Le *Corpus de référence du français contemporain* (CRFC), un corpus massif du français largement diversifié par genres

Siepmann, Dirk[1], & Bürgel, Christoph[2], & Diwersy, Sascha[3]

1 Universität Osnabrück, Institut für Anglistik/Amerikanistik
dirk.siepmann@uni-osnabrueck.de
2 Universität Paderborn, Institut für Romanistik
christoph.buergel@upb.de
3 Praxiling, UMR 5267, Université Paul-Valéry Montpellier 3
sascha.diwersy@univ-montp3.fr

Résumé. Cet article porte sur le Corpus de référence du français contemporain (ci-après abrégé en CRFC), un nouveau corpus qui, tout en présentant une taille considérable, vise, dans sa conception, à un équilibre en termes de genres textuels. La première version du corpus, qui sera enrichi au fur et à mesure de la disponibilité de nouveaux documents, compte 310 millions de mots du français tel qu'il se parle et s'écrit en France pour une période comprise entre 1945 et 2014, avec plus de 90 % de textes remontant aux deux dernières décennies. Ce corpus est destiné à représenter la langue française de telle manière qu'il réponde aux besoins des apprenants, des enseignants et des chercheurs en français contemporain.

Abstract. The Corpus de référence du français contemporain (CRFC) is a new purpose-built genre-diverse corpus for investigating modern French. The 310-million-word corpus is the first collection of French to incorporate a substantial amount of spontaneous speech (approx. 30 m words) and 'pseudo-spoken' data (approx. 125 m words); it is evenly divided between spoken/pseudo-spoken and written sources. The present article discusses major issues relating to the design of the corpus and the sources used in compiling it.

1 Introduction

L'analyse outillée du français enregistre un retard par rapport à d'autres langues majeures en termes de diversité et de disponibilité de corpus ainsi que de sophistication de l'analyse statistique (Deulofeu & Debaisieux, 2012:36). Par conséquent, il n'a pas été possible, à ce jour, d'élaborer des grammaires du français fondées sur corpus ni de parvenir à une description lexico-statistique fiable des collocations et des colligations récurrentes en français parlé (Siepmann, 2015), pour ne retenir que deux exemples.

Cet article porte sur le *Corpus de référence du français contemporain* (ci-après abrégé en CRFC), un nouveau corpus de grande taille, équilibré et largement diversifié par genres textuels, qui a été réalisé dans le but de remédier aux défaillances de l'état actuel mentionnées ci-dessus. La première version du corpus, qui sera enrichi au fur et à mesure de la disponibilité de nouveaux documents, compte 310 millions de mots du français tel qu'il se parle et s'écrit en France pour une période comprise entre 1945 et 2014, avec plus de 90 % de textes remontant aux deux dernières décennies. Ce corpus est destiné à représenter la langue française de telle manière qu'il réponde aux besoins des apprenants, des enseignants et des chercheurs en français contemporain. Il se distingue des précédents corpus du français à plusieurs égards :

1. C'est le plus vaste corpus du français ne reposant pas exclusivement sur des sources internet
2. Il couvre un large éventail de genres textuels divers.
3. C'est le premier corpus du français à réserver une place importante à la parole spontanée (environ 30 millions de mots) et « pseudo-orale » (à savoir la langue qui se caractérise par« l'immédiat communicatif » [voir Section 2.2 pour de plus amples détails] ; environ 125 millions de mots)
4. Le corpus sera mis à jour en permanence et enrichi chaque année, à compter de 2014, d'environ 25 millions de mots à partir de données orales et écrites.

À l'instar du projet britannique *Cobuild* lancé à l'Université de Birmingham dans les années 80 et 90 (voir Sinclair, 1987), le CRFC a pour objectif de servir à l'élaboration de dictionnaires, de grammaires et de supports pédagogiques dans le cadre d'une approche à dominante inductive ou « corpus-driven » (pour reprendre les termes de Tognini Bonelli, 2001). Il permettra de prendre en considération certaines questions auxquelles ne permettaient pas de répondre les corpus existants, qu'il s'agisse des variations lexico-grammaticales selon différents genres textuels ou de la phraséologie du français parlé.

Le CRFC sera consultable en ligne fin 2018, à l'issue d'une utilisation réservée aux compilateurs pendant trois ans. Avant cette date, les chercheurs individuels peuvent obtenir l'autorisation d'y accéder sur demande. Le CRFC est actuellement installé sur SketchEngine (accès privé), mais, une fois expiré le délai de trois ans, il deviendra accessible sur la plateforme PrimeStat développée par S. Diwersy (voir Diwersy, 2013) aux universités de Cologne et de Montpellier.

La partie suivante traite des principales caractéristiques structurelles du CRFC. Pour une analyse critique des autres corpus du français, le lecteur se référera à Siepmann (2015), Siepmann & Bürgel (2015) et Siepmann et al. (2015).

