UNE INTRODUCTION AU *BRITISH NATIONAL CORPUS* DANS SON ÉDITION XML[*1]

Lou Burnard
Oxford University Computing Services

Le *British National Corpus* (*BNC*) a eu au cours de la dernière décennie une influence considérable sur la construction des corpus de langue, au moins comme point de référence majeur. Ce corpus est, pour ainsi dire, le point culminant d'une tradition de recherche remontant à 1964 et au corpus Brown d'un million de mots. Mais sa structuration et ses techniques de production à une échelle industrielle permettent surtout d'envisager un nouveau monde, où l'ingénierie centrée sur le langage et le développement de logiciels est enfin au cœur de la société de l'information, plutôt que de s'enferrer dans des développements théoriques stériles.

Cet article revient sur les problèmes de conception et de gestion, ainsi que sur les options prises lors de la construction du *BNC* ; il justifie enfin la pertinence toujours actuelle de sa version la plus récente, au format XML.

1 QU'EST-CE DONC, LE *BNC* ?

Le *BNC* est un corpus de cent millions de mots d'anglais britannique moderne, produit au départ, entre 1990 et 1994, par un consortium regroupant des éditeurs de dictionnaires et des chercheurs universitaires. Le consortium réunissait les Presses Universitaires d'Oxford, Longman et Chambers, en qualité d'éditeurs de dictionnaires, ainsi que les centres de recherche des universités de Lancaster, d'Oxford et de la *British Library*. Le projet fut au départ conçu dans le cadre du « Programme commun pour les Technologies de l'Information », une initiative partiellement subventionnée du gouvernement britannique, dont l'objectif était de faciliter la coopération entre le monde de l'entreprise et celui de la recherche universitaire. Pour cela, partenaires commerciaux et recherche universitaire étaient financés par le Ministère Brittanique de l'Industrie et du Commerce et par celui de «Science and Engineering Research Council », respectivement à hauteur de 50 et 100 %.

L'histoire sociale a qualifié ces années 1990 de diverses manières ; puisque notre propos est lié à l'équipement informatique, je suggère quant à moi de leur attribuer le qualificatif de « néotoniques ». Il est utile de se souvenir que, dans les magazines informatiques du début des années 1990, la discussion principale comparait les mérites relatifs des traitements de textes *WordPerfect* (version 5) et *WinWord* (un ancêtre de l'omniprésent *Microsoft Word* actuel). Si vous étiez raisonnablement pourvu par votre université, vous pouviez alors avoir sur votre bureau un ordinateur personnel doté du rapide processeur Intel 386, avec au moins 50 Mb de mémoire disque – juste assez somme toute pour lancer le *Windows 3.1* de *Microsoft*. Mais le traitement informatique réel était réalisé par votre laboratoire ou par le service informatique central, sans soute pourvu d'un système Unix, quel qu'il fut, ou d'un ordinateur VAX. À cette

---

[*] Pour faire référence à cet article : Burnard Lou, « Une introduction au British National Corpus dans son édition XML », revue électronique *Texte et corpus*, n°3 / août 2008, Actes des Journées de la linguistique de Corpus 2007, p. 17-34 (disponible sur http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_burnard.pdf).

[1] Une première version de cet article (en anglais) a été publiée sous le titre « The BNC: Where did we go wrong », in : Kettemann B. & Markus G., *Teaching and learning by doing corpus analysis*, Amsterdam : Rodopi, p. 51-71.

La traduction de cet article a été réalisé par N. Dugalès et D. Seguin, avec la collaboration de C. Millon et G. Williams.

