude :

- les emplois pronominaux du verbe : en retirant le pronom *se* lorsqu'il s'agit du mot qui précède immédiatement le verbe,
- les emplois personnels du verbe : en enlevant les pronoms personnels en fonction sujet et les substantifs situés juste avant le verbe,
- les constructions infinitives et les gérondifs : en enlevant les prépositions (*en, de, à*) situées juste avant le verbe qui les introduit,
- les emplois du verbe au sens propre : en enlevant les prépositions *à* et *avec*, juste après le verbe *jouer*.

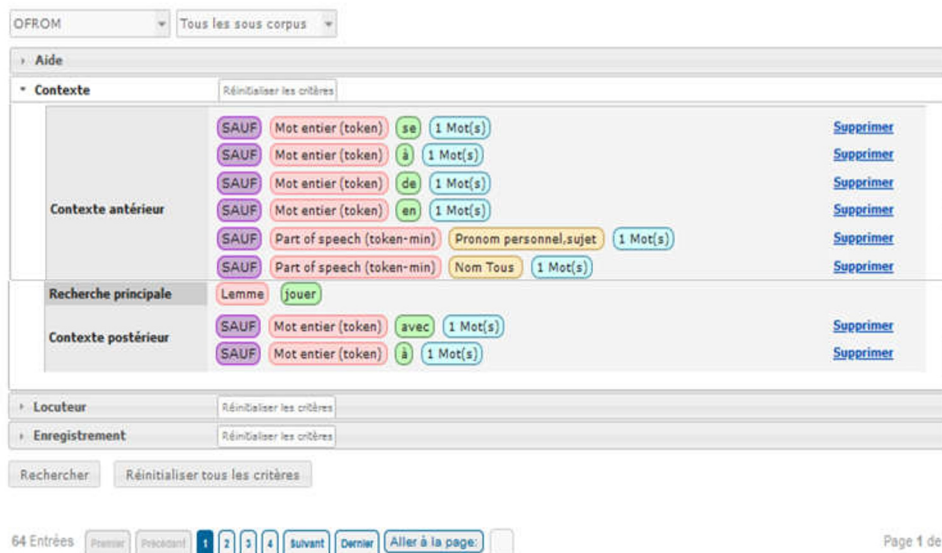


Figure 8. Requête complète pour la construction *ça joue*

Les résultats obtenus sont au nombre de 64 dans l'état actuel du corpus. Les segments qui contiennent l'expression recherchée s'affichent sous la forme d'une liste numérotée, avec un paginateur. Si l'on examine de plus près la pertinence des résultats obtenus, on relève six occurrences du verbe qui ne relèvent pas de l'objet de la recherche. La recherche dans le corpus nous permet toutefois d'observer que la construction impersonnelle et intransitive *ça joue* se décline dans des contextes variés : des interrogatives (a), des propositions subordonnées conditionnelles (b), avec une négation (c) ou une modalisation, introduite par *devoir* ou *pouvoir* au conditionnel (d), ou par le semi-auxiliaire *vouloir* (e) :

- (6) (a) est-ce que **ça joue** pour vous si on refait la même chose (unifr19-030)
 (b) si **ça joue** pour ses parents (unifr19-036)
 (c) c'est ça le problème **ça va pas jouer** (unine16-020)
 (d) mais **ça** devrait **jouer** (unine15-046)
 (e) si je le fais c'est que **ça** veut **jouer** (unine15-934)

3.3 Recherche de la distribution des comparatifs d'égalité *autant* et *aussi*

Dans les grammaires du français contemporain, on considère que l'emploi des adverbes quantificateurs *aussi* et *autant* dans une structure comparative d'égalité à deux termes observe une distribution selon la catégorie morphosyntaxique du terme sur lequel ils portent. Si *aussi* renvoie au degré d'une propriété et se retrouve devant des adjectifs et des adverbes (*aussi jolie* et *aussi vite*), *autant* se rapporte à la quantité ou à un procès et est employé devant des noms (souvent précédés de *de* : *autant de pommes*) ou des verbes (*travailler autant* ; voir Fuchs 2009, à qui nous reprenons ces exemples). De manière convergente, la *Grande Grammaire du Français* (Abeillé et al. 2021 : 1650-1651) considère que « *autant* n'est pas possible avec un adjectif gradable (**Cette affaire est autant importante que la dernière*) », ni avec un adverbe (**Paul va autant souvent qu'avant à la piscine*) ; nous verrons qu'OFROM donne une tout autre image de la distribution de cette forme (*infra*, exemples 7 et 8)⁸.