2 Conception et traitement du corpus

Cette section présente la composition du CRFC (2.1) ainsi que l'éventail des sources sélectionnées (2.2) avant d'aborder brièvement la méthodologie utilisée pour son uniformisation et son annotation (2.3).

2.1 Composition et principes d'échantillonnage

Le CRFC a été conçu pour répondre aux besoins de plusieurs catégories d'utilisateurs :

1. des linguistes qui souhaitent mener des recherches sur le français contemporain parlé et écrit tel qu'il se manifeste dans divers genres ;

2. des didacticiens de la langue qui peuvent utiliser le corpus comme outil pour la préparation de dictionnaires, de livres de vocabulaire, de grammaires et de supports pour l'enseignement d'autres langues ;
3. des professeurs de français qui pourraient se servir du corpus pour la préparation et l'évaluation de leurs cours ;
4. des traducteurs et des rédacteurs techniques susceptibles de se poser des questions précises sur l'emploi de certains termes ;
5. des étudiants qui ont besoin d'être sensibilisés aux régularités linguistiques

L'une des principales préoccupations consistait à réunir un corpus assez important dans un délai raisonnable et à un coût abordable. Cela veut dire qu'il a fallu faire des concessions au niveau de l'annotation du corpus. On a décidé, par exemple, de ne pas annoter ce qui relève de la sémantique ou de la prosodie.

La composition du CRFC s'est inspirée des deux corpus de référence les plus importants de l'anglais, à savoir le BNC et le COCA[1], la spécificité du CFRC étant que la diversité des genres y excède de loin celle de n'importe quel corpus antérieur. L'objectif est de garantir un niveau raisonnable d'équilibre et de représentativité, tout en sachant qu'il s'agit là d'un idéal statistique qui, comme l'ont montré de manière convaincante Atkins et al. (1992) et Evert (2006), ne s'applique qu'imparfaitement aux faits langagiers.

Si déplorable que cela puisse paraître, il semble effectivement assez illusoire de vouloir constituer un corpus de langue usuelle en respectant à la lettre tous les principes théoriques d'échantillonnage et d'inférence préconisés dans les manuels canoniques de méthodologie statistique. Ceci tient non seulement au fait qu'il s'avère très difficile de délimiter de manière rigoureuse la population entière qu'un tel corpus est censé représenter, mais aussi à ce qu'il n'existe aucune unité évidente du langage susceptible d'être échantillonnée et utilisée pour définir cette population (Atkins et al., 1992:7).

Dans ces conditions, une stratégie d'échantillonnage défendable consiste au moins à à diversifier largement les genres discursifs en visant à ce que Corbin a appelé « un équilibre appréciable entre contrôle et naturel » (2005 : 131), et le CRFC, qui comporte, en proportions égales, aussi bien des sources écrites que des sources orales ou pseudo-orales, est sans doute l'un des premiers corpus résultant de la mise en œuvre d'une telle stratégie. En témoigne le tableau 1, qui donne un aperçu des sous-échantillons qui constituent le CRFC dans sa version actuelle.

Catégorie macro-générique	**Sous-échantillon**	**Taille (en millions de mots)**
oral	interactions par oral spontané	30
pseudo-oral (écrit mimétique de l'oral)	pièces de théâtre et scénarios de film	30
	sous-titres de films et de feuilletons télé quotidiens	2,5
	textos/*chats*	2,5
	forums de discussion	60
pseudo-écrit (oral mimétique de l'écrit)	oral formel (allocutions, discours, informations ...)	30
		155
écrit	universitaire et scientifique	30
	ouvrages non universitaires	30
	roman et fiction en prose	30
	journaux	45
	revues et magazines	10
	journaux intimes et blogs	5
	lettres et courriels	1
	textes divers	4
		155

Tableau 1 : Composition du CRFC

Les critères qui sous-tendent la classification des genres nécessitent quelques explications. Par « pseudo-oral », on entend les genres de textes qui, tout étant produits à l'écrit, revêtent une conception orale, relevant ainsi davantage de ce qui a été qualifié par Koch & Oesterreicher (2011) de « proximité communicative ». Cette dernière s'oppose à la « distance communicative » en tant que pôle extrême d'un continuum défini par une configuration de paramètres qui se distinguent comme suit:

	Proximité	**Distance**
1.	communication privée	communication publique
2.	interlocuteur familier	interlocuteur inconnu
3.	émotionnalité forte	émotionnalité faible
4.	contextualisation	dé-contextualisation
5.	ancrage référentiel dans la situation	éloignement référentiel par rapport à la situation
6.	co-présence spatiale et temporelle	séparation spatiale et temporelle
7.	coopération communicative intense	coopération communicative minimale
8.	dialogue	monologue

	Proximité	**Distance**
9.	communication spontanée	communication préparée
10.	liberté thématique	fixité thématique

Tableau 2 : Paramètres de la proximité et de la distance communicative selon Koch & Oesterreicher (2011:13)