époque également, certaines personnes commencèrent à parler d'un nouveau concept d'hypertexte, nommé « *World Wide Web* », et à tester une nouvelle et impressionnante interface graphique appelée *Mosaic*…

Malgré tout, l'art de construire un corpus était déjà bien maîtrisé dans ces années 1990, tout au moins par les Européens. C'est avec une indéniable surprise que Leech déclare, dans la préface des *Actes de l'ICAME* de 1990 : « Les corpus sont en train de devenir le courant principal ». Nous pouvons distinguer trois courants intellectuels, ou trois approches distinctes, émergeant alors : l'école traditionnelle, initiée par le corpus Brown, institutionnalisée à travers le *LOB* (*Lancaster-Oslo/Bergen corpus*) et perpétuée à travers l'*ICAME* (*International Computer Archive of Modern and Medieval English*) ; l'École de Birmingham qui, participant au projet *Cobuild* (*Collins Birmingham University International Language Database*) vers la fin des années 1980[2], construit des collections toujours plus importantes de matériaux textuels ; et l'approche américaine dont l'expression la plus célèbre a été celle de Mitch Marcus : « there's no data like more data » (« rien de tel que toujours plus de données »). La naissance de la lexicographie assistée par ordinateur est la conséquence la plus visible de la combinaison de ces différentes traditions de recherche, bénéficiant non seulement de la mise à disposition de dictionnaires traditionnels - tels que le *Dictionnary of Contemporary English* de Longman et, bien sûr, la numérisation du *Oxford English Dictionnary* lui-même, mais aussi d'un intérêt sans cesse croissant, au sein de la communauté des linguistes, pour la linguistique computationnelle (Atkins, 1992).

Par ailleurs, ce début des années 1990 constitue une période particulièrement stimulante au regard de la synergie existant alors au sein de la recherche d'applications pour les technologies de l'information. Les « sciences humaines computationnelles » et la « linguistique computationnelle » remportèrent en ce temps-là leur premier (et à ce jour, unique) succès commun : l'établissement d'une norme de balisage des textes appropriée à la naissance de l'ère digitale[3]. Les termes d'« ingénierie du langage » ont été utilisés non pour décrire une douteuse forme de politique sociale, mais une nouvelle technologie en quelque sorte plus « sexy ». C'est dans ce contexte que la création du *BNC* fut financée trois années durant, avec un budget d'environ 1,5 million de livres sterling.

Ce projet d'établissement du *BNC* est né d'une convergence inhabituelle d'intérêts entre lexicographes, éditeurs, chercheurs et gouvernement. Parmi les éditeurs, les Presses Universitaires d'Oxford et Longman envisageaient déjà les potentiels bénéfices de l'utilisation d'un corpus. Le succès des dictionnaires *Collins COBUILD* (le premier, publié en 1987, fut probablement le premier dictionnaire majeur pour apprenants entièrement acquis aux principes du corpus) stimula également fortement la motivation de ces éditeurs rivaux que sont OUP et Longman. Pour le gouvernement, un facteur clé était le désir de stimuler l'industrie en ingénierie de la langue anglaise, dans le contexte d'un intérêt croissant pour ce domaine en Europe. Nous détaillerons un peu plus bas que, pour les chercheurs d'Oxford et de Lancaster, cette synergie fortuite était également l'occasion inespérée de pousser encore plus loin les limites de la construction de corpus. Pour la *British Library*, le corpus était enfin l'un de ces projets exploratoires permettant d'expérimenter les nouveaux médias à l'aube de l'ère des bibliothèques numériques (pour trouver d'autres exemples, *cf.* Carpenter, 1998).

Les buts prescrits du projet *BNC* étaient dès le départ clairement explicites : il s'agissait de créer un corpus de langue au moins aussi grand que tout ce qui était librement utilisable

[2] Pour un résumé de ce travail, préfigurant en bien des aspects le *BNC*, voir Renouf (1986) et les nombreuses autres publications de son mentor, J. Mc H. Sinclair (ex. Sinclair 1987).