Le formulaire de recherche d'OFROM permet de faire apparaître la distribution de ces deux comparatifs d'égalité selon la catégorie qui les suit immédiatement. Dans cette requête, nous nous limitons aux comparatives où *aussi* et *autant* s'appliquent à des adjectifs. Pour ce faire, l'objet de la recherche principale, l'adverbe *aussi* ou *autant*, est saisi dans le volet « Contexte », sous la forme de « mot entier ». Ensuite, un critère secondaire s'y ajoute avec l'opérateur booléen « ET » qui permettra de chercher l'intersection des deux ensembles. En l'occurrence, il s'agit de l'une des parties du discours (« part of speech ») de la

catégorie choisie, adjectif par exemple, en précisant qu'elle doit se trouver immédiatement après le comparatif d'égalité, soit « dans l'intervalle de 1 mot(s) ». Le bouton « Ajouter » permet de valider la prise en compte de ce critère et la requête prend forme dans un encadré situé juste sous ce bouton.

Figure 9. Requête à propos de *autant* suivi d'un adjectif

Le formulaire ne permet toutefois pas de cumuler deux objets de recherche principale : il n'est donc pas possible de rechercher en même temps les contextes d'apparition des adverbes *aussi* et *autant*. Cela dit, la simple substitution de l'objet de « Recherche principale » permet de conserver tous les autres paramètres et de relancer rapidement une nouvelle recherche. De même, les boutons « Supprimer » dans l'encadré récapitulatif de la recherche et « Ajouter » permettent d'ajuster les spécifications d'une recherche, sans devoir la recommencer entièrement.

Pour éviter de récupérer parmi les résultats les formes *d'autant* et *pour autant*, il est possible d'exclure de la requête les cas où *autant* est immédiatement précédé de *d'* et de *pour*.

Dans l'ensemble du corpus, on retrouve 14 occurrences d'*autant* + adjectif, sur les 202 occurrences de *autant* dans le corpus. En comparaison, on obtient 121 résultats d'*aussi* + adjectif (le corpus contient 4099 occurrences de cet item très polysémique). Pour la recherche concernant *autant* + adjectif, deux autres occurrences doivent cependant être exclues de la liste de résultats. Elles sont dues à un étiquetage erroné de la partie du discours : *j'avais jamais autant réfléchi* (unine18-012, où le participe passé *réfléchi* est analysé comme un adjectif) ; et *il est tellement exigeant mais autant envers lui-même qu'envers les autres* (unine15-068), avec une mésanalyse de la préposition *envers*. Ces deux résultats rendent attentif aux limites de l'étiquetage morphosyntaxique automatique, réalisé pour OFROM par le logiciel DisMo. Il faut encore soustraire de la liste de résultats une occurrence : *parler autant mal*, où *autant* est antéposé à l'adverbe *mal*, étiqueté comme un adjectif. Le fait qu'*autant* puisse commuter avec *aussi* dans ce contexte ne fait pas partie de la distribution de référence décrite par les grammaires.

Les occurrences obtenues montrent que la structure où *autant* est antéposé à un adjectif gradable est bien attestée en Suisse romande ; la *Grande Grammaire du Français* ne tient pas compte de cette variation. La structure comparative est observable lorsque l'adjectif est épithète (7a) et lorsqu'il est attribut (exemples 7b à 7f). La comparaison peut explicitement présenter les deux termes (7b-7c) ou bien être plus implicite (7d-7e).

- (7) (a) en vrai je suis pas habituée à des chaises **autant** confortables (unine19-005)
- (b) je m'attendais pas à ce que ce soit **autant** bon (unine08a02m)
- (c) je voulais pas **autant** grand mais j'ai pas eu le choix (unine 12-gmd)
- (d) j'ai jamais été **autant** malade en avion (unine11-sda)

(e) tu vas causer euh avec euh une personne toute simple qui est **autant** intelligente que l'un que l'autre hein (unine15-007)

(f) je l'ai jamais vu **autant** bourré déjà hier soir (unifr18a02d)

Dans le corpus, on observe également qu'*autant* peut modifier un adverbe – contrairement là aussi à ce que prescrit la *Grande Grammaire du Français* :

(8) (a) pis le détenu t'inquiète pas quand il est en balade il est il est **autant** bien que dans sa cellule hein (unifr19a14d)