En tenant compte de ces paramètres, il semble raisonnable de présumer que les films, les pièces, les *chats* et les forums de discussion tendent vers le pôle de proximité du continuum tandis que les journaux intimes, les blogs, les lettres et les emails sont légèrement plus proches du pôle de distance. Ces décisions concernant la classification des genres sont corroborées lorsqu'on analyse les sous-échantillons respectifs en termes de la distribution fréquentielle de certains traits généralement considérés soit comme formels, soit comme informels. Considérons, par exemple, la distribution du terme familier *mec*, dont la fréquence relative est nettement plus élevée dans les genres pseudo-oraux que dans les journaux intimes et blogs ou dans les lettres et courriels. Ceci est indiqué par le tableau 3 qui donne les chiffres calculés pour *mec* au moyen de l'application «Relative Text Type Frequency » de la plate-forme *SketchEngine*, que nous utilisons en ce moment pour l'exploitation du CRFC :

Sous-échantillon	**Fréquence relative (*Relative Text Type Frequency*)2**
CRFC SMS [CRFC textos]	542,50
CRFC discussion forums.txt [CRFC forums de discussion]	286,10
CRFC Film.txt [CRFC films]	279,10
CRFC Plays.txt [CRFC pièces de théâtre]	137,10
CRFC spoken informal TV.txt [CRFC langue parlée familière à la télévision]	123,40
CRFC Fiction.txt	60,40
CRFC spoken informal except TV.txt [CRFC langue parlée familière (hors télévision)]	31,60
CRFC magazines.txt	10,00
CRFC Diaries and Blogs.txt [CRFC journaux intimes et blogs]	7,80
CRFC Newspapers.txt [CRFC journaux]	5,80
CRFC Acad Merge.txt [CRFC tous textes universitaires confondus]	3,90

Sous-échantillon	**Fréquence relative (*Relative Text Type Frequency*)2**
CRFC letters and e-mails.txt [CRFC lettres et courriels]	1,90
CRFC Non-acad.txt [CRFC non universitaire]	1,10
CRFC spoken formal.txt [CRFC langage soutenu]	0,60
CRFC Miscellaneous.txt [CRFC textes divers]	0,30

Tableau 3 : Fréquence relative de *mec* selon genres discursifs (tels que représentés par les sous-échantillons respectifs du CRFC)

2.2 Sources sélectionnées

On notera en particulier l'ampleur de la partie orale informelle, un domaine où on devait jusqu'à présent se contenter de corpus de taille limitée et représentant une gamme peu diversifiée de situations langagières (principalement des interviews ; voir Debaisieux, 2010, Gadet et al., 2012). Le corpus de parole informelle se compose à 75 % de transcriptions de dialogues ou de monologues en majorité spontanés, provenant de plus de 200 types différents d'émissions télévisées qui ont été diffusées par France 2, France 3 et France 5 entre 2013 et 2014, au total plus de 6 000 émissions, soit plus de 3 000 heures d'une parole qui s'exprime plutôt librement et naturellement (on en trouvera quelques illustrations plus loin). On a retenu sans discrimination une année entière d'émissions afin d'éviter de fausser le contenu du corpus en privilégiant tel jour de la semaine ou telle période de l'année (voir Kennedy, 1998:75). Comme l'a affirmé Meißner (2006:248-249) pour tenter de répondre à la question « quel français enseigner? », la télévision a depuis longtemps fixé la règle statistique en donnant à entendre et à voir aux masses une multitude d'idiolectes et de variétés linguistiques, tout en les présentant de telle sorte que ces dernières puissent être comprises par la majeure partie des téléspectateurs et des auditeurs dans les zones linguistiques et de radiodiffusion concernées. Par conséquent, rien ne s'oppose en principe à ce qu'un compilateur de corpus utilise toute la gamme des variétés langagières parlées dans les médias comme point de référence pour la composition de corpus oraux et le développement de supports pédagogiques.

Parmi les émissions retenues, on mentionnera les débats-spectacles comme *Toute une histoire*, les émissions politiques comme *C à dire ?!*, les émissions culturelles comme *Entrée libre*, les jeux culinaires comme *Dans la peau d'un chef*, diverses émissions sportives ou scientifiques, ainsi qu'un très grand nombre de documentaires sur toutes sortes de sujets. On doit inévitablement recourir à des données empruntées à la télévision (ou à la radio) quand on cherche à fournir un corpus étendu et varié de la langue parlée ; comme le dit Davies, ce matériau s'avère « représenter assez bien la parole spontanée telle qu'on l'entend dans la vie de tous les jours » (http://corpus.byu.edu/coca/). En fait, l'authenticité de la partie du CRFC réservée au français parlé informel dépasse même celle du COCA, car il contient une grande quantité de séquences documentaires montrant des échanges de la vie courante entre interlocuteurs, y compris de réelles conversations privées. Des analyses de corrélation, des algorithmes de classifications et des tests d'indépendance à base de n-grammes tirés de corpus de sous-titres montrent que la langue employée pour le sous-titrage au cinéma ou à la télévision ressemble fortement à celle de la conversation informelle (voir Levshina, à paraître) ; le même constat vaut pour la langue des séries télévisées (voir Quaglio, 2009). Il faut cependant reconnaître qu'il existe de légères différences entre les données télévisuelles et les autres données du langage oral, les

premières étant marquées par une plus grande fluidité de l'expression et moins de phénomènes d'hésitation. Voici quelques exemples à titre d'illustration :