[3] L'introduction de l'ouvrage de Zampolli (1994) explicite cette connexion.

jusque-là[4]. Ce nouveau corpus se devait d'être synchronique et contemporain, et constituer un échantillonnage conséquent de l'ensemble de la langue anglaise, à la fois parlée et écrite. Les débats et discussions ont beaucoup porté sur cette notion d'échantillon, en particulier pour la construction du corpus. À la différence d'autres collections de données langagières alors populaires, on peut définir le projet du *BNC* comme « non opportuniste », sa conception ne visant pas directement des objectifs de rentabilité économique. Afin de créer un corpus utilisable par tous, il devait être entièrement étiqueté - de manière automatique, et contenir de plus une information contextuelle très détaillée. Ces trois caractéristiques, ajoutées à sa grande taille et à sa disponibilité générale, rendaient le *BNC* unique au sien des autres collections de données langagières, et justifiaient par ailleurs le « national » de son titre (ce qualificatif ne fut initialement inclus qu'en compensation de son partiel financement public).

D'autres visées existaient, mais ces dernières n'étaient pas clairement définies même si elles étaient présentes de manière implicite dès la conception du projet. Pour les partenaires commerciaux, la raison principale de leur investissement substantiel, en temps et argent, était bien sûr la production de meilleurs dictionnaires pour apprenants ainsi que, peut-être, l'espoir de gains en terme de reconnaissance. Pour les partenaires universitaires, un objectif implicite était d'établir un nouveau modèle de développement de corpus, au sein d'une industrie des langues alors en pleine émergence en Europe. Il s'agissait également de tester des idées nouvelles quant à la standardisation du balisage des textes et son interprétation. Mais, plus que tout cela, il y avait le simple souhait de construire un très, très gros corpus !

2 L'ORGANISATION DU PROJET

Cette coopération entre industriels et chercheurs a eu une conséquence inattendue et intéressante, celle de souligner la nécessité de trouver un accomodement entre désir de perfectabilité des chercheurs et impératifs financier et de temps des partenaires commerciaux (pour plus de détails, *cf.* Burnard, 1999). Par la mise en place d'un système de production de texte à une échelle industrielle, l'instauration du projet lui-même fut inévitablement l'objet de compromis tant aux niveaux conceptuel que exécutif. La Figure 1 illustre la chaîne de production du *BNC*, surnommée par le manager du projet, Jeremy Clear, « *the BNC Sausage Machine* » (*La machine à saucissonner du BNC*)

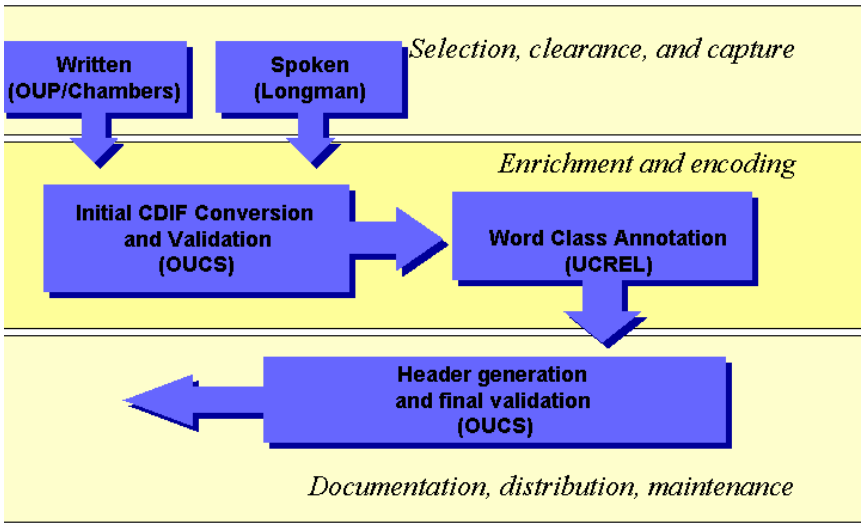

Figure 1 : « The BNC Sausage Machine » (la machine à saucissonner du BNC)

[4] Bien qu'incontestablement plus grand, le corpus *Bank of English* développé au cours du projet *Cobuild* n'a pas été créé à l'origine pour la distribution ou l'utilisation à l'extérieur du projet. À ce jour, les droits de propriétés intellectuels et autres restrictions en limitent effectivement l'accès.

