Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

Description sur corpus
Quelques réflexions autour des données et des instruments pour le français (parlé) à travers la description de *cause* et *causer*

Christophe Benzitoun

ATILF, Nancy Université & CNRS
Christophe.Benzitoun@atilf.fr

Paul Cappeau

FoReLL, Université de Poitiers & CNRS
Paul.Cappeau@univ-poitiers.fr

1 Introduction

Dans l'étude de nombreux phénomènes grammaticaux, les descriptions qui prennent appui sur des données orales ont souvent permis d'apporter des éclairages inédits ou complémentaires (Blanche-Benveniste, 1986 sur le passif, Deulofeu, 1986 sur *que*, Miller & Weinert, 1998, etc.). Si l'on considère que la description de la langue nécessite de tenir compte de données orales tout autant qu'écrites, se pose alors la question de la part à réserver aux corpus oraux et du contenu qu'ils doivent avoir. L'exemple de Biber et alii (1999) qui constitue à ce jour le travail le plus abouti sur corpus montre un fort déséquilibre en faveur des corpus écrits[1]. De même, Sinclair (1991) et Renouf (1984) avancent les chiffres de 6 millions de mots d'écrit et 1,3 millions de mots d'oral dans le cadre du projet COBUILD. Cette situation tient pour partie à l'énergie et au temps que la collecte et la transcription de données orales engloutissent. De fait, on peut considérer comme souhaitable un meilleur équilibre entre productions orales et écrites. Or, à l'heure actuelle, pour disposer d'un corpus de français parlé 1) équilibré, 2) diversifié, 3) échantillonné (ou au moins comportant des parties documentées) et 4) de taille significative (d'au moins deux millions de mots), il semble difficile de travailler uniquement à partir de données homogènes réalisées par une même équipe, pour un même projet et si possible à la constitution desquelles l'utilisateur a participé. En témoigne, la création récente, par le CNRS, des deux Centres de Ressources pour la Description de l'Oral (CRDO) servant de plateformes de dépôt et d'archivage. Pourtant, cela aurait le grand avantage de faire reposer les descriptions sur des ressources connues, homogènes et surtout ayant été élaborées dans un objectif précis, gages d'un travail rigoureux sur corpus.

Du coup, si l'on souhaite mener des descriptions grammaticales et/ou lexicales sur le français parlé, il nous faut nous diriger vers une solution alternative. Celle-ci consiste à voir comment s'accommoder au mieux des données orales disponibles souvent hétérogènes et éclatées (sur des sites différents)[2] et à recourir plus à une archive qu'à un réel corpus (Rastier, 2005 ; McEnery et alii, 2006) en les concaténant. En effet, les transcriptions récupérables actuellement possèdent des formats extrêmement variés (texte brut, fichiers au format *praat* ou *transcriber*, etc.), suivent des conventions de transcription pas nécessairement compatibles entre elles et ont un contenu diversifié et pas toujours bien documenté (les méta-données vont de l'absence totale à des tableaux en passant par des fiches en PDF). Nous nous proposons, dans le présent article, de mener une étude linguistique portant sur des regroupements de ce type afin de tester concrètement si l'on rencontre des problèmes insurmontables et dans quelle mesure les résultats obtenus sont exploitables. L'étude en question portera sur la description des lexèmes *cause* et *causer* à l'interface entre la grammaire et le lexique. Ces deux unités ont été choisies pour leur diversité d'emplois, parfaitement adaptée à la phase expérimentale dans laquelle nous sommes actuellement.

En plus de la réflexion sur les données, nous poursuivons l'objectif plus ambitieux de nous doter des bases indispensables à des études sur corpus à grande échelle, ce qui inclut bien évidemment l'établissement d'une méthodologie précise. De nos jours, cette réflexion passe nécessairement par un état

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

des lieux des instruments existants et par une analyse approfondie des fonctionnalités que devraient avoir des logiciels d'aide à la description linguistique (notamment ceux permettant l'interrogation de corpus, tels les concordanciers). Le constat qui s'impose c'est qu'il existe finalement peu d'outils répondant aux besoins réels que la recherche dans ce domaine met à jour, ce qui oblige à disposer d'un savoir-faire non trivial dans le domaine de l'informatique.

Ce papier est en définitive une contribution à la réflexion autour du linguiste à l'instrument lancée par Habert (2005). Cette question est centrale car de nombreuses contraintes pèsent sur la linguistique française sur corpus, notamment au niveau de l'existence et de la disponibilité des données et des outils. Ces deux facteurs sont pourtant indispensables à l'émergence d'une discipline outillée, cumulative et structurée.

Dans un premier temps, nous présentons les données sur lesquelles nous basons notre étude. Puis nous exposons la démarche que nous avons suivie, avant de détailler l'ensemble des paramètres pertinents dans le cadre de la description de *cause / causer*. L'exposé de ces paramètres est important car il donne des indications sur les traitements que pourraient effectuer des logiciels d'aide à la description linguistique. Ensuite, nous évoquons le problème de la présentation des résultats et son incidence sur la mise en lumière des phénomènes pertinents. Et pour finir, nous détaillons quelques-unes des fonctionnalités qui nous auraient été utiles dans le cadre du présent travail.

Nous tenons à souligner que notre démarche est essentiellement à visée théorique et méthodologique. Les descriptions de *cause* et *causer* ne sont là que pour insérer la réflexion dans une étude concrète. Nous avertissons donc le lecteur qu'il ne trouvera pas, dans les pages qui suivent, une monographie finalisée de ces deux lexèmes. Tel n'est pas l'objectif poursuivi ici.

2 Les données

2.1 Présentation générale

Nous présentons ici les choix que nous avons faits dans la récupération « opportuniste » des données plus ou moins disponibles, pleinement conscients des biais qu'induit notre démarche. Mais c'est justement l'influence de ces biais sur les résultats que nous souhaitons évaluer.

En premier lieu, bien que notre étude privilégie le français parlé, il nous a semblé indispensable d'inclure du français écrit, et ce pour au moins trois raisons. Tout d'abord, il n'y a pas lieu de défendre une description syntaxique qui ne porterait que sur l'oral, ce serait amputer la double dimension de la langue. Ensuite, il est nécessaire d'avoir une base de comparaison pour éclairer les résultats. Enfin, la limitation actuelle des données orales rend le recours à l'écrit incontournable. Nous avons donc réuni, en complément de l'archive orale, un corpus de textes écrits échantillonné et équilibré (100.000 mots pour chaque tranche) émanant du *Corpus Evolutif de Référence du Français*[3] (CERF).

En outre, trois paramètres principaux ont guidé l'élaboration de cette archive : la taille minimale requise, l'équilibre et la diversité. Concernant la taille, l'étude de *cause / causer* nécessite une quantité importante de données. En guise d'illustration, signalons que l'archive orale *Corpaix*[4], qui fait environ 1.000.000 mots, contient seulement 5 occurrences du verbe *causer*. De même, la partie du CERF comportant des textes journalistiques et faisant la même taille, en contient seulement 38. Certes, la taille de la partie écrite peut être largement augmentée, mais la partie orale est loin d'être extensible à volonté. Dans ces conditions, certains détails n'ont pas pu être observés ou approfondis, du moins à l'oral. On peut donc déjà en déduire que les études pouvant être réalisées sur ces données seront nécessairement limitées aux unités et phénomènes ayant une fréquence relativement élevée.