(b) je pense qu'il est plus possible de parler **autant** mal qu'avant (unine19-002)

(c) c'est des personnes qui peuvent pas manger **autant** vite que nous (unine09-tba)

Ces premiers résultats donnent à penser qu'il existe, du moins à l'oral et en Suisse romande, un certain flottement de l'emploi des comparatifs par rapport à la catégorie morphosyntaxique sur laquelle ils portent. Il faudrait recueillir davantage d'exemples de cette construction pour en comprendre le fonctionnement et l'origine, mais ce bref « avant-goût » montre que le corpus OFROM apporte de précieuses ressources de première main et que son formulaire de recherche permet une analyse préalable des matériaux, en travaillant directement au niveau des étiquettes morphosyntaxiques.

En attendant une étude plus approfondie, on peut faire l'hypothèse que cet emploi de *autant* constitue un cas de maintien d'un usage ancien. Dans l'histoire de la langue en effet, la distribution des formes *aussi* et *autant* n'a pas toujours été aussi nette, à l'intérieur des structures comparatives. Ainsi, jusqu'au XVII^e siècle, l'adverbe *autant* pouvait précéder un adjectif (voir Combettes & Kuyumcuyan 2008).

Dans le système des comparatifs d'égalité du français, *aussi* et *si* sont réputés s'appliquer à un adjectif, et *autant* et *tant* à un verbe. Or l'emploi observé dans OFROM témoigne d'un phénomène de flottement sur l'appartenance à une classe. En effet, *autant* (ainsi que *tant*⁹ et *beaucoup*) – s'applique parfois à un adjectif (ex. 7), alors qu'*aussi* (ainsi que *très* et *si*) s'applique parfois à des verbes (ex. : *Personne n'a aussi souffert que lui* ; presse, *Le Figaro* : golf.lefigaro.fr, 9.3.2015). Ces unités en distribution complémentaire tendent dès lors à rejoindre la classe de celles qui se placent indifféremment devant un adjectif ou un verbe (comme *plus*, *moins*, *peu*, *assez*, *trop*). Les cas marqués sont réanalysés comme des termes neutralisant l'opposition (portée sur un adjectif vs portée sur un verbe).

4 OFROM Enseignement

4.1 Position du problème

Malgré l'évolution des technologies du son depuis plus d'un siècle, les avancées prometteuses de la linguistique de corpus et de la recherche sur le français parlé ces dernières décennies, leur impact sur l'enseignement des langues ne s'est pas fait ressentir autant qu'attendu (Surcouf & Ausoni 2021), du moins si l'on observe les méthodes de FLE sur le marché, qui s'appuient la plupart du temps, pour le développement des compétences orales, sur des documents fabriqués et interprétés par des comédiens (Ravazzolo *et al.* 2015 : 111). Dans ce secteur de l'édition, les corpus oraux brillent par leur absence¹⁰. On peut imputer ce désintérêt à divers obstacles, comme la difficulté de manipulation des outils et d'interprétation des données brutes, le manque de formation des enseignants sur la langue parlée, l'effort nécessaire à une transposition didactique ou le problème de l'exploitation commerciale des données (Johnsen 2022). En outre, l'idée que l'oral, en particulier sous sa forme spontanée, est une version imparfaite, déformée voire fautive de la langue est encore bien ancrée non seulement dans l'opinion collective, mais également dans le milieu de l'enseignement (Gagnon & Benzitoun 2020). La dimension régionale n'est pas en reste, avec l'idée répandue, y compris chez les locuteurs de Suisse romande, que plus on s'écarte du centre de diffusion parisien, moins « bon » est l'usage du français : ce jugement contribue de ce fait au sentiment d'insécurité linguistique des locuteurs des régions périphériques (Singy 1996, 2004). Toutefois, sous l'impulsion des spécialistes du français parlé ou de didacticiens du FLE, des projets pédagogiques visant à illustrer la variation du français en s'appuyant sur les corpus oraux ont vu le jour,

surtout dans le domaine du FLE (Detey *et al.* 2010¹¹, André 2016, Surcouf & Ausoni 2018, Ravazzolo & Etienne 2019, Laboratoire ligérien de Linguistique 2023), dont les outils sont essentiellement proposés sur des plateformes en ligne.