Comme le montre cette figure, la production des différents types de matériau était répartie entre les différents acteurs du projet. Longman était chargé de la collecte et de la transcription du matériau oral. Les OUP se concentraient quant à elles sur la transcription du matériau écrit, soit par l'utilisation de différents systèmes de reconnaissance optique de caractère, la saisie informatique ou bien encore en utilisant des données déjà disponibles sous forme numérique. Le rôle des services informatiques de l'Université d'Oxford, Oxford University Computing Services (OUCS), était ensuite d'effectuer la conversion de tous ces matériaux en un seul et même format, et d'en valider la structure. OUCS maintenait également à jour une base de données, renseignant sur les sources et flux de travail. Les annotations linguistiques du matériau étaient effectuées par l'équipe de Lancaster, qui a utilisé pour cela l'étiqueteur déjà bien connu *CLAWS* (*cf. infra.* et Garside, 1996). Aux textes ainsi obtenus fut finalement associée une description normalisée des méta-données, extraite de la base de données sous la forme d'un document conforme aux recommandations de la TEI (*Text Encoding Initiative*) (Sperberg & McQueen, 1994).

Bien entendu, la « machine à saucissonner du *BNC* » n'a pas tourné à vitesse constante tout au long du projet, et il y eu d'inévitables interruptions ou blocages. Le développement du corpus était réparti entre cinq groupes de travail, auxquels participaient de manière variable les équipes de chaque membre du consortium. Nous pourrions résumer le travail de chacun de ces groupes de la façon suivante :

- autorisations : conception (et envoi) de courriers standards pour l'autorisation d'inclure dans le corpus tous les documents protégés par des droits de propriété intellectuelle ;
- critères d'élaboration : définition des différents types de textes à intégrer au corpus et dans quelles proportions ;
- enrichissement et annotations : mise en œuvre de l'annotation linguistique et contextuelle des textes du corpus ;
- Encodage et l'étiquetage : définition de la structure de balisage, servant de référence à l'ensemble du corpus final, et des procédures permettant de l'appliquer aux différents formats de collecte des textes.
- Logiciel d'extraction : définition et mise en œuvre d'un logiciel d'extraction simple de l'ensemble des informations encodées dans le corpus.

Chacun de ces sujets sera abordé de façon plus détaillée dans les sections suivantes.

2.1 Le problème des autorisations.

Comme indiqué ci-dessus, le *BNC* était le premier corpus de cette taille conçu pour être largement accessible. Ceci n'a été en grande part possible que grâce au travail fourni par ce groupe de définition d'un contrat d'utilisation qui fasse l'unanimité, entre les propriétaires des droits, le consortium et les utilisateurs. Pour les propriétaires des droits, ceux-ci devaient non seulement autoriser l'intégration de leur matériau au corpus sans exiger pour autant de compensation financière, mais également accepter une licence d'utilisation encore en vigueur aujourd'hui. Le caractère innovant du concept et le prestige associé au projet ont peut-être, dans une certaine mesure, facilité un tel accord. Tous les propriétaires contactés n'ont cependant pas, loin de là, accepté immédiatement une cession gracieuse pour un usage, sans limitation de durée, des versions numérisées. Certains préférèrent éviter de s'engager de quelque façon que ce soit ; d'autres refusaient tout accord gratuit.

En matière d'autorisation, le matériau oral soulevait deux problèmes spécifiques. Les participants ayant été assurés que leur identité demeurerait secrète, d'importants efforts durent être déployés afin de trouver le meilleur moyen pour préserver l'anonymat des contributions sans compromettre outre mesure leur utilité linguistique. Des allusions trop explicites à l'identité des personnes furent en de nombreux cas supprimées et, si la possibilité d'employer

des pseudonymes (linguistiquement semblables) fut un temps envisagée, cela s'est avéré pratiquement impossible.