Pour ce qui est de la question de l'échantillonnage des données et de leur équilibre, les problèmes rencontrés sont encore plus importants. En effet, il n'existe pas, à l'heure actuelle, de vaste corpus (i.e. de plusieurs millions de mots) de français parlé échantillonné et équilibré. De plus, l'établissement de critères pour effectuer des regroupements pertinents représente un secteur de recherche en soi

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

particulièrement peu exploré pour l'oral, et encore largement débattu pour l'écrit (cf., parmi d'autres, Biber, 1988 ; Malrieu & Rastier, 2001). Il nous a donc été impossible de procéder à un échantillonnage de manière systématique. Cette situation tendrait donc à nous priver provisoirement de la possibilité de faire ressortir des symétries et des disparités de fonctionnements ou d'aborder la question de l'incidence des genres à l'oral, du moins en apparence.

En fait, dans le contexte actuel, nous considérons que la typologie des textes doit aller de pair avec les analyses. Les regroupements émaneront des proximités de fonctionnements linguistiques observées issues d'un grand nombre de travaux descriptifs. Il s'agit donc d'un des résultats attendus de nos analyses et non d'un préalable. Concrètement, nous comptons mettre en relation des phénomènes linguistiques convergents et des types de données, comme l'ont fait par exemple Bilger et Cappeau (2004) et Biber (1988).

Le travail à partir du *Corpus de Référence du Français Parlé*[5] (CRFP) a tout de même permis d'observer l'influence des situations de parole, mais ces données restent de taille trop modeste pour permettre des analyses détaillées (quelques dizaines de milliers de mots). Nous avons également pu compléter la description grâce à des données supplémentaires homogènes provenant de la collection personnelle de P. Cappeau. Mais là encore elles ont une dimension extrêmement réduite, ce qui limite grandement la possibilité d'observer des phénomènes autres que massifs lorsque l'on s'intéresse aux fréquences lexicales ou aux liens entre le lexique et la syntaxe.

Au final, nous avons choisi des proportions égales entre oral et écrit. Cette décision va à l'encontre des choix faits pour le *British National Corpus*, par exemple, mais il nous semblait opportun de ne pas ajouter une variable supplémentaire (le déséquilibre) aux autres problèmes déjà soulevés. Il a donc fallu nous contenter d'une archive réduite, à savoir environ 4.000.000 mots (oral et écrit confondus). Cette archive nous sert uniquement de base de comparaison. Nous sommes extrêmement prudents sur les résultats obtenus à partir de ces textes provenant de sources variées, notamment pour ce qui est de l'oral (cf. Cappeau & Gadet, 2007 sur le lien entre constitution du corpus et visée de l'étude). Il faut être d'autant plus prudent que ces données présentent une diversité importante. Pour la partie orale, nous avons affaire à du français majoritairement non planifié, mais dont nous ne connaissons que partiellement le contenu (il serait de toute façon bien naïf de croire que l'on peut cerner précisément le contenu d'un enregistrement un peu long, quel qu'il soit) et assez mal les proportions de chaque « genre » (ce qui supposerait résolues des questions encore en débat : Adam & Heidmann, 2006). Tant qu'un vaste corpus, même imparfait, de français équilibré et échantillonné n'existera pas, nous serons condamnés à travailler avec une variable d'imprécision difficile à quantifier et à pondérer. Nous insistons particulièrement sur le problème que représente l'absence d'un « corpus de référence » du français car il s'agit d'une question récurrente dans les travaux contemporains en linguistique française.

Ainsi, à l'heure actuelle, l'urgence est selon nous de disposer de données stables à partir desquelles nous pourrons mener des descriptions. En conséquence, nous réfléchissons à la manière d'uniformiser les formats et d'aborder la question de la diversité des conventions de transcription. Ces points doivent être traités précisément si l'on souhaite pouvoir tirer profit de l'ensemble des ressources mises à disposition et ne pas donner des résultats comportant trop de bruit ou de silence dus à des conventions distinctes. Pour l'instant, notre intervention s'est limitée à l'ajout de balises afin de délimiter chaque transcription. On sait donc à tout moment de quels fichiers sont issus les exemples et il est possible de revenir systématiquement à la transcription source, si jamais un fonctionnement tout à fait singulier est repéré ou lorsque l'on soupçonne une erreur de transcription. A terme, la vérification à partir du fichier son devrait d'ailleurs être possible pour plus de la moitié des données.

2.2 Références

Pour l'oral, notre archive est pour l'instant composée du *Corpus de Français Parlé Parisien* (CFPP2000), de *Corpaix* dans sa version de mai 2000, du *Corpus de Référence du Français Parlé*, et d'une partie de la base *Phonologie du Français Contemporain* (PFC). Il s'agit d'une archive ouverte qui sera enrichie au fur et à mesure que de nouvelles données seront disponibles, notamment celles en cours de constitution au

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

sein du laboratoire ATILF. Plusieurs paramètres ont été pris en compte dans ce choix : données à notre disposition, libres pour certaines, le plus possible composées de productions non planifiées, productions d'adultes ou d'adolescents et dans une aire géographique limitée, à savoir la France. Ainsi, dans PFC, nous n'avons récupéré que les discussions libres et seulement de locuteurs français. Toutefois, des transcriptions ne respectant pas certains de ces critères se trouvent forcément dans l'archive, mais nous espérons les repérer dans le cadre des descriptions que nous ferons. En effet, nous émettons l'hypothèse que ces transcriptions devraient se singulariser du point de vue linguistique. Par la même occasion, cette incertitude subie quant au contenu peut s'avérer utile dans le cadre de recherches sur la typologie des données.

Nous avons également eu recours, en plus de cette archive constituée dans un but comparatif avec l'écrit, à des données orales complémentaires plus homogènes. Il s'agit de corpus d'une taille comprise entre 12.000 et 70.000 mots qui sont souvent "thématiques" (des récits de films, des recettes, des visites guidées, des locuteurs parlant de la langue, des émissions judiciaires, etc.). Ces corpus sont identifiés par la côte P+2chiffres.

Pour l'écrit, nous avons aussi opté pour un empan assez large. Pour le constituer, nous avons sélectionné des tranches dans le *Corpus Evolutif de Référence du Français*, à savoir :

- De la presse : Courrier international, Le Monde, Le Monde diplomatique, le Nouvel Observateur ;
- Des discours politiques : J. Chirac, L. Jospin, F. Mitterrand ;
- Des textes scientifiques : ouvrages parus aux éditions du CNRS, revues publiées chez Hermès, Pour la science, Sciences et avenir ;
- Des textes institutionnels : Assemblée Nationale, textes juridiques et législatifs ;
- Divers autres textes : Philosophie, critiques littéraires, critiques cinéma, nouvelles de science fiction, romans.

Là encore, nous nous autorisons la consultation de ressources écrites complémentaires pour vérifier des hypothèses. Il faut donc concevoir l'ensemble présenté ici comme notre ressource principale mais pas exclusive.