> *Ça a commencé comment? O. Pruvost: Au milieu des années 90, bêtement, j'ai l'impression. Ça a commencé en faisant du cheval. J'ai l'impression d'avoir tapé sur la selle, continuellement, pendant cette promenade de cheval qui a duré 2 heures. Ça a provoqué un tassement des disques. Ça a été le point de départ de ces douleurs. Marina: Vous étiez déjà sportif? O. Pruvost: Oui. Michel: C'était quelle douleur? O. Pruvost: Ça ne s'est jamais vraiment déclaré au niveau des sciatiques. Par contre, en crise, c'était épouvantable. Je me laissais glisser contre un mur sur le sol pour apaiser la douleur. Michel: C'était dans le bas du dos, les lombaires? O. Pruvost: Oui. Marina: C'était déclenché par quoi? De mauvaises positions?* (émission de santé)
>
> *-Appuie ! Voilà ! Allez ! -Il faut le pencher dans l'autre sens, le bidon ! Il faut anticiper ! Avec la bascule, il faut mettre le goulot vers l'avant. -Réfléchis, Nath ! -Penche-le au maximum vers l'avant ! Penche le bidon vers l'avant ! -OUAIS ! -C'est beau ! -Allez ! Vas-y, championne ! -T'es à mi-parcours ! -Tranquille. Voilà. -C'est bon ! C'est bon ! -C'est bien, Nath !* (jeu télévisé)
>
> *-Là, je suis en train de faire le pain. Pain bio au levain maison. Vu que les ravitaillements ici, c'est tous les mois, ben le pain, on arrive difficilement à le garder pendant un mois, donc à l'héliportage, on monte du pain et après, on le fait ici. Quand il y a du monde, c'est tous les jours. Sinon, avec moins de monde, tous les 2 jours. Avant tout, il faut aimer la montagne. Sinon, passer 5 mois ici dans l'année, on ne tient pas longtemps.* (documentaire)

Les 25% restant de la partie consacrée au langage familier comportent divers corpus oraux accessibles sous la licence Creative Commons, comprenant, par exemple, le corpus TALN, le CoLaJE et des parties des corpus ESLO 1 et 2, avec des données recueillies entre la fin des années soixante et nos jours.

La partie orale formelle se compose de discours politiques et autres, de transcriptions de conférences académiques, ainsi que de débats parlementaires à l'Assemblée Nationale ou au Sénat.

La partie du corpus réservée au français pseudo-oral a été compilée à partir de pièces de théâtre et de scénarios de film, de sous-titres, de forums de discussion et de textos ou de messages de *chat*. Les œuvres théâtrales et les scénarios de films ont été extraits de plusieurs sites internet, tels que www.leproscenium.com et http://theatrentreprise.free.fr, et les sous-titres téléchargés à partir de www.subsynchro.com/tous-les-films-francais.html. On a ciblé différents forums de discussion, parmi lesquels www.doctissimo.fr, qui propose sur ses forums une large palette de sujets allant de la psychologie au voyage et aux sports en passant par la santé. Les messages de *chat* sont issus du corpus de Falaise (2005), les textos, eux, du récent corpus de SMS français (Panckhurst et al., 2013).

La partie écrite du corpus se répartit en six catégories : universitaire et scientifique, ouvrages non-universitaires, roman et fiction en prose, journaux, revues et magazines, journaux intimes et blogs, lettres et emails, documents divers. On a pris soin de ne pas privilégier une catégorie au détriment des autres, comme on l'a souvent observé avec la langue de la presse écrite dans de nombreux corpus antérieurs. La partie du corpus portant sur les écrits universitaires a été calquée sur le COCA. Pour ces articles, on a pioché dans les revues et publications scientifiques consultables en ligne afin de couvrir un large éventail de sujets en arts et en sciences. L'intégralité de livres non universitaires et des échantillons de ces derniers ont été téléchargés à partir de sites tels que http://livreslib.com ; on a donné la préférence à la langue générale plutôt qu'à la langue spécialisée. Des romans entiers et des œuvres de fiction plus courtes, écrits entre 1945 et aujourd'hui, ont été tirés de différents sites, dont www.edition-grasset.fr ; les compilateurs ont aussi saisi quelques livres de littérature de jeunesse. On a obtenu, sur leurs sites respectifs, les éditions d'un quotidien national (*Le Monde*) et de divers journaux régionaux (*Sud-Ouest, L'Est Républicain, Le télégramme de Brest, La Voix du Nord*)

pour la période 2012-2013. On a également téléchargé des extraits d'un certain nombre de magazines et revues diverses, en veillant à équilibrer les domaines de spécialité (informations économiques, presse féminine, informatique, droit, finance, santé, décoration et jardinage, défense de la langue française, etc.). Pour des raisons d'accès aux ressources, les parties du corpus écrit sont plus ou moins longues, la catégorie « divers » étant la plus modeste (par exemple articles de lois et règlements, modes d'emploi, dépliants destinés aux patients, bulletins météo, rapports techniques, lettres d'information, etc.). Le corpus est conçu pour permettre à la fois la consultation par sous-corpus et des comparaisons entre ces derniers.