Un problème plus embarrassant est venu du fait que, pour les personnes qui furent interviewées dans l'objectif d'intégrer leur discours à l'échantillon conversationnel du corpus[5] (« *The demographically sampled part of the corpus* », composante principale du corpus oral du *BNC*), ne fut demandée (et accordée) que l'autorisation d'inclure les versions retranscrites de leurs discours, non ce discours lui-même. Il eu fallut demander cette autorisation aux locuteurs originaux eux-mêmes ; l'efficacité de l'anonymisation des procédures a rendu ensuite la chose beaucoup plus complexe.

Deux facteurs supplémentaires ont freiné l'empressement des détenteurs de droits à nous donner leur accord : tout d'abord, le fait que le corpus n'intégrerait aucun texte complet et ensuite, que l'exploitation, ou la distribution, des matériaux du corpus n'était l'objet d'aucune intention commerciale. Ceci n'excluait toutefois pas l'utilisation commerciale de produits dérivés, créés pour permettre l'accès au corpus. Cette distinction, explicite dans la licence utilisateurs, est essentielle pour permettre à la fois une disponibilité à long terme du corpus pour la recherche, et un emploi dans le secteur commercial par exemple comme plateforme d'expérimentation d'autres projets, du modeste correcteur orthographique à la plus sophistiquée mémoire de traduction. Pour souligner ce positionnement initial de ne pas distribuer le corpus lui-même sur une base commerciale, l'un des membres universitaires du consortium, OUCS, a été nommé comme l'unique agent à même de fournir les licences, et de rapporter en outre tout cas litigieux au consortium lui-même. Initialement limitée à l'Union Européenne, la distribution du corpus en dehors de l'Europe fut finalement possible à partir de 1998.

2.2 Les critères de conception.

Le caractère « non opportuniste » du *BNC* est rapidement évoqué ci-dessus (*supra.* Section 1) ; un retour sur le contexte historique de l'époque permettra de réellement percevoir cette spécificité comme remarquable et distinctive. Durant les années 1990, même si toutes sortes de supports textuels faisaient l'objet d'une maquette au format numérique avant leur impression sur papier, l'idée que cette forme numérique elle-même pouvait avoir une quelconque valeur n'était pas monnaie courante. En outre, la numérisation était à cette époque précédant le commerce électronique encore loin d'être une constante, tant dans sa couverture que dans les formats. La conséquence était que les chercheurs succombaient à une tendance, bien compréhensible, de sauter sur n'importe quel texte électronique disponible sans prendre plus en compte leur statut spécifique par rapport à la langue en générale. Pour prendre un exemple célèbre, une grande partie du *Wall Street Journal* était alors déjà disponible sous forme numérique et guettait le danger suivant : celui de s'appuyer sur cet unique type d'écrit, d'un registre bien spécifique, pour servir de base à la linguistique computationnelle naissante et en déduire ainsi des généralisations abusives à l'ensemble de la langue.

Pour corriger ceci, le projet *BNC* s'appuya dès sa conception sur le respect d'un certain nombre de critères précis, pas uniquement leur disponibilité sous forme numérique, pour collecter ses échantillons de la langue. Ces critères (judicieusement résumés dans Atkins, 1992) ont défini les caractéristiques exactes des textes ainsi que leur proportion, pour la constitution de l'échantillon de matériaux écrits du corpus. L'objectif premier du *BNC* était de permettre de dire quelque chose sur la langue en général. Mais cette langue est-elle ce qui est