3 La démarche

Etant donné qu'à notre connaissance il n'existe pas d'outil informatique capable d'effectuer automatiquement des descriptions linguistiques, il a fallu tirer parti des logiciels à notre disposition. Pour l'instant, nous sommes obligés de recourir à un traitement majoritairement manuel, même si certains logiciels facilitent grandement la recherche des occurrences et la manipulation des données. De toute façon, compte tenu du caractère exploratoire de nos recherches, un compromis entre le travail manuel et l'assistance des outils informatiques était incontournable afin de :

- Contrôler au mieux les différentes étapes menant à l'obtention des résultats ;
- Pouvoir faire émerger les besoins en logiciels ;
- Avoir la possibilité d'observer l'ensemble des phénomènes intéressants.

Par ailleurs, nous nous situons plutôt dans le cadre d'une DAO (Description linguistique Assistée par Ordinateur) et non dans celui d'une DA (Description linguistique Automatique). En effet, la plupart des traitements automatiques font intervenir des théories linguistiques ayant la malencontreuse conséquence de transformer en profondeur les données primaires et de générer des erreurs. L'étiquetage automatique en parties du discours est emblématique de cette situation. Les informations additionnelles doivent donc être manipulées avec précaution dans une approche guidée par les données (corpus driven), dont l'objectif n'est pas de vérifier une hypothèse préexistante. Dans ce type d'approche, moins les données sont transformées, moins elles seront orientées par une théorie extérieure à l'analyste.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

Pour effectuer les requêtes, nous avons utilisé le concordancier *Contextes*[6] développé par J. Véronis. Il s'agit d'un logiciel disposant des fonctions de base de la plupart des concordanciers et qui est bien adapté à notre étude. Ainsi, nous avons recherché tous les « mots » débutant par les caractères `caus`. La restriction aux cas où il s'agissait du nom *cause* ou du verbe *causer* a nécessité une phase extrêmement rapide de filtrage manuel. Et mise à part la forme *cause*[7], il n'y a pas d'ambiguïté entre le verbe et le nom. Les deux emplois sont donc, le plus souvent, aisés à distinguer.

Certes, ces résultats auraient pu être obtenus directement en effectuant des requêtes sur des textes préalablement lemmatisés et étiquetés en parties du discours. Mais pour les raisons évoquées ci-dessus, nous ne souhaitions pas y recourir pour l'instant. De plus, le logiciel dont nous disposions pour effectuer les recherches dans des fichiers « enrichis » (à savoir *LoX*[8]) ne permettait pas de conserver les divisions dans lesquelles les occurrences ont été retrouvées, information fondamentale à nos yeux[9]. Cela met en lumière une fois de plus la question du format informatique des données, qui nécessite des transformations systématiques pour les rendre compatibles avec un logiciel particulier. D'où l'intérêt d'unifier le format initial pour faciliter ces transformations. Il faut également bien maîtriser les formats d'entrée des logiciels pour pouvoir les utiliser et limiter les risques d'erreurs.

Les concordances de *cause* et de *causer* ont ensuite été exportées dans le tableur *Excel* afin d'être manipulées à l'aide des fonctionnalités de tris et de filtres que propose ce logiciel. Elles sont présentées sous la forme suivante (colonnes de gauche à droite) :

- Les deux premières colonnes stipulent les références de l'occurrence ;
- Les trois suivantes, le contexte gauche, la cible ainsi que le contexte droit ;
- La dernière contient une clé de tri gauche afin de pouvoir trier en commençant par le premier mot directement à gauche de la cible.

Afin de pouvoir les exploiter plus aisément, nous avons ajouté une colonne dans laquelle nous avons apposé manuellement des étiquettes correspondant aux paramètres linguistiques décrits ci-dessous. Le principe que nous avons suivi était de prendre en compte, à chaque étape, un paramètre supplémentaire et ainsi de parvenir à des résultats de plus en plus détaillés par petites touches successives.

4 Les paramètres de l'analyse

Dans cette partie sont présentés les paramètres pouvant, d'une façon ou d'une autre, avoir une influence sur le fonctionnement des formes *cause / causer*. Ils sont classés en deux rubriques principales : 1) les paramètres linguistiques qui tiennent compte des propriétés des éléments recensés (catégorie, particularités syntaxiques, etc.) et 2) les paramètres liés aux caractéristiques des données retenus pour l'étude. Pour exploiter ces paramètres, nous avons utilisé quelques-unes des fonctionnalités d'*Excel* (tris et filtres) et procéder à un travail manuel conséquent, notamment pour apposer les étiquettes sur chacun de nos exemples.

Dans cette partie, notre objectif est d'énumérer quelques-uns des paramètres pertinents pour une description distributionnelle sur corpus. Il s'agit de mettre en lumière certains phénomènes observés par les linguistes afin de lancer la rédaction d'un « cahier des charges » destiné à décrire les contours d'instruments informatisés d'assistance à la description.

4.1 Les paramètres linguistiques

Il importe en premier lieu de s'intéresser à la catégorie (ou partie du discours) à laquelle se rattache les formes *cause* et *causer* puis de voir, pour chacune d'elles, quels phénomènes mettent en lumière les corpus.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

4.1.1 Catégorie de l'élément et segmentation

Si les formes conjuguées à partir de *causer* sont en général assez facilement identifiables, il n'en est pas de même de la forme *cause* qui peut entrer dans de multiples environnements et être rattachée à des catégories distinctes. On peut ainsi recenser un nom (1), une préposition (2), un verbe (3) :

(1) *le premier respect que j'ai et la première euh* ***cause*** *de remerciements que je peux évoquer c'est envers mes parents* [CRFP]

(2) *ce que je vous expliquais tout à l'heure principalement* ***à cause du*** *style* [P96]

(3) *bon je ne* ***cause*** *plus la parole à toi tu te débrouilles avec ta feuille* [Corpaix]

En plus de *à cause de*, il y a de nombreux autres emplois locutionnels pour lesquels le rattachement à une catégorie se calcule en tenant compte du fonctionnement d'ensemble de la locution (la syntaxe externe du groupe et pas de *cause* tout seul) : *gain de cause* peut être rattaché au nom ; *mettre / remettre en cause* relèvent du verbe ; *pour cause de* peut être traité comme une préposition ; *en tout état de cause* et *en (toute) connaissance de cause* sont plutôt des adverbes. *Et pour cause,* quant à lui, semble proche d'un rôle adverbial ou d'une interjection.

Le choix de traiter des unités linguistiques comme des mots isolés ou comme des blocs indissociables fait donc partie de la tâche du linguiste. Or, avec des étiqueteurs morphosyntaxiques, le découpage en locutions et mots isolés est réalisé en amont par le logiciel, ce qui a une incidence directe sur l'analyse.

4.1.2 Autres paramètres en fonction de la catégorie

Pour chacune des trois catégories précédentes, il est possible de retenir un certain nombre de caractéristiques de construction.