Avec OFROM Enseignement, nous nous inscrivons donc dans le sillage de ces démarches en proposant des ressources pédagogiques qui illustrent et thématisent la variation, à la fois dans ses dimensions diatopique, diaphasique et diamésique. Bien que les référentiels internationaux (Conseil de l'Europe 2001, 2018) ou émanant d'instances officielles en Suisse (CIIP, 2010) intègrent l'approche de la variation dans les objectifs à atteindre, les moyens qui tiennent compte des derniers résultats de la recherche et qui soient réellement adaptés aux besoins des apprenants restent peu répandus. À travers le volet Enseignement, nous souhaitons ainsi mettre du matériel à la disposition des enseignants et des apprenants qui permette d'aborder le français dans la variété de ses usages via l'exposition à des extraits authentiques de français parlé en Suisse romande. Les ressources ont été conçues en lien avec des enseignements universitaires de FLE ou des projets de recherche en cours à l'Université de Neuchâtel.

4.2 Propositions

En l'état actuel, les efforts se déploient dans trois directions : apprendre en autonomie (§ 4.2.1) ; diversifier les genres de parole (§ 4.2.2) ; répondre aux questions fréquentes (FAQ) sur le français en Suisse et sur le français parlé en situation (§ 4.2.3).

4.2.1 Apprendre en autonomie : outils de recherche

Moyennant une compétence numérique de base, ainsi qu'une familiarisation avec l'interface et avec le moteur de recherche, un apprenant (ou un enseignant dans une phase « teacher-controlled ») peut utiliser les fonctionnalités ordinaires de recherche linguistique dans une perspective d'apprentissage sur corpus (*Data-driven-learning*, désormais DDL) : les interrogations réalisées par l'intermédiaire du moteur de recherche peuvent servir à faire émerger chez « l'apprenti linguiste » l'observation de régularités linguistiques à travers la visualisation des résultats. Le processus d'apprentissage se voit individualisé, les types de requêtes variant selon les objectifs de chaque apprenant. Celles-ci sont susceptibles d'avoir trait à la syntaxe, au lexique, à la pragmatique ou encore à la prononciation. À titre d'exemple, observons une série de résultats du marqueur *en vrai*, courant à l'oral spontané, mais pratiquement introuvable dans les grammaires ou les manuels destinés à l'enseignement :

Locuteur	Texte	
1 unifr14-jba	en vrai j'ai envie soit d'être ambulancière de rentrer à l'école d'ambulancière à Lausanne.	Détails
2 unifr18-010	bon en vrai je suis déjà top soixante-sept c'est mieux.	Détails
3 unifr18-010	en vrai faudrait faire des écoles privées	Détails
4 unifr18-014	à part les décorations on a mis hyper longtemps après à faire les décorations mais en vrai c'est hyper rapide	Détails
5 unifr19-008	mais en vrai ça s'applique après forcément dans chaque	Détails
6 unifr19-034	mais en vrai c'est sympa le truc de euh un peu genre chasse au trésor ou je sais pas ce que c'est mais tu sais	Détails
7 unifr19-035	Je pense mais en vrai je sais pas.	Détails
8 unifr19-035	mais en vrai tu peux à part ça	Détails
9 unifr20-002	en vrai tu devrais t'acheter un vélo électrique hein	Détails
10 unifr22-070	mais après je me suis dit en vrai je crois en plus ils ont une soirée ce soir et pis elle sera genre juste dans ma chambre tu vois elle va a- arriver à vingt-trois heures elle va repartir à neuf heures quoi	Détails
11 unifr22-070	parce qu'en vrai euh un temps fallait vr- enfin c'était un peu comme le Popu quoi fallait	Détails
12 unifr22-070	en vrai m- moi ouais ouais moi j'aime bien mais c'est vrai que il y a un petit côté un peu euh	Détails
13 unifr22-070	ah non non en- en vrai il y en a assez c'est juste qu'on va souvent boire des cafés	Détails
14 unifr22-070	non mais en- oui oui moi ça me va bien je me disais on peut genre regarder deux trois frites si il y a parce que en vrai à Genève je suis jamais allée tu vois je sais pas exactement	Détails
15 unifr22-070	oui moi j'ai p- en en vrai j'aimerais juste trouver un pull blanc enfin blanc crème	Détails
16 unifr22-071	ouais exactement bon en vrai on se plaint mais ça va encore	Détails

Figure 10. Premiers résultats de la requête de la suite de caractères *en vrai*

D'abord, la possibilité d'écoute offre une expérience immersive dans une situation authentique, apportant des informations prosodiques et pragmatiques essentielles sur l'utilisation du marqueur en contexte. Ensuite, la visualisation des contextes met en évidence la prédominance de son antéposition (par rapport au noyau verbal) ainsi que des combinaisons privilégiées ou possibles (*mais en vrai, bon en vrai, parce qu'en vrai*). Cette démarche, qu'elle soit accompagnée par l'enseignant ou réalisée en autonomie par l'apprenant, favorise le développement d'une conscience langagière et de stratégies métacognitives favorables au processus d'appropriation (André & Ciekanski 2018).