Outre l'organisation en fonction des types de médium et de genre, nous travaillons actuellement à une classification du contenu par grands domaines thématiques et sujets particuliers, à la manière du *Longman/Lancaster English Language Corpus* (Summers, 1993) et de l'*Oxford English Corpus*. Summers a fait remarquer (1991:7 ; 1993) que cette répartition par thèmes facilite les recherches lexicales. Le corpus Longman/Lancaster repose sur 16 « super-domaines » thématiques ; le nombre est identique pour notre corpus, mais nous avons organisé les thèmes différemment : arts ; affaires et économie ; politique, gouvernement et lois ; informatique ; environnement ; sciences et recherche ; santé et médecine ; croyances et pensées ; psychologie et rapports sociaux ; loisirs, divertissements, sports ; alimentation ; vêtements et mode ; voyages, tourisme et transports ; décoration et jardinage ; histoire ; communication et mass media. Ainsi, on pourra étudier les structures lexico-grammaticales utilisées dans le football sous l'intitulé assez vaste « loisirs, divertissements, sports » ; on y trouvera des livres et des articles sur le football, des discussions tirées de forums sur le football, des commentaires télévisuels autour du même sujet, etc.

Comme il ressort clairement de ce qui précède, le CRFC est un corpus de textes complets. On assiste actuellement à un débat en linguistique de corpus sur le risque de fausser les corpus en suivant une politique du texte intégral (voir McEnery & Hardie, 2012:152-153), puisque les textes autonomes sont susceptibles de comporter un lexique très idiosyncratique ou spécialisé. Nous avons cherché à contourner ce problème en faisant en sorte qu'aucun des textes retenus ne dépasse une longueur disproportionnée. Hormis le fait qu'elle permet aux linguistiques d'étudier la collocation textuelle (Hoey, 2005:115 ff.), la prise en compte de textes entiers dans la constitution d'un corpus a pour avantage évident que celui-ci se prête à des utilisations variées , les textes abrégés de manière aléatoire ne fournissant pas de données exploitables dans le cadre d'études relevant de domaines aussi divers que la linguistique textuelle, l'analyse du discours ou la didactique du FLE.

2.3 Normalisation et annotation du corpus

La compilation du corpus a nécessité un important travail de préformatage automatique, y compris la suppression des « boilerplates » contenus dans les documents issus du web[3]. La structuration du corpus a été effectuée selon un schéma XML simpliste[4], qui ne distingue pour l'instant que les différents textes et les sous-échantillons définis par genre textuel qui les regroupent. Quant à l'annotation linguistique, la version actuelle du corpus a été catégorisée et lemmatisée à titre provisoire au moyen de l'étiqueteur *TreeTagger* (Stein & Schmid, 1995). Dans l'avenir, cette annotation fera l'objet d'une extension[5] ainsi que d'une révision qui visera à augmenter la couverture des formes inconnues du catégoriseur (coquilles, néologismes, formes « non-standard »...).

À ce stade, nous avons eu pour préoccupation majeure d'éviter, entre autres, qu'un même document revienne plusieurs fois. Gardons à l'esprit, cependant, qu'éliminer les redites est une décision qui altère dans une certaine mesure les données originales. En élaborant le CRFC, nous avons pris soin d'éviter à tout prix les doublons d'un texte (celui-ci pouvant, évidemment, se réduire au mot unique d'un message dans le cas des forums de discussion sur la toile) dans son intégralité. Néanmoins, on a gardé les répétitions si elles étaient considérées comme un élément normal de la production du discours dans un domaine particulier. On retiendra en guise d'illustration :

1. l'intratextualité : on a remarqué, par exemple, que le même fragment de texte sur la Croix Rouge revenait deux fois dans un livre sur le vieillissement, d'abord dans le chapitre consacré à « la recherche d'une maison de retraite », puis dans celui sur comment « rédiger son testament et faire une donation ».
2. l'intertextualité : on a pu constater qu'un court extrait tiré d'un livre sur la poésie, reproduit dans la partie du corpus réservée aux productions universitaires, figurait également dans une encyclopédie citée dans la partie sur les textes non-universitaires.
3. l'autocitation : répandue dans le journalisme et les encyclopédies, cette pratique consiste à réutiliser des formulations ou de courtes parties de textes pour traiter de sujets similaires ; de même, on a remarqué que les politiciens se répètent facilement d'un discours à l'autre.
4. duplication appartenant au format même : certains formats audiovisuels impliquent la redondance en leur sein (par exemple les bandes-annonces pour les feuilletons et les débats télévisés qui résument les épisodes précédents ou donnent un aperçu de celui que l'on va voir)
5. duplication fondée sur les exemplaires : on peut observer une répétition massive des mêmes tournures dans des cadres extrêmement conventionnels (par exemple, les formules de salutation et de politesse au début et la fin des lettres ou des courriels ; les salutations dans la conversation)