4.1.2.1 Pour le nom

Parmi les variables que les données permettent de faire ressortir (en plus du sens), on peut noter :

a) Le nombre. Le décompte selon ce paramètre suppose une analyse qui va au-delà de la simple reconnaissance formelle. En effet, faire la différence entre *cause* et *causes* impose généralement de distinguer emploi nominal isolé et emploi locutionnel. Dans l'emploi isolé, les occurrences de *cause* au pluriel sont bien plus nombreuses que celles au singulier (31 contre 15 sur les corpus oraux retenus, incluant la partie additionnelle). Cette tendance ne correspond pas à celle que l'on observe pour les noms en général où le pluriel est moins attesté que le singulier. Ainsi, dans le CRFP, on peut opposer : 33 occurrences de *solution* contre 19 de *solutions* et 196 occurrences de *problème* contre 105 de *problèmes*. Elle illustre, en creux, l'importance que les locutions occupent dans la distribution de la forme *cause* et montre la nécessité d'un travail d'analyse approfondi puisque, si l'on regroupe toutes les occurrences (nom autonome et locution nominale) sans les distinguer, la tendance s'inverse largement.

b) La fonction syntaxique. La fonction syntaxique peut être observée à deux niveaux distincts. D'une part, la position que le nom occupe dans le syntagme nominal lui-même : est-il la tête du SN (*rechercher des causes*) ou est-il complément d'un autre nom (*un réseau de causes*) ? C'est l'emploi tête qui, à l'oral, domine (26 contre 8). D'autre part, quelle fonction remplit le SN avec *cause* par rapport à un verbe ? La fonction sujet (ou couplé au sujet) est assez peu utilisée (6 sujets contre 28 pour les autres fonctions). Il n'est pas sûr que cette variable soit très pertinente (elle s'éloigne trop peu, semble-t-il, des tendances générales observées pour les noms). La prise en compte de la forme du verbe (ou du nom) recteur peut également révéler des phénomènes intéressants.

c) L'émergence de locutions ou de collocations. Il s'agit de cerner l'unité lexicale de traitement, qui peut être soit une locution (cf. exemples ci-dessus), soit une collocation (ex. (4)), soit un mot autonome (ex. (5)) :

(4) ***Les relations de cause à effet*** *n'étaient pas toujours évidentes.* [Roman_C]

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

(5) *Encore convient-il de s'attaquer aux **causes** profondes de l'instabilité et du désordre…* [Mitterrand]

Les deux exemples ci-dessus ont été pris à dessein afin de montrer les frontières plus ou moins floues qui séparent ces trois types d'emplois. (4), bien qu'ayant l'apparence d'une locution, tolère des variations : *liens réels de cause à effet*, *relation directe de cause à effet*, *le problème de la cause et de l'effet*, etc. Il faudrait sans doute comptabiliser tous ces exemples ensemble et les considérer plutôt comme des collocations. Quant à (5), on a le sentiment d'un lien étroit entre *causes* et *profondes*, sans que l'on puisse réellement l'objectiver. Sur ce point, l'apport de très grands corpus pourrait être décisif grâce à la prise en compte de la fréquence (cf. 4.1.2.4 ci-dessous).

4.1.2.2 Pour le verbe

Pour le verbe, on peut retenir les variables suivantes :

a) Le sens. Il nous a paru pertinent de distinguer seulement deux sens : un sens proche de « parler » (6) et un autre proche de « provoquer » (7).

(6) *tout le monde se met à **causer** avec son petit voisin* [CRFP]

(7) *Mais je ne veux pas **causer** le malheur de ta vie.* [Roman_A]

b) Les variations temporelles. Elles sont assez peu représentées dans le corpus oral qui contient pour l'essentiel des emplois au présent et au passé composé. Le corpus écrit de littérature classique permet d'observer des attestations beaucoup plus variées : passé simple, imparfait, etc. soit une exploitation bien plus diversifiée du paradigme. Quelques formes temporelles peuvent alors être vues comme des marqueurs intéressants puisqu'elles permettent de relier la forme observée à des productions bien identifiées.

c) Le type de complément. Plusieurs schémas de construction sont attestés et recoupent des sens distincts :

- emploi intransitif (*causer* + Ø) : c'est le cas le plus fréquent lorsque le verbe a le sens de « parler » (*j'aime entendre causer* [P97])

- emploi transitif (*causer* + C1 (+ C2)) : Tout va dépendre du paradigme auquel se rattache le complément construit par le verbe. Quand c'est un complément de type *ainsi*, c'est le sens de *parler* qui est activé (*il cause pas bien*), quand c'est un complément direct, *causer* a le sens de *provoquer* (*les sports qui causent moins de problèmes*). Dans ce dernier cas, le verbe construit, à l'oral, presque systématiquement un complément indirect en *lui/me/te* (*si vous voulez **me** causer des problèmes*).

d) Les lexèmes se trouvant dans la position de complément. La prise en compte de cette variable revient souvent à faire une étude des collocations. Pour *causer*, il nous a d'ailleurs fallu utiliser un paramètre supplémentaire. Il s'agit de ce que Sinclair (2004) à la suite de Louw (1993) appelle la « prosodie sémantique » (*semantic prosody*), c'est-à-dire une propension qu'ont certaines unités lexicales à sélectionner des unités ayant un sens plutôt positif ou négatif. Dans le cas du verbe *causer* (au sens de « provoquer »), les compléments sélectionnés sont clairement négatifs (*préjudices, dégâts*, etc.)[10] dans la majorité des cas. En français, on cause donc majoritairement des événements négatifs. Mais il faut préciser que le travail sur la prosodie sémantique et les collocations doit se faire à partir de constructions syntaxiques identiques (ce que Stefanowitsch & Gries, 2003 nomment « collostruction »). En effet, la fréquence des collocatifs ne doit pas être calculée sur la totalité des occurrences mais seulement sur celles ayant le même patron syntaxique. Le mélange de tous les emplois de *causer* fausserait immanquablement le calcul des fréquences et pourrait même aller jusqu'à rendre invisibles des phénomènes pourtant cruciaux. Comme pour le nombre grammatical, un classement préalable est donc indispensable et il paraît difficile d'envisager une automatisation de cette tâche à l'heure actuelle, à moins de disposer de grands corpus annoter syntaxiquement.

e) Lexème de substitution. Il semble pertinent d'observer si la fréquence d'emploi de *causer* n'est pas corrélée à celle de verbes proches tels que *parler* et *provoquer*. Cela impose de travailler sur des

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

contextes probablement plus larges que ceux qu'offre par défaut un concordancier. (voir 4.2.2. c sur ce point).

4.1.2.3 Pour la préposition

Tout emploi confondu, c'est l'emploi *à cause de* qui est, de loin, le plus fréquent à l'oral. Deux paramètres peuvent être observés :

a) Les éléments mis en relation. Dans le contexte antérieur, c'est le verbe *être* qui se rencontre le plus régulièrement à l'oral (mais c'est une indication de faible portée puisque c'est aussi la forme verbale la plus utilisée). Pour le contexte postérieur, il semble difficile de dégager une régularité autre que le fait que l'on trouve souvent du lexique non humain (mais là encore difficile de déterminer si cette observation est pertinente).

b) La place du groupe prépositionnel. On rencontre un faible nombre d'exemples dans lesquels *à cause de* est en position frontale comme dans :

(8) *mais euh* ***à cause de cette phrase qui a été soulignée au rouge*** *ah bè dam {mots en patois} j'ai eu neuf* [P97]

4.1.2.4 Paramètres définitoires des locutions

La fréquence brute d'une suite de mots n'est pas toujours un critère pertinent pour déterminer si nous avons affaire à une locution. Cependant, elle peut faciliter leur repérage. Par exemple, *(re)mettre/(re)mise en cause* est extrêmement fréquent. Nos données en comportent 179 occurrences. Mais cette fréquence significative est liée en grande partie au fait que nous avons choisi de regrouper ces deux tournures, ce qui les rend plus visibles.