[5] Dans l'objectif de rendre compte de l'ensemble des discours oraux produits, le *BNC* ne pouvait limiter son corpus oral à la seule retranscription de matériaux conversationnels recueillis auprès d'une part représentative de la population brittanique (« *the demographically sampled part of the corpus* »). Un second sous-corpus fut donc élaboré afin de prendre en compte les situations plus cadrées, l'échantillon de discours produits en contexte dirigé (« *the context-governed sample part of the corpus* »). *Cf.* Section 2.2 [ndt].

réceptionné (lu ou entendu) ou ce qui est produit (écrit ou parlé) ? En bons pragmatiques anglo-saxons, les concepteurs du *BNC* ont décidé d'ignorer la traditionnelle dichotomie Saussurienne en tentant de tenir compte des deux perspectives.

L'objectif était de définir un échantillon à plusieurs niveaux répondant aux différents critères choisis. Même s'il est délicat de déclarer le corpus statistiquement représentatif de l'ensemble du langage, en terme de production ou de réception, le *BNC* ambitionne cependant de représenter des variabilités reconnues de la langue, définies selon certaines dimensions spécifiques telles que le mode de production (oral ou écrit), le média (livre ou journal, etc.), le domaine (fiction, scientifique, loisirs etc.), le contexte social (formel, informel, entreprenorial, etc.) et ainsi de suite.

Cet article n'a pas pour propos de répéter en détail les raisons ayant guidé les choix opérés dans le *BNC*[6] pour la classification des textes. Par exemple, les « textes oraux » peuvent être caractérisés soit par l'âge, le sexe, ou la classe sociale (de la personne collectant ces données, non de son interlocuteur), soit par le domaine, la région ou le type de discours enregistré ; les « textes écrits » de leur côté seront également être caractérisables selon leur auteur (âge, sexe ou type), leur public (diffusion, statut) ou encore, comme nous l'avons déjà vu ci-dessus, le média ou le domaine. Certaines de ces catégories relevaient de critères initiaux de sélection, exigeant donc de respecter des valeurs et des proportions prédéfinies ; d'autres critères n'auront qu'une valeur descriptive, puisqu'aucune proportion particulière n'était recherchée si ce n'est une volonté de maximiser les variables au sein de chaque catégorie. Il est essentiel de remarquer que ces indications ont pour seul objectif d'améliorer plus encore la documentation du corpus, non d'en faciliter l'accès ou de créer des sous-corpus en lien avec quelque théorie que ce soit.

Les objectifs poursuivis, qu'une telle conception du projet justifiait, ont en tout état de cause du être tempérés au regard des réalités de la vie économique. Pour résumer grossièrement cette conjecture, il suffit d'indiquer que le coût de la retranscription d'un matériel oral est 10 fois plus important que celui de l'ajout de textes, de proportion équivalente en nombre de mots, en provenance d'un quotidien. La proportion entre matériaux écrit et oral est ainsi de dix contre un, même si nombreux sont ceux qui suggéraient que, l'oral et l'écrit ayant selon eux la même importance en terme de signification dans la langue, ils devaient être présents en quantité égale dans le corpus. Au sein du corpus oral, il fut tout de même tenté de représenter à la fois différents types de discours élaborés dans un cadre normé (« *The context-governed sampled part* ») et différentes formes d'anglais conversationnel (« *The demographically sampled part* »).

Ce sont, de manière similaire, des questions pragmatiques qui conduisirent à la présence prédominante, au sein du corpus écrit, de livres publiés et de périodiques. Si cependant ce type de textes - publiés sous la forme de livres, de magazines, etc. - ne peut être totalement représentatif de la langue écrite produite (puisque écrire en vue d'une publication est une activité relativement spécialisée, dans laquelle finalement peu de personnes s'engagent), il demeure tout de même largement représentatif de la langue écrite telle qu'elle est reçue par la plupart des personnes. Une quantité non négligeable d'autres productions textuelles fut en outre ajoutée au corpus (notamment des matériaux non publiés, lettres ou littérature grise). Enfin, au sein d'un type de textes aisément accessible comme les journaux, nous avons pris tout de même le soin de constituer des échantillons, intégrant différents quotidiens et tabloïds, à la fois nationaux et régionaux, et faisant en sorte que les plus faciles à collecter (les quotidiens nationaux) ne noient pas les autres sous leur nombre.