Pour obtenir davantage de contexte, la fenêtre « Détails » – outre un élargissement de l'empan de l'extrait – permet de prendre connaissance à la fois du profil socio-démographique des locuteurs, de la relation entre les interlocuteurs (nature et degré de proximité) et du type d'interaction auquel appartient l'extrait, qui précisent ainsi les conditions d'usage de tel ou tel élément recherché en contexte authentique de parole.

4.2.2 Diversifier les genres de parole : le sous-corpus OFROM-Multigenres

L'orientation didactique adoptée par le volet Enseignement d'OFROM a permis d'intégrer de nouveaux types d'interaction au corpus, à des fins de représentation de la variation diaphasique. La liste des types d'interaction évolue au fil de l'alimentation du corpus selon les intérêts de l'équipe de recherche et n'est à ce titre pas échantillonnée de manière équilibrée. Cela dit, des genres allant de la distance communicative (y compris certains illustrant de l'écrit oralisé, par ex. les bulletins d'information) à la proximité y sont représentés (conversations, messages vocaux).

Il est possible d'effectuer une requête sur un type d'interaction, à partir du filtre « Enregistrement » qui donne accès au menu déroulant des types présents dans la base :

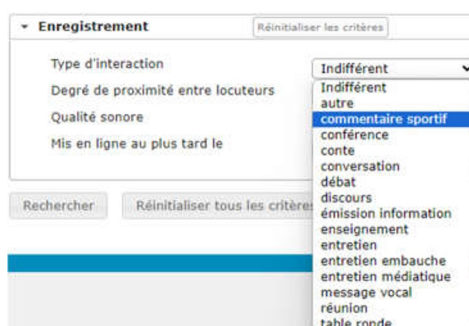


Figure 11. filtre « Enregistrement ».

À titre d'exemple, on peut songer à la recherche du pronom clitique *on*, dont la polysémie n'a pas d'équivalent dans les autres langues, pour observer ses différents emplois dans les genres « conférence » vs « message vocal ».

À moyen terme, un accès à une sélection d'extraits (audio et transcription) contextualisés et décrits (type, participants, degré de proximité, structure, actes de langage) est prévu à des fins d'utilisation en classe par les enseignants. De même, des rubriques thématiques permettront d'illustrer et de documenter un ensemble de phénomènes remarquables au niveau des rituels d'interaction, du lexique et de la grammaire notamment. L'exposition à des documents authentiques reflétant différentes situations de parole est l'un des éléments que l'équipe juge cruciaux pour favoriser les compétences de compréhension, interactionnelles et plus généralement sociolinguistiques des apprenants.

4.2.3 Répondre aux questions fréquentes (FAQ) sur le français en Suisse romande et sur le français parlé en situation

La nature même d'OFROM, à savoir le caractère oral et situé des données que le corpus rassemble, nous a poussés à proposer des ressources destinées à documenter et à illustrer le français sous des angles divers

pour un public de non spécialistes. À cet effet, deux foires aux questions (FAQ), accessibles sur le site web via l'onglet *Enseignement*, ont été élaborées dans le but de répondre aux interrogations possibles des utilisateurs (enseignants, apprenants ou simples intéressés) en évitant autant que possible la terminologie spécialisée. Pour chacune d'entre elles, un développement vulgarisé est proposé, accompagné d'illustrations sonores et cartographiées, lorsque cela s'avère pertinent. Outre des explications de nature encyclopédique, ces notices abordent les limites de certaines notions, dans une démarche qui engage une posture critique. La FAQ « Le français en Suisse romande » s'interroge sur l'origine, la nature et les caractéristiques du français parlé en Suisse romande, ainsi que sur les représentations de cette variété de français. La FAQ « Le français parlé en contexte » aborde quant à elle le statut de l'oral en regard de l'écrit, les représentations des locuteurs sur l'oral, les différences avec l'écrit et les conditions de production de l'un et l'autre médium. Les réponses cherchent notamment à nuancer les représentations sur la nature présumée fautive de l'oral, en comparaison avec les formes idéalisées de l'écrit. Elles permettent également de se livrer à des comparaisons sonores immersives entre des genres et des situations à l'oral diversifiées. Ces ressources sont rendues accessibles au fur et à mesure de l'avancement des travaux de l'équipe.