Pour *être en cause*, la fréquence est nettement moins importante. En effet, nous n'avons relevé que 20 occurrences et le verbe *être* est une unité très présente dans tous les corpus, ce qui rend très hypothétique la possibilité d'un repérage lié à une cooccurrence significative. Ainsi, dans nos futurs travaux et pour l'élaboration de notre classement final, il faudra prendre garde à ne pas restreindre les locutions et les collocations au seul critère de fréquence (même en tenant compte de la fréquence respective de chaque collocatifs). Mais il est possible qu'en changeant d'échelle et en travaillant sur des corpus de taille nettement plus importante, ce problème se pose différemment.

De plus, nous avons rencontré des difficultés pour assigner systématiquement des étiquettes grammaticales aux locutions. Même si on peut facilement considérer *(re)mettre en cause* comme un verbe et *(re)mise en cause* comme un nom, pour *en toute connaissance de cause*, la question est moins évidente : s'agit-il d'une locution adverbiale ?

De même, la délimitation exacte des locutions et la frontière avec les collocations ne sont pas des problèmes résolus. Pour *gain de cause*, par exemple, on peut considérer qu'il s'agit d'une locution nominale ne prenant jamais de déterminant. Mais faut-il intégrer également les verbes précédents qui se limitent à *avoir* et *obtenir* ? Et ces deux verbes font-ils partie de la locution ou s'agit-il de collocation ? Du coup, dans la présentation des résultats, faut-il distinguer les collocations des locutions ou doit-on les regrouper ?

4.2 Les facteurs liés aux types de données retenus

En plus des paramètres linguistiques, il faut tirer profit de la composition des corpus. La prise en compte du contenu ou de la situation de parole est pertinente dans le cadre de l'observation de la distribution des usages et elle a une forte influence sur les résultats.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

4.2.1 Différences selon la distinction écrit / oral

Dans un premier temps, on doit s'interroger sur les répartitions que l'on souhaite mettre en avant. Un comptage global, par exemple, cacherait des distinctions pourtant fondamentales. Pour *cause*, l'emploi majoritaire (écrit et oral confondus) est celui de nom isolé. Pourtant, on observe de grandes différences en fonction des types de textes allant de l'emploi très majoritaire de *à cause de* à celui de *(re)mettre en cause* en passant par l'emploi nominal.

Dès lors que l'on recourt à des ressources écrites et orales, il semble intéressant d'exploiter cette opposition. La forme étudiée se prête bien à une telle approche qui fait ressortir une différence quantitative importante en faveur de l'écrit. Ainsi, pour *causer* on trouve 82 occurrences à l'écrit contre 17 à l'oral (soit presque 5 fois moins) ; le déséquilibre se retrouve, de façon un peu moins forte, pour *cause* : 617 occurrences à l'écrit contre 176 à l'oral.

Mais cette présentation des résultats globaux fournit des indications d'un intérêt somme toute assez limité. En effet, il n'est pas rare que des différences de distribution soient observées entre écrit et oral. La description au niveau des formes seules et en prenant appui sur des corpus dans leur globalité (écrit *vs* oral) n'éclaire que peu l'analyse de *cause* / *causer*. A titre de comparaison, on trouve des différences quantitatives importantes dans les diverses tranches du corpus CERF pour le lexème *solution* : 14 pour la littérature classique et 275 pour les textes politiques. Il est donc nécessaire d'aller voir dans le détail à quoi sont dues ces variations de fréquences et de proposer des échantillons plus fins.

4.2.2 Différences selon les genres

Cette partie est la plus dépendante des corpus additionnels homogènes. La possibilité d'exploiter des corpus écrits et oraux dont le contenu est mieux cerné permet de présenter une répartition plus sensible aux genres (Biber, 1988 ; Rastier, 2005). Il devient alors possible de relier certains emplois des formes *cause* / *causer* à des situations codifiées des productions (tant orales qu'écrites), ce que la présentation globale masquait. Ce n'est plus la différence écrit / oral qui prime mais d'autres traits plus fins. Ainsi, nous avons observé les phénomènes suivants :

a) Les corpus juridiques (textes législatifs, audition d'une commission d'enquête, etc.) comportent en quantité des emplois qui se rencontrent peu dans les autres corpus : c'est le cas de *(re)mettre en cause* et de *causer* (au sens de « provoquer »). De même *en tout état de cause* est quasi absent des productions orales sauf lorsqu'il s'agit de corpus dans lesquels interviennent des avocats. Voilà certainement une locution qui relève de la langue technique et qui transcende clairement le clivage oral/écrit.

b) Autre phénomène remarquable : sur les 15 occurrences du verbe *causer* que comportent les textes juridiques et législatifs, 10 ont pour collocatif *préjudice*. L'intégralité des exemples de *causer un préjudice* est d'ailleurs dans cette tranche.

c) Il y a un suremploi du verbe *causer* au sens de « parler » dans la tranche des romans du XIX^{e}-début XX^{e} s. Sur les 15 occurrences du verbe, 11 ont ce sens-là. Et il est intéressant de souligner que dans notre portion contenant des romans contemporains, il n'y a aucun exemple de ce type et seulement deux dans les nouvelles contemporaines de science fiction. Il faudrait donc faire une étude complémentaire pour voir si *causer* (au sens de « parler ») n'est pas en perte de vitesse dans la littérature française contemporaine.

d) Une autre observation conduit à se demander si le thème abordé dans le corpus n'est pas un facteur sensible. C'est en tout cas cette variable-là qui semble pertinente quand on observe que le verbe *causer* au sens de « parler » se rencontre à l'oral lorsque les locuteurs parlent de la langue, des dialectes, de leur façon de parler, etc. Dans une des portions additionnelles faisant seulement 21.241 mots, il y a 13 occurrences du verbe *causer* au sens de « parler ». Cet emploi est quasi inexistant dans les autres données orales. Les deux emplois (*parler* et *causer* dans ce sens-là) se rencontrent conjointement dans les corpus. Ainsi, dans la tranche citée, on relève 230 occurrences de *parler*, alors que dans d'autres corpus plus volumineux, on rencontre une trentaine d'occurrences de ce verbe.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

e) Pour ce qui est de *causer* (au sens de « provoquer »), la notion de prosodie sémantique évoquée ci-dessus occupe une place de choix. En effet, dans la plupart des registres, les compléments de ce verbe ont une valeur clairement négative :

(9) *Il est vrai que tout autre schéma pourrait causer un* ***scandale*** *en France* [NouvelObs]

Mais les textes philosophiques se distinguent sur ce point en présentant uniquement des emplois « neutres » :

(10) *Bien entendu, ma croyance qu'il pleut cause le désir de ne pas être trempé* [Philo]

(11) *Ressentir de la douleur, par exemple, c'est se trouver dans un état mental interne pouvant résulter de certaines entrées sensorielles comme la sensation de brûlure (en posant par exemple la main sur une plaque chauffante) lequel, à son tour, cause ou engendre certaines actions ou comportements (retirer sa main en vitesse).* [Philo]

On peut d'ailleurs remarquer l'incongruité de l'usage neutre dans un autre registre :

(12) *À UN PEU PLUS d'un mois des championnats du monde, Stéphane Diagana a causé la surprise, mercredi 5 juillet à Lausanne, en battant le record d'Europe du 400 mètres haies en 47 s 37.* [LeMonde]

Dans ce contexte, on s'attendrait plus à l'utilisation d'un verbe tel *créer*. Et l'exemple oral ci-dessous présente une reformulation de *causer beaucoup de plaisir*, sans doute parce que le locuteur a saisi l'incongruité de la tournure employée :