[6] Ceci est abordé de façon exhaustive dans Atkins (1992) pour le matériau écrit, et dans Crowdy (1995) en ce qui concerne le matériau oral. Des commentaires et tableaux détaillés sont également fournis dans le « BNC User Reference Guide » (Burnard, 1995, revu en 2000, http://www.natcorp.ox.ac.uk/docs/userManual/)

Dans sa version actuelle, l'édition XML du *BNC* (*BNC XML Edition*) intègre 4 049 textes pour environ 4.5 Gb (balisage XML inclus). Il contient au total un peu moins de cent millions de mots (il y a 98 363 784 éléments lemmatisés, soit un peu plus de six millions (6 026 284) de segments, tels qu'identifiés par *CLAWS*). Le corpus se décompose de la façon suivante (*cf. Tableau 1*) :

	Texts	W-units	%	S-units	%
Spoken demographic (Anglais conversationnel représentatif de la population brittanique)	153	4 233 955	4,30	610 557	10,13
Spoken context-governed (Paroles énoncées en contexte dirigé)	755	6 175 896	6,27	427 523	7,09
Written books and periodicals (Ecrits en provenance de livres publiés ou de périodiques)	2 685	79 238 146	80,55	4 395 581	72,94
Written-to-be-spoken (Ecrits destinés à être énoncés)	35	1 278 618	1,29	104 665	1,73
Written miscellaneous (Ecrits divers)	421	7 437 168	7,56	487 958	8,09

Tableau 1 : Le Corpus BNC actuel par type de textes.

« texts » : indique le nombre d'échantillons distincts, chacun n'excédant pas les 45 000 mots.
« S-units » : indique le nombre de segments identifiés par le système CLAWS (correspond globalement au nombre de phrases)
« W-units » : indique le nombre d'éléments « w » identifiés par le système CLAWS (correspond globalement au nombre de mots).

Au sein de la partie écrite du corpus, étaient définies au préalable les proportions désirées dans la répartition entre les différents types de médias et les différents sujets. Le *Tableau 2* en détaille quelques-unes :

	Texts	W-Units	%	S-Units	%
imagination	476	16 496 420	18,75	1 352 150	27,10
Instruction : sciences dures et naturelles	146	3 821 902	4,34	183 384	3,67
Instruction : sciences appliquées	370	7 174 152	8,15	356 662	7,15
instruction : sciences sociales	526	14 025 537	15,94	698 218	13,99
instruction: monde des affaires	483	17 244 534	19,60	798 503	16,00
instruction: commerce et finance	295	7 341 163	8,34	382 374	7,66
instruction: arts	261	6 574 857	7,47	321 140	6,43
instruction: croyances et opinions	146	3 037 533	3,45	151 283	3,03
instruction: loisir	438	12 237 834	13,91	744 490	14,92

Tableau 2 : Répartition des différents types d'écrits dans le corpus.

La partie orale du corpus est elle-même composée de deux sous parties. Une première moitié (approximativement) comprend des conversations informelles enregistrées auprès d'environ 200 volontaires, recrutés pour le projet par une agence privée d'étude de marché. Ce premier ensemble (« *the demographically sampled part* ») forme un échantillon représentatif selon des critères d'âge, de sexe, de zone géographique, et de classe sociale. Cette méthode d'échantillonnage reflète la répartition démographique de la langue parlée, mais en raison de sa taille modeste, il exclut probablement du corpus de nombreuses variations linguistiquement significatives relatives au contexte. Pour compenser cela, l'autre moitié du corpus oral (« *the context-governed sampled part of the corpus* ») se compose de discours enregistrés dans diverses situations prédéfinies, des réunions publics ou semi-publics par exemple, des interviews professionnelles, des débats (semi-)formels en contexte universitaire, entreprenorial ou de loisirs, ...