5 Conclusion

OFROM présente plusieurs spécificités dans le panorama des corpus de français parlé. D'une part, le corpus est totalement en libre accès. D'autre part, le système permet d'élargir le contexte et découpe les sons lui-même, ce qui est particulièrement commode pour les utilisateurs. Il donne l'accessibilité à la fois à la transcription et au fichier audio correspondant, et fournit une documentation détaillée sur le profil de chaque locuteur. Enfin, les données ne sont pas issues d'un grand centre urbain en particulier, mais recueillies dans des agglomérations de différentes tailles. On peut ajouter que, moyennant parfois quelques ajustements, les données sont exploitables quel que soit le niveau d'analyse considéré. De ce point de vue également, l'archive OFROM est complémentaire d'autres corpus qui présentent une orientation clairement identifiée (ESLO, orienté sociolinguistique ; PFC, orienté phonologie ; FLORALE, orienté FLE, etc.)

Le corpus a été exploité pour un nombre conséquent d'études sur le français tel qu'il est pratiqué dans cette communauté de locuteurs et nombreux sont aujourd'hui les linguistes à l'utiliser dans le cadre de leurs recherches et de leurs enseignements, en Suisse comme à l'étranger.

Références bibliographiques

- André V. (2016). « FLEURON : Français Langue Étrangère Universitaire – Ressources et Outils Numériques. Origine, démarches et perspectives », *Mélanges Crapel*, 37, 69-92.
- Arrivé M, Galmiche M. & Gadet F. (1986). *La grammaire d'aujourd'hui : Guide alphabétique de linguistique française*. Paris : Flammarion.
- Avanzi M., Béguelin M.-J. & Diémoz F. (éds) (2016a). *Corpus de français parlé et français parlé des corpus*, numéro thématique de la revue *Corpus* n°15.
- Avanzi M., Béguelin M.-J. & Diémoz F. (2016b). « De l'archive de parole au corpus de référence. Le corpus oral de français de Suisse romande (OFROM) », *Corpus* 15, 309-342.
- Avanzi M., Béguelin M.-J., Corminbœuf G., †Diémoz F. & Johnsen L. A. (2012-2024). *Corpus OFROM – Corpus oral de français de Suisse romande*. Université de Neuchâtel, <ofrom.unine.ch>
- Béguelin M.-J., Corminbœuf G. & Johnsen L. A. (2014). « Réflexions sur le statut de la métanalyse », *Verbum* XXXVI-1, 3-16.
- Benzitoun C. & Cappeau P. (à paraître). « Les corpus et leur exploitation », *Encyclopédie grammaticale du français*. <www.encyclogram.fr>
- Berrendonner A. (2021). « Les 'adverbes d'énonciation' », *Verbum*, XLIII/2, 227-244.
- Biggs P. & Dalwood M. (1976). *Les Orléanais ont la parole : Teaching Guide and Tapescript*. Londres : Longman (Livre du maître).