(13) *L2 oui oui quand même oui est-ce que euh il y en a un que vous avez fabriqué et qui vous a causé euh comment dire euh beaucoup de plaisir enfin que vous avez fait avec plaisir* [CRFP]

La prosodie sémantique serait également à étudier pour le nom *cause* et la locution *à cause de* qui, sans forcément régir des lexèmes négatifs, imposent tout de même une saisie plutôt négative. Mais par manque de temps, nous n'avons pas encore pu nous atteler à cette tâche.

f) *A cause de* est particulièrement bien représenté à l'oral (121 occurrences sur les 176 que compte la forme *cause*). En fait, on observe que plus les procédés formels décrits ci-dessus sont présents, plus le nombre de *à cause de* est faible. C'est ainsi par exemple que dans les textes juridiques et législatifs on ne trouve que deux occurrences de *à cause de*. Du coup, on peut se demander par quelle tournure est remplacée *à cause de*, étant donné qu'il n'y a aucune raison de ne pas exprimer la notion de cause dans les textes juridiques et législatifs à l'aide d'une préposition causative. *En raison de* paraît être une alternative possible pour exprimer la notion de cause. Cependant, nous n'avons pas observé de corrélation évidente entre les fréquences de ces deux unités.

Par ailleurs, notre partie de corpus oral qui distingue paroles publique, professionnelle et privée présente une distribution intéressante. La partie privée (280.000 mots) comporte presque uniquement *à cause de* alors que la partie publique (80.000 mots) n'en comporte aucun, et seulement des noms autonomes et des *(re)mettre en cause*. Cependant, le manque d'exemples et le non équilibre entre les sections nous obligent à une très grande prudence.

5 Présentation des résultats

Dans cette partie, nous synthétisons les principaux résultats. Toutefois, il apparaît qu'aucune solution concernant la présentation des résultats n'est neutre ou indifférente et que cette étape doit faire l'objet d'une attention particulière. La réflexion sur ce point est pour l'instant très superficielle.

Sans pouvoir affirmer que nous avons épuisé le sujet, la plupart des phénomènes pertinents ont été relevés modulo les limites inhérentes à nos données. Il nous faut maintenant présenter les principaux résultats. Et à ce stade, il ne suffit pas de lister de manière exhaustive toutes les données que nous avons patiemment récoltées. Il faut au contraire les hiérarchiser et réfléchir à une présentation permettant de mettre en

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

lumière les phénomènes saillants. Pour cela, nous avions plusieurs options, toutes plus ou moins discutables.

La forme verbale ne pose pas de problème particulier, comme le montre le tableau suivant :

Causer 97	Oral 15	Ecrit 82
	Sens « provoquer » 9	Sens « provoquer » 67
	Connotation négative Se retrouve plutôt dans les corpus juridiques	Connotation négative sauf pour PHILO Fréquences remarquables : PHILO 16 JURILEGIS *causer* 15 (*préjudice* 10)
	Sens « parler » 6	Sens « parler » 15
	Particulièrement présent lorsque l'on parle de la langue	Fréquence remarquable : ROMAN (19^{e} - début 20^{e}) 11

Tableau 1. Principaux résultats pour *causer*

Mais il n'en va pas de même des autres emplois, pour lesquels on hésite entre la possibilité de les regrouper ou pas. Les questions centrales sont ici celles de savoir quels sont les éléments réellement comparables et quelle comparaison est la plus pertinente. Nous avons fait le choix de séparer chaque emploi en considérant qu'il s'agissait de locutions distinctes à opposer à un emploi libre du nom *cause*. Dans un premier temps, nous avions des résultats plus « spectaculaires » en mentionnant le pourcentage d'occurrences d'un emploi donné par rapport à l'ensemble des occurrences de la chaîne de caractères `cause` (avec ou sans -s). Ainsi, cela montrait une forte concentration dans certains types de textes. Mais en faisant cela, nous comparons des fonctionnements pas nécessairement comparables aussi directement. C'est d'ailleurs ce qui est fait dans d'autres travaux, sur la forme *contre* par exemple. Dans Bilger & Cappeau (2003), les auteurs mettent sur le même plan la proportion d'emplois de la préposition *contre* et de la locution *par contre*. Or, il s'agit de deux unités distinctes constituant vraisemblablement deux entrées séparées du lexique français comme *cause* et *à cause de*. Du coup, on doit se poser la question de savoir quelle conclusion tirer d'une telle comparaison. Que nous apprend une différence significative de pourcentage d'emploi entre *contre* et *par contre* ?

Dans le tableau ci-dessous, nous avons donc eu recours à un compromis. Nous présentons les fréquences brutes de chaque emploi ainsi que les fréquences remarquables dans certaines tranches. Nous mentionnons également la fréquence globale. Le lecteur pourra ainsi proposer son propre classement s'il le souhaite et disposera des mêmes informations que nous.

Cause 790	Oral 173	Ecrit 617
cause (nom isolé)	21	235 PHILO 53 CNRSEd 45
à cause de	121	90 ROMAN (20^{e}) 11 ROMAN (19^{e} - début 20^{e}) 9 Nouvelles Science fiction 5
(re)mettre/mise en cause	17 Plutôt dans les corpus juridiques	162 JURILEGIS 31
être en cause	1	19
le N en cause	0	39 JURILEGIS 27
en tout état de cause	4	33

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

	Corpus juridiques	
avoir/obtenir gain de cause	3	2
et pour cause	3 (uniquement dans corpus avocat)	5
pour cause de	1	6
en (toute) connaissance de cause	0	10
à cause que	1	1
de cause à effet	1	7 + *relation* (adj)
faire cause commune	0	3
en désespoir de cause	0	2
prendre fait et cause	0	2
hors de cause	0	1

Tableau 2. Principaux résultats pour *cause*

Cette présentation devra faire l'objet d'une réflexion plus approfondie, car en l'état, elle ne nous satisfait pas pleinement. Elle n'est, en effet, pas très « parlante » et mélange sans doute quelques paramètres inconciliables.

6 Pistes pour un instrument

En effectuant le présent travail, nous nous sommes rapidement aperçus que les outils informatiques dont nous disposions étaient très limités. Finalement très peu étaient adaptés à nos besoins. Pour l'anglais, il existe des outils permettant de travailler sur des corpus échantillonnés, dont les fonctionnalités vont de la sélection du corpus de travail au calcul de la fréquence des formes rencontrées dans chaque tranche en passant par l'obtention automatique des collocations et le tri aléatoire lorsque le nombre d'occurrences est trop important. *Sketch Engine*[11] et *BNCWeb*[12] en sont de bonnes illustrations. Ils sont généralement accompagnés d'un langage de requêtes très puissant (basé sur le *Corpus Query Processor* pour les deux logiciels précités). Pour le français, c'est *Stella* le moteur de recherche de *Frantext* qui se rapproche le plus à l'heure actuelle de tels outils. Mais on ne peut pas l'utiliser avec ses propres données.