On peut regretter après coup que certaines catégorisations (l'origine ethnique de l'auteur, par exemple) aient été trop partiellement définies. La précision et la constance avec laquelle toutes ces variables sont en réalité présentes dans les en-têtes des textes ont sérieusement

souffert à la fois des contraintes de production et du manque d'informations disponibles. Même le concept apparemment aussi neutre que celui de la datation n'était pas sans poser problème pour les textes écrits : devions-nous prendre en compte la date de la version utilisée ou bien celle de la première édition ? De la même façon, quand nous parlions de « l'âge de l'auteur », devions-nous considérer son âge au moment de la publication ou bien celui à l'impression de l'ouvrage ?

D'autres corpus avant le *BNC* ont bien sûr été conçus selon des méthodes semblables, quoiqu'à une moindre échelle. Néanmoins, les méta-données accompagnant ces corpus étaient alors généralement considérées comme un élément strictement distinct du corpus lui-même, pouvant être consulté par quelques curieux dans le *Manuel d'information*. La TEI impliquait une innovation, adoptée par le *BNC* : celle d'intégrer à chaque fichier texte du corpus un en-tête distinct certes mais respectant une convention unique pour l'ensemble du corpus. Cet en-tête contient des informations d'identification et de classification de chaque texte ; il caractérise par exemple les locuteurs ou bien encore fournit des informations plus techniques quant à la taille, les dates de mises à jour, etc. Suivant toujours les principes de la TEI, toutes les définitions à jour de ces catégorisations du corpus *BNC* ont été renseignées dans un unique fichier d'en-tête applicable à l'ensemble du corpus, permettant la lecture des en-têtes spécifiques de chaque texte individuel[7].

Durant le proces de production, les différents protagonistes, ci-dessus mentionnés, ont tout naturellement considéré le travail de classification et l'emploi des autres méta-données comme partie intégrante de leur tâche de collecte. Chaque texte individuel fut donc stocké comme tel. Ce n'est qu'ensuite que OUCS a compilé l'ensemble de ces informations dans une unique base de données, à partir de laquelle furent générés les en-têtes conformes aux normes de la TEI. Avec toute la meilleure volonté du monde, il était donc difficile d'éviter les incohérences dans la saisie des méta-données et, en conséquence, d'assurer avec une complète certitude, au moment de leur intégration, leur uniformité.

À la suite d'une réévaluation approfondie sur l'ensemble du corpus des catégorisations appliquées à chaque texte individuel, un étudiant de Lancaster, David Lee, procéda à la vérification de toutes les classifications en place et proposa également une taxonomie plus détaillée des types de textes. Cette recherche systématique (Lee, 2001) nous a certes permis de fiabiliser nos critères de catégorisation initiaux, mais aussi et surtout d'adopter une classification totalement nouvelle agrémentée d'une taxonomie infiniment plus précise, telle que définie par Lee. À la même époque, tandis que pour le *BNC-1* l'association de mots-clefs thématiques à chaque texte ne se faisant de manière aussi systématique, dans la nouvelle version un ensemble de mots-clefs descriptifs, inspirés de la nomenclature des bibliothèques[8], est automatiquement associé à chaque texte écrit. Ces multiples codes de classification ont été simplifiés, de façon à permettre la partition du corpus selon 8 types principaux de textes (*Tableau 3*).

	Texts	W-units	%	S-units	%
Ecrits universitaires	497	696 038	11.55	15 781 859	16.04
Fictions et vers	452	1 323 573	21.96	16 143 913	16.41
Journalisme et information	486	508 609	8.43	9 412 174	9.56
Autres écrits publiés	711	1 021 633	16.95	17 970 212	18.26
Ecrits