Nous avons considéré que tous ces types de redites font partie intégrante de la façon dans les être humains créent du discours et nous les avons donc conservés. En revanche, nous avons décidé de ne pas retenir ou de supprimer du corpus, à un stade ultérieur, les doublons relevant des cas de figure suivants :

6. reprise intégrale de textes
7. reprise partielle de textes (sauf exceptions mentionnées plus haut)
8. redite à l'intérieur d'un document dans le cas de fils de discussion sur un forum proposant une fonction « citation »

On a écarté ce dernier type de répétition en utilisant un logiciel de « web scraping »[6] qui nous a permis d'identifier avec précision les éléments de contenu à extraire d'une page web particulière et toutes les autres pages web de teneur similaire ; nous n'avons pas conservé les contenus repris. Il est possible de trouver un très petit nombre de redites, étant donné que quelques rares internautes reproduisent les citations dans leurs propres contributions sur les forums au lieu d'utiliser la fonction « citer ».

3 Conclusion

Tout compte fait, nous espérons avoir démontré dans les sections précédentes l'intérêt que revêt le CRFC en tant que ressource susceptible d'ouvrir des pistes de recherche dont l'exploration systématique était jusqu'à présent plutôt hors de portée. Parmi ces pistes, on mentionnera surtout le traitement lexicographique exhaustif du français familier et la description lexico-grammaticale de la langue française dans la lignée d'une approche mise en œuvre par Biber et al. (1999) à l'égard de l'anglais. Certes, à l'instar de tous les corpus existants, le CRFC est le résultat d'un compromis entre faisabilité et rigueur désirable, ce qui, dans notre cas précis, vaut surtout pour le traitement assez (peut-être trop) rudimentaire auquel nous avons soumis la grande quantité de données orales intégrées au sein du corpus. Il n'en reste pas moins qu'à ce jour, il existe, pour le français, très peu de corpus massifs équilibrés et largement diversifiés par genres textuels : le CRFC est désormais là pour combler un peu plus cette lacune.

Références bibliographiques

Dictionnaires cités

Collins = Collins/Robert [French-English &] *English-French Dictionary* ed. by B.T. Atkins et al. Glasgow/London: Collins. 9th edition 2010.

DDF = Rey-Debove, J. (éd.). (1999). *Dictionnaire du français*. Le Robert & Cle International.

DAFLES = Verlinde, S., Selva, T., Bertels, A. & Binon, J. *Dictionnaire d'apprentissage du français langue étrangère ou seconde*, Leuven, Institut Interfacultaire des Langues vivantes. Internet : http://ilt.kuleuven.be/blf/search.php

HS Slang = Nicholson, K. & Pilard, G. (2012). Harrap's Slang. *Dictionnaire d'argot et d'anglais familier*. Paris: Larousse.

LHF = Langenscheidt-Redaktion (2010). *Langenscheidts Handwörterbuch Deutsch-Französisch/Französisch-Deutsch*. Berlin: Langenscheidt.

LWUF = Meißner, F. J. (1992). *Langenscheidts Wörterbuch der französischen Umgangssprache*, Berlin : Langenscheidt.

HS = Stevenson, A. (2007). *Harrap's Unabridged PRO French/English-English/French*. Laurier Books Ltd.

PONS = *PONS Großwörterbuch Französisch: Deutsch-Französisch/Französisch-Deutsch*. Stuttgart: PONS.

PR = Robert, P. Rey-Debove, J. Rey, A. (éds.). (2008). *Nouveau Petit Robert: dictionnaire alphabétique et analogique de la langue française*. Paris : Le Robert.

TLF = TLF: *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789–1960)*, publié sous la direction de Paul Imbs. Paris: Éditions du Centre National de la Recherche Scientifique 1971-1994. Internet : http://atilf.atilf.fr

Ouvrages cités

Abeillé, A. & Barrier, N. (2004). *Enriching a french treebank.* In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Atkins, B. T. S., Clear, J. & Ostler, N. (1992). Corpus Design Criteria. *Journal of Literary and Linguistic Computing, 7(1)*, 1-16.

Beacco, J., Bouquet, S. & Porquier, R. (2004). *Niveau B2 pour le Français (utilisateur, apprenant indépendant) - un référentiel*. Paris: Didier.

Benzitoun, C., Fort, K. & Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2: TALN, 99-112.

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Blanche-Benveniste, C. (1991). Analyses grammaticales dans l'étude de la langue parlée. In Dausendschön-Gay, U, Gülich, E. & Krafft, U. (éds.), *Linguistische Interaktionsanalysen*, Tübingen: Niemeyer, 1-18.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français.* Gap-Paris: Ophrys.