- Blanche-Benveniste C. & Jeanjean C. (1987). *Le français parlé : transcription et édition*, Paris, Didier Érudition.
- Blanche-Benveniste C. (2001). « Grammaticalisation d'un terme de lieu : *quelque part* et *mis à part* ». *Recherches sur le français parlé* 16, 83-101.
- Boersma P. & Weenink D. (2023), « Praat, v. 6.3.14 », accessible à <<https://www.fon.hum.uva.nl/praat/>>
- Bruxelles S., Mondada L., Simon A. C. & Traverso V. (éds) (2009). *Grands Corpus de français parlé : Bilan historique et perspectives de recherche* (= *Cahier de linguistique de Louvain* 33-2), Louvain, Presses universitaires de Louvain.
- 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 », *Proc. LREC*, 3902-3907.
- CIIP (2010). *Plan d'études romand* (PER), Neuchâtel, <<https://www.plandetudes.ch/>>
- Combettes B. & Kuyumcuyan A. (2008). « Comme dans les comparaisons d'égalité : la corrélation 'aussi/autant... comme' jusqu'à l'époque classique », *Langue Française* 159, 16-32.
- Conseil de l'Europe (2001). *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Paris : Didier.
- Conseil de l'Europe (2018). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Volume complémentaire avec de nouveaux descripteurs*. Strasbourg : Conseil de l'Europe.
- Corminbœuf G., Rothenbühler J. & Sauzet M. (éds) (2020). *Français parlés et français 'tout court'*. *Studia Linguistica Romanica* n°4. <<https://studialinguisticaromanica.org/index.php/slr/issue/view/4>>
- Detey S., Durand J., Laks B & Lyche, Ch. (2010). *Les variétés du français parlé dans l'espace francophone. Ressources pour l'enseignement*. Paris/Gap : Ophrys.
- Diémoz F., Rothenbühler J. & Sauzet M. (2020), « Que nous apprend OFROM du français régional d'aujourd'hui en Suisse romande ? Premières investigations entre pratiques et métadiscours », *Studia Linguistica Romanica* 4. <<https://studialinguisticaromanica.org/index.php/slr/article/view/46>>
- DSR = Thibault A. (2004²). *Dictionnaire suisse romand*. Carouge : Zoé.
- Durand J., Laks B. & Lyche Ch. (2002). « La phonologie du français contemporain : usages, variétés et structure », in Pusch K. & Raible W. (dir.), *Romance Corpus Linguistics – Corpora and Spoken Language*. Tübingen : Gunter Narr Verlag, 93-106.
- Durand J., Laks B. & Lyche Ch. (2009). « Le projet PFC (phonologie du français contemporain) : une source de données primaires structurées », in Durand J., Laks B. & Lyche Ch. (eds), *Phonologie, variation et accents du français*. Paris : Hermes, 19-62.
- Fuchs C. (2009), « Les marqueurs de comparaison du français : un air de famille avec l'italien », in C. Calvo et al. (éds.), *Miradas cruzadas : Estudios franco-italianos*, Presses de l'université de Valencia, 93-111.
- Gagnon R., & Benzitoun Ch. (2020). « Le français parlé comme objet d'enseignement ? Regards croisés d'un linguiste et d'une didacticienne », *Formation et pratiques d'enseignement en questions - Revue des HEP de Suisse romande et du Tessin*, 26, 37-51.
- Groupe de Fribourg (2012). *Grammaire de la période*. Bern et al. : Peter Lang.
- Johnsen L. A. (2020). « OFROM, *Corpus Oral de français de Suisse romande* : une ressource pour la recherche... et pour l'enseignement/apprentissage du français », *Babylonia*, 1, 45-53. Disponible en ligne : <<http://babylonia.ch/fr/archives/2020/numero-1/>>
- Johnsen L. A. (2022). « Approche du français parlé dans une perspective d'enseignement : quelques pistes d'exploitation du corpus OFROM », *Travaux de linguistique* 84-85, 211-229.
- Kleiber G. & Gerhard-Krait F. (2006). « *Quelque part* : du spatial au non spatial en passant par l'indétermination et la partition », *Journal of French Language Studies* 16-2, 147-166.

- Laboratoire Ligérien de Linguistique (2023). *ESLO-FLEU : Enquêtes Sociolinguistiques à Orléans – Fle et Linguistique pour l’Enseignement Universitaire* [Corpus]. ORTOLANG – www.ortolang.fr, v1, <<https://hdl.handle.net/11403/eslo-fleu/v1>>
- Mertens, P. (2004). « Un outil pour la transcription de la prosodie dans les corpus oraux », *Traitement Automatique des langues*, 45/2, 109-130.
- Ravazzolo, E. & Étienne, C. (2019). « Nouvelles ressources pour le FLE à partir des études en interaction », *Linx*, 79.
- Ravazzolo, E., Traverso, V., Jouin, É. & Vigner, G. (2015). *Interactions, dialogues, conversations : l’oral en français langue étrangère*, Paris : Hachette.
- Riegel, M., Pellat J.-C. & Rioul, R. (2018). *Nouvelle grammaire méthodique du français*, Paris, PUF.
- Rousset F. & Maikal K. (2022). *Le monde francophone. Parcours digital*. Copenhague : Praxis Forlag. <<https://praxis.dk/le-monde-francophone>>
- Singy, P. (1996). *L’image du français en Suisse romande. Une enquête sociolinguistique en Pays de Vaud*. Paris : L’Harmattan.
- Singy P. (2004). *Identités de genre, identités de classe et insécurité linguistique*. Berne : Peter Lang.
- Surcouf C. & Ausoni A. (2018), « Création d’un corpus de français parlé à des fins pédagogiques en FLE : la genèse du projet FLORALE », *EDL (Études en didactique des langues)*, 31, 71-91. <<https://florale.unil.ch/>>
- Surcouf C. & Ausoni A. (2021). « Du phonographe aux corpus numériques oraux : aperçu historique de la compréhension du français oral quotidien en FLE », *Bulletin suisse de linguistique appliquée*, 2, 31-47.