Cependant, il est tout à fait possible d'utiliser directement le moteur sous-jacent au *BNCWeb* et à *Sketch Engine*, à savoir le *Corpus WorkBench*[13], et ainsi profiter de leur puissance. Mais l'absence totale d'interface pour guider l'utilisateur dans le dédale des fonctionnalités est assez déroutante et le langage utilisé le rend très difficile à maîtriser pour des linguistes peu férus d'informatique. Cela limite en tout cas grandement le nombre de ces utilisateurs potentiels. Pourtant, le *Corpus WorkBench* couvre au moins une partie des besoins listés ci-dessous, ce qui rend sa difficulté de prise en main d'autant plus dommageable.

Afin de combler ce manque, nous espérons que les quelques propositions faites ci-dessous concernant les fonctionnalités utiles émergeant de notre étude serviront à l'élaboration d'un logiciel d'interrogation de corpus performant pour le français. Une telle réflexion est actuellement menée au sein du laboratoire ATILF et devrait aboutir à la réalisation d'un instrument pour les linguistes.

En résumé, nous aurions besoin d'un logiciel qui :

- permette à l'utilisateur de sélectionner, dans un ensemble plus vaste, les données à partir desquels il souhaite travailler et donc d'élaborer ses propres sections ;
- calcule le nombre total d'occurrences du motif recherché ;

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

- permette de formuler des requêtes sur les lemmes (ce qui suppose que le fichier soit préalablement lemmatisé) ;
- signale la taille totale des données et le nombre de sections différentes sur lesquelles porte la requête (avec la taille de chacune de ces sections) ;
- propose une normalisation de la fréquence par n milliers (ou millions) de mots globalement et dans chaque section ;
- permette la visualisation de plusieurs sections à la fois pour essayer de mettre en évidence des similarités ou des fréquences significatives ;

  [Exemple : *causer* (sens « provoquer ») : 16 dans les textes philosophiques et 15 dans les textes juridiques et législatifs sur 76 (ou 67 uniquement pour l'écrit). Ces deux tranches représentent donc à elles seules près de la moitié des *causer* au sens de « provoquer » (à l'écrit).]
- renvoie le nombre de sections dans lesquelles le motif a été trouvé, afin d'avoir une idée de la dispersion du phénomène à décrire ;
- donne la localisation des occurrences recherchées dans chaque section ;
- donne la possibilité, lorsqu'il y a un déluge d'occurrences à traiter, de faire une sélection aléatoire et de trier les exemples de manière aléatoire (dans chaque section et de manière globale) ;
- permette d'avoir une vision rapide des principales collocations en fonction de divers algorithmes. En fait, il faut que l'on puisse visualiser les mots qui apparaissent souvent dans un entourage déterminé, et ce quelle que soit la position. Mais aussi pouvoir le faire uniquement sur le contexte droit ou sur le contexte gauche. Ce calcul doit pouvoir être effectué à partir des formes lemmatisées ;
- permette de tester des collocations à toutes les étapes ;
- permette de nous reporter au texte intégral à tout moment de l'analyse.

Une fois le travail de description fait par le biais d'étiquettes apposées manuellement ou semi-automatiquement, il faudrait :

- disposer de la fréquence de chaque étiquette mise par l'utilisateur et pouvoir la visualiser pour chaque texte ou regroupement de textes ;
- avoir les résultats couplant étiquettes et formes lemmatisées ;
- connaître la fréquence des étiquettes dans chaque section et dans chaque regroupement effectué par l'utilisateur ;
- pouvoir visualiser tous les exemples annotés avec une certaine étiquette dans une section déterminée ;
- calculer les collocations en fonction de l'étiquetage (ce qui peut revenir à travailler sur des collostructions). On doit donc pouvoir observer les collocations des unités linguistiques ayant une même étiquette apposée par l'utilisateur. Ce type de calcul doit pouvoir se faire sur les lemmes et/ou les mots. Dans notre étude, cela devrait revenir à observer que *préjudices* et *dégâts* apparaissent souvent avec *causer* au sens de « provoquer », quelle que soit leur position autour de *causer* ;
- pouvoir facilement comparer des résultats provenant de différentes requêtes. Par exemple, visualiser simplement que les textes philosophiques se singularisent pour *causer*, mais aussi pour *cause*, de même pour la tranche comportant les romans fin XIX^{e}-début XX^{e} siècle.

Il faudrait également disposer de logiciels performants pour l'annotation manuelle ou semi-automatique. En effet, lorsque l'on travaille sur des phénomènes que l'on ne peut pas rechercher à partir de chaînes de caractères (les parataxes, par exemple), l'obtention de concordances est inenvisageable. Il faut effectuer le travail manuellement. Mais étant donné l'investissement en temps que cela nécessite, on souhaiterait conserver l'annotation en cours et pouvoir annoter de manière collaborative. Ainsi, une communauté de chercheurs pourrait enrichir collectivement une base de données et/ou comparer leurs analyses directement.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

7 Discussion

A la fin de cette étude, quelques aspects semblent encore mériter une discussion. Tout d'abord, les données orales à partir desquelles nous avons travaillé sont en nombre trop limité pour aborder de manière détaillée l'ensemble des paramètres. C'est ici patent pour la description de l'emploi verbal (seulement 15 occurrences). Mais si l'on compare avec des projets antérieurs tels que COBUILD (1,3 millions de mots d'anglais parlé), il apparaît qu'il n'est pas indispensable de disposer d'une quantité astronomique de données orales pour parvenir à mener des études intéressantes ou élaborer des ressources linguistiques. La catégorisation interne et/ou externe des données nous semble plus urgente à l'heure actuelle, à l'image du travail fourni par Lee (2001). De même, la mise en forme des méta-données disponibles est un autre impératif. C'est d'ailleurs l'articulation peu pratique entre données et méta-données qui nous a le plus gêné pour l'exploitation de la partie orale, nos tranches de référence étant presque exclusivement limitées à la transcription.

Par ailleurs, notre choix initial d'équilibrer la taille des portions orale et écrite, à la réflexion, n'était pas forcément le plus judicieux. En effet, nous aurions pu recourir à des données écrites plus volumineuses vu qu'elles sont plus aisées à se procurer. Cela aurait eu une incidence sur la comparabilité des résultats entre oral et écrit, mais sans que cela soit nécessairement rédhibitoire. Des tendances supplémentaires auraient alors pu être dégagées. De plus, on peut s'interroger sur la proportion optimale d'oral, indépendamment du coût élevé de telles données. Nous ne pourrons répondre à cette question qu'en multipliant les études. Pour ce faire, une stabilisation des données permettra de travailler à partir d'une ressource unique et ainsi de mener des tests à grande échelle.

D'autres questions importantes se posent au sujet de la taille que doit avoir chaque transcription et surtout celle de leur délimitation. En effet, si des chercheurs tels que J. Sinclair ont éprouvé le besoin de travailler sur des textes intégraux pour l'écrit, à l'oral, la pratique généralisée à l'heure actuelle est celle de transcrire des parties de conversation. Or, on peut se demander quelle incidence cela a sur les données. Faut-il que chaque transcription fasse la même taille ou que chaque enregistrement soit transcrit en intégralité ? Seule une étude approfondie de l'incidence de ces paramètres sur les résultats pourra apporter des réponses à ces interrogations.