Blanche-Benveniste, C. (2010). *Approches de la langue parlée en français.* Gap-Paris: Ophrys.

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Brosens, V. (1999). ELICOP, Etude LInguistique de la COmmunication Parlée : Bilan et perspectives. In *Actes du Colloque TALN'99*, Cargèse, Corse, 12-17 juillet 1999, *Corpus et TAL : Pour une réflexion méthodologique* (Atelier Thématique), 15-25.

Candito, M.-H., Nivre, J., Denis, P. & Henestroza Anguiano, E. (2010). Benchmarking of Statistical Dependency Parsers for French. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Cappeau, P. & Magali, S. (2005). *Les corpus oraux en français* (inventaire 2005 v.1.1). Internet : http://www.culture.gouv.fr/ culture/dglf/recherche/corpus_parole/ Inventaire.pdf.

Cappeau, P. & Gadet, F. (2007). Où en sont les corpus sur les français parlés. *Revue française de linguistique appliquée, 12(1)*, 129-133.

Carroll, J. B., Davies, P. & Richman, B. (1971). *The American heritage word frequency book*. Boston: Houghton Mifflin.

Corbin, P. (2005). Des occurrences discursives aux contextualisations dictionnairiques. Eléments d'une recherche en cours sur l'expression en français d'expériences du football. In Heinz, M. (éd.), *L'exemple lexicographique dans les dictionnaires français contemporains*, Tübingen: Niemeyer, 343-356.

Cresti, E. & Moneglia, M. (éds.). (2005). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins.

Debaisieux, J.-M. (2010). *Corpus Oraux – Problèmes méthodologiques de recueil et d'analyse de données*. Nancy: Presses Universitaires de Nancy.

Delais, E. & Durand, J. (éds.). (2003). *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse: Presses Universitaires du Mirail.

Denis, P. & Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*. Hong Kong, China.

Deulofeu, H.-J.& Debaisieux, J.-M. (2012). Une tâche à accomplir pour la linguistique française du XXIe siècle : élaborer une grammaire des usages du français. *Langue française, 176*, 27-46.

Diwersy, S. (2013-). *Varitext - Plateforme d'analyse lexico-statistique de variétés linguistiques*. Université de Cologne, Institut des Langues Romanes / Praxiling, UMR 5267, Université Paul-Valéry Montpellier 3. [Accessible en ligne à l'adresse http://syrah.uni-koeln.de/varitext/]

Equipe DELIC (2004). Présentation du *Corpus de référence du français parlé. Recherches sur le français parlé, 18*, 11-42.

Eshkol-Taravella, Iris et al. (2011). Un grand corpus oral „disponible" : le corpus d'Orléans 1968-2012. *TAL, 53/2*, 17-46.

Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik, 54(2)*, 177-190.

Falaise, A. (2005). Constitution d'un corpus de français tchaté. In *Actes de Récital 2005, 6-10 juin, Dourdan, France*. Internet : http://pro.aiakide.net/publis/2005RECITALPaper-Falaise.pdf

Ferraresi, A., Bernardini, S., Picci, B. & Baroni, M. (2010). Web Corpora for Bilingual Lexicography: A Pilot Study of English/ French Collocation Extraction and Translation. In Xiao, R. (éd.), *Using Corpora in Contrastive and Translation Studies*, Newcastle: Cambridge Scholars Publishing, 337-362.

Gadet, F. et al. (2012). CIEL_F: choix épistémologiques et réalisations empiriques d'un grand corpus de français parlé. *Revue Française de Linguistique Appliquée, 17(1)*, 39-54.

Gougenheim, G., Michea, R., Rivenc, P. & Sauvageot, A. (1956). *L'Elaboration du Français élementaire*. Paris: Didier.

Hoey, M. (2005). *Lexical priming: a new Theory of Words and Language*. London: Routledge.

Ide, N., Suderman, K. & Simms, B. (2010). ANC2Go: A Web Application for Customized Corpus Creation. In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC 2010)*. Valletta, Malta.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Harlow: Longman.

Koch, P. & Oesterreicher, W. (2011). *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Berlin e.a. : de Gruyter.

Levshina, Natalia. (à paraître) Subtitles as a Corpus: An n-gram approach. *Corpora*. Version provisoire accessible sur http://www.natalialevshina.com/articles/Levshina_SubtitlesAsCorpus.pdf

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Meißner, F.-J. (2006). Linguistische und didaktische Überlegungen zur Entwicklung von Kompetenzaufgaben im Lernbereich Mündlichkeit (Schwerpunkt Hörverstehen). *französisch heute, 37*, 240-282.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M. & Verine, B. (2013). Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. *Épistémè - revue internationale de sciences sociales appliquées, 9*, 107-138.

Nivre, J., Hall, J., Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Quaglio, P. (2009). *Television dialogue: the sitcom Friends vs. natural conversation.* Amsterdam: Benjamins.