¹ La base de données OFROM a été conçue à l’origine par Marie-José Béguelin, professeure ordinaire en linguistique française à l’Université de Neuchâtel, et Mathieu Avanzi, alors assistant. Réalisé par Pierre Ménétreay, le site Internet d’OFROM a rendu le corpus officiellement accessible dès le début de l’année 2013, muni d’un moteur de recherche multicritères. Au départ à la retraite de Marie-José Béguelin, Federica Diémoz, devenue directrice du Centre de dialectologie et d’étude du français régional en 2014, a repris la direction d’OFROM, avant de disparaître brutalement en 2019. À ce moment-là, Laure Anne Johnsen (Université de Neuchâtel) et Gilles Corminbœuf (Université de Fribourg), secondés par l’équipe mise en place par Federica Diémoz (François Delafontaine, Julie Rothenbühler et Maguelone Sauzet), ont assuré le pilotage du projet – rejoints par Mathieu Avanzi, nommé en 2022 directeur du Centre de dialectologie. Depuis ses débuts, le corpus OFROM est le fruit d’une collaboration pérenne entre les Universités de Neuchâtel et de Fribourg, qui ont apporté un soutien financier périodique pour le développement de la base.

² Les principales conventions de transcription sont disponibles sur le site web.

³ Au plan légal, chaque témoin signe une autorisation pour l’exploitation des données. Celles-ci sont anonymisées et un droit de rétractation est accordé au témoin, qui peut demander le retrait de l’enregistrement du corpus.

⁴ Certaines informations non renseignées sont notées « NR » dans les métadonnées. La Figure 4. ne prend bien sûr pas en compte les données étiquetées NR.

⁵ Le test du clivage n’est pas entièrement probant, ce qui tient à la signification indéterminée du marqueur locatif, difficilement compatible avec l’extraction (Blanche-Benveniste 2001 : 89). De même pour l’équivalence avec *là*, qui n’est pas très bonne, en raison de la tension entre la valeur déterminée de *là* et l’indétermination de notre forme (*ibid.*).

⁶ Une occurrence ambiguë (ex. 5, *infra*) a été écartée des comptages, ce qui explique que l’addition des chiffres donne 78 et non 79.

⁷ Les tracés intonatifs ont été réalisés avec l’outil Prosogramme (Mertens 2004).

⁸ Une commutation des deux adverbes est possible dans certains cas où l’analyse de leur portée est soumise à l’interprétation variable du locuteur. Par exemple, dans une structure où le syntagme verbal comporte un participe passé conjugué avec l’auxiliaire *être* (*il est aussi / autant perturbé par la situation...*). La *Grammaire méthodique du français* (Riegel *et al.* 2018 : 623) signale le cas où *autant* est postposé au nom, et endosse un sens qualitatif ([...] *cette femme froide et tranchante autant que l’acier*, Balzac). De même, *La Grammaire d’aujourd’hui* (Arrivé *et al.* 1986 : 207), après avoir souligné qu’*autant* fonctionne dans un syntagme verbal, mentionne le cas où, « postposé à un adjectif, il se substitue à *aussi* : *il est travailleur autant qu’intelligent* ». Les exemples du type de ceux tirés d’OFROM, où *autant* est antéposé à l’adjectif (cf. *autant intelligente*, voir *infra* sous 7), ne sont en revanche pas mentionnés dans ces deux grammaires.

⁹ Dans le Midi de la France, des constructions comme *elle est tant grande comme toi* existaient encore récemment à l'oral : la comparaison d'égalité introduite par l'adverbe *tant* selon un schéma corrélatif s'explique par l'influence de l'occitan, où il s'agit de la construction courante.

¹⁰ On peut citer quelques exceptions : *les Orléanais ont la parole* (à partir du corpus ESLO1) (Biggs & Dalwood 1978), *les Variétés de français dans l'espace francophone* (Detey *et al.* 2010), un manuel de grammaire du français pour apprenants danois basé sur OFROM (Rousset & Maikal 2022).

¹¹ PFC se veut plus généraliste et ouvre ses ressources didactiques à l'enseignement du FLM également.