Du côté de la méthode d'investigation, du chemin reste encore à parcourir pour la stabiliser. Pour mener nos investigations, nous nous sommes appuyés en grande partie sur Sinclair (1991, chap. 3). Le travail a été fait de manière artisanale. Il nous faut désormais explorer les possibilités d'assistance informatique pour accompagner la découverte et l'exploitation des facteurs linguistiques ainsi que la navigation des résultats en fonction des types de données, et ce en attendant de disposer d'un logiciel dédié. Par exemple, on pourra évaluer dans quelle mesure la fonctionnalité *Word Sketch* du logiciel *Sketch Engine* peut nous aider dans le cadre de nos recherches.

object	66462	5.9	subject	29308	4.0	modifier	8945	1.1	and/or	544	0.1	pp_by-p	12251	30.5
damage	3559	9.86	virus	352	7.53	intentionally	50	6.87	exacerbate	11	6.58	bacterium	130	7.53
harm	1120	8.83	bacterium	163	7.05	thereby	61	6.67	aggravate	12	6.54	virus	238	7.42
problem	3813	8.33	driving	159	7.0	some	52	6.4	perpetuate	7	6.52	exposure	97	6.83
death	1834	8.25	root	157	6.74	likely	94	6.21	worsen	14	6.5	earthquake	77	6.74
trouble	1142	8.21	negligence	104	6.72	recklessly	23	6.17	leak	9	5.47	lack	118	6.61
delay	655	7.93	smoking	153	6.67	possibly	95	6.02	contribute	16	4.65	infection	115	6.59
cancer	937	7.79	rain	199	6.62	partly	74	5.96	permit	11	4.56	clot	43	6.53
chaos	508	7.73	infection	161	6.55	mainly	82	5.88	attempt	9	4.03	smoking	88	6.49
havoc	443	7.71	earthquake	98	6.3	apparently	112	5.76	spread	8	3.26	leak	46	6.34

Figure 1. Extrait des résultats de la fonctionnalité *Word Sketch* pour le verbe *to cause* en anglais

Pour finir, notre présentation actuelle des résultats ne nous satisfait pas pleinement. Pour ce qui est de la forme *cause*, elle ne fait pas suffisamment ressortir les phénomènes saillants. La réflexion devra donc également se poursuivre dans ce domaine pour obtenir une présentation standardisée des résultats qui soit plus « parlante » et agréable à lire.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

Références bibliographiques

Adam, J.-M. & Heidmann, U. (2006). Six propositions pour l'étude de la généricité. *La Licorne*, 79, 21-34.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London and New York : Longman.

Bilger, M. & Cappeau P. (2003). Les emplois de *contre* dans les corpus de français parlé et de presse écrite. In Péroz, P. (éd.), Contre*, identité sémantique et variation catégorielle*, *Recherches Linguistiques*, 26, 91-111.

Bilger, M. & Cappeau P. (2004). L'oral ou la multiplication des styles. *Langage & société*, 109/3, 13-30.

Blanche-Benveniste, C. (1986). La notion de contexte dans l'analyse syntaxique des productions orales : exemples des verbes actifs et passifs. *Recherches sur le Français Parlé*, 8, 39-57.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F. & Pires, M. (2009). *Corpus de Français Parlé Parisien des années 2000* (CFPP), *Discours sur la ville*.

Cappeau, P. & Gadet, F. (2007). L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale. *Revue Française de Linguistique Appliquée*, 2007/1, 121, 99-110.

Delic (2004). Présentation du *Corpus de Référence du Français Parlé*. *Recherches sur le français parlé* n°18, 11-42.

Deulofeu, J. (1986). Syntaxe de *que* en français parlé et le problème de la subordination. *Recherches sur le français Parlé*, 8, 79-104.

Durand, J., Laks, B. & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. Pusch, C. & Raible, W. (eds.), *Romanistische Korpuslinguistik - Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, Tübingen, Gunter Narr Verlag, 93-106.

Durand, J., Laks, B. & Lyche, C. (2005). Un corpus numérisé pour la phonologie du français. Williams, G. (éd.), *La linguistique de corpus*, Rennes : Presses Universitaires de Rennes, 205-217.

Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texto!* [en ligne], vol. X, n°4.

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund, Y. (2008). *Corpus Linguistics with* BNCweb - *a Practical Guide.* Frankfurt am Main: Peter Lang.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle, *Language Learning & Technology*, 5-3, 37-72.

THE BNC JUNGLE

Louw, W.E. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds), *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins, 157-176.

Malrieu, D. & Rastier, F. (2001). Genres et variations morphosyntaxiques. *TAL*, 42-2, 547-577, Hermès.

McEnery, T., Xia, R. & Tono, Y. (2006). *Corpus-Based Language Studies. An advanced resource book*. New York: Routledge Applied Linguistics.

Miller, J. & Weinert, R. (1998). *Spontaneous spoken language. Syntax and discourse*. Oxford: Clarendon Press.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. In Williams, G. (éd), *La linguistique de corpus*, Rennes : PUR, 31-45.

Renouf, A. (1984). Corpus development at Birmingham University. In Aarts, J. and Meijs, W. (eds.).

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford University Press.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse*. Routledge.

Stefanowitsch, A. & Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8:2, John Benjamins, 209-243.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Pluri-thématique
DOI 10.1051/cmlf/2010237

[1] Dans leurs données principales, l'écrit représente environ 80 % du total.

[2] Le paysage concernant les corpus oraux se modifie assez rapidement en France. Depuis quelques années un appui institutionnel a permis la réalisation de nombreux projets d'envergure (CRFP, PFC, CLAPI, CFPP2000) et d'autres sont en cours (CIEL-F).

[3] Le CERF est un corpus qui a été constitué sous la direction de Jean Véronis et qui comporte 9 tranches écrites de 1.000.000 de mots (récupérés en partie sur Internet) + une tranche orale (*Corpaix*).

[4] *Corpaix* est une archive constituée par l'équipe du GARS.

[5] Le CRFP (ou *Corpus de Référence du Français Parlé*) a été présenté dans la revue *Recherches Sur le Français Parlé* n° 18 (2004). Il s'agit d'un corpus de plus de 400.000 mots dont les enregistrements ont été effectués dans diverses villes de l'hexagone et selon trois situations de parole : parole privée, parole publique, parole professionnelle.

[6] Une version de démonstration de ce logiciel est disponible à l'adresse :
http://sites.univ-provence.fr/veronis/logiciels/Contextes

[7] Il y a également la forme *causes*, mais nous n'en avons pas trouvé d'attestation.

[8] *LoX* est un logiciel développé par L. Audibert et qui permet de faire des requêtes sur des corpus annotés. Il peut être téléchargé à l'adresse : http://www-lipn.univ-paris13.fr/~audibert/lox.php

[9] Nous avons tout de même souhaité tester la requête à partir des données préalablement étiquetées à l'aide du logiciel *Cordial Analyseur 2008*. Les concordances obtenues ne sont pas parfaites et réclament une phase de correction minimale. Le travail n'aurait donc nullement été facilité.

[10] Stubbs (1996) a décrit ce même phénomène pour le verbe anglais *to cause*.

[11] *Sketch Engine* (http://www.sketchengine.co.uk) est un logiciel qui permet d'utiliser ses propres corpus et ce, quelle que soit la langue. Il constitue à l'heure actuelle une piste à explorer pour les linguistes travaillant sur le français.

[12] Des informations concernant BNCWeb se trouvent à l'adresse http://bncweb.info et dans Hoffmann et al. (2008). Pour l'utiliser, il suffit de s'inscrire à : http://bncweb.lancs.ac.uk/bncwebSignup.

[13] http://cwb.sourceforge.net