Renouf, A. & Sinclair, J. (1991). Collocational Frameworks in English. Ajimer, K. & Altenberg, B. (éds.), *English Corpus Linguistics*, Cambridge: Cambridge University Press, 128-143.

Serpollet, N., Bergounioux, G., Chesneau, A. & Walter, R. (2007). A large reference corpus for spoken French: ESLO 1 and 2 and its variations. In Davies, M., Rayson, P., Hunston S. & Danielsson, P. (éds.), *Proceedings of the Corpus Linguistics Conference 2007*. University of Birmingham, UK.

Siepmann, D. (2007). Collocations and examples: their relationship and treatment in a new corpus-based learner's dictionary. *Zeitschrift für Anglistik und Amerikanistik, 3/2007*, 235-260.

Siepmann, D. (2015). Dictionaries and spoken language: a corpus-based review of French dictionaries. *International Journal of Lexicography, 2015/2*, 139-168. doi: 10.1093/ijl/ecv006.

Siepmann, D. & Bürgel, C. (2015). L'élaboration d'une grammaire pédagogique à partir de corpus : l'exemple du subjonctif. Tinnefeld, T. (éds.), *Grammatikographie und Didaktische Grammatik – gestern, heute, morgen.* Saarbrücker Schriften zur Linguistik und Fremdsprachendidaktik (SSLF), Saarbrücken: htw saar. Internet : http://grammatikographie.blogspot.fr/search/ label/4%20Siepmann.

Siepmann, D. & Bürgel, C. (à paraître). Les unités phraséologiques fondamentales du français contemporain. In Kauffer, M. & Keromnes, Y. (éds.), *Approches théoriques et empiriques en phraséologie*, Stauffenberg: Tübingen.

Siepmann, D., Bürgel, C., & Diwersy, S. (2015). The Corpus de référence du français contemporain (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography*, doi: 10.1093/ijl/ecv043. Internet : http://ijl.oxfordjournals.org/content/early/2015/12/23/ijl.ecv043.abstract.

Sinclair, J. (1987). *Looking up: an account of the COBUILD Project in Lexical Computing*. Birmingham: Collins COBUILD.

Stein, A. & Schmid, H. (1995). Etiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues, 36 (1-2)*, 23-35.

Summers, D. (1991). *Longman/Lancaster English Language Corpus. Criteria and Design*. Manuscrit inédit.

Summers, D. (1993). Longman/Lancaster English Language Corpus. Criteria and Design. *International Journal of Lexicography, 6 (3)*, 181-208.

TEI Consortium (éds.). (2015). *TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 2.9.1, 2015-10-15*. TEI Consortium. Internet : http://www.tei-c.org/Guidelines/P5/ ([Date of access]).

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Urieli, A. & Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*. Les Sables d'Olonne, France.

[1] Comme dans le cas du COCA, et à la différence d'un corpus de référence librement téléchargeable comme le *Open American National Corpus* (OANC ; cf. Ide et al., 2010), les échantillons qui constituent le CRFC ne seront accessibles qu'à travers une interface web, la diffusion du corpus par téléchargement n'étant pas prévue pour l'instant.

[2] La valeur de fréquence relative nommée « Relative Text Type Frequency » dans le cadre de *SketchEngine* se calcule en divisant la fréquence relative du résultat de la requête par la taille relative de la section en question (supposons que le mot « test » apparaisse 2000 fois dans le corpus, dont 400 fois dans la section « divers », qui représente 10 pour cent du corpus ; sa fréquence relative serait (400/2000)/0.1 = 200 pour cent).

[3] Sur le sens et l'utilisation des « boilerplates », voir http://blocnotes.iergo.fr/concevoir/les-outils/a-la-decouverte-dun-boilerplate/

[4] Cette simplicité se justifie dans la mesure où le schéma en question est censé s'appliquer avant tout au niveau global d'un corpus constitué par une grande variété de textes, qui, de par leur appartenance à des genres différents, se caractérisent par une structuration interne fort divergente. Ceci dit, il n'est bien sûr pas exclu de prévoir, dans l'avenir, un schéma plus complexe (et conforme aux diverses consignes de la TEI (TEI consortium, 2015)), au fur et à mesure qu'il s'avérera nécessaire de restreindre l'exploitation du corpus à l'analyse d'un genre textuel spécifique.

[5] Comme il est prévu d'annoter les textes au niveau des relations de dépandance syntaxique, nous testons en ce moment l'utilisation d'autres outils comme *Talismane* (Urieli & Tanguy, 2013), *mate-tools* (Bohnet, 2010) ou encore la chaîne de traitement *bonsai* (Candito et al., 2010), qui intègre l'étiqueteur morpho-syntaxique *MElt* (Denis & Sagot, 2009) et l'analyseur *MaltParser* (Nivre et al., 2006). A la différence de TreeTagger, les modèles d'annotation mis en œuvre par ces outils ont tous été générés au moyen de ressources dérivées du corpus arboré French Treebank (Abeillé & Barrier, 2004).

[6] Il s'agit du logiciel *Virtual Web Ripper 2.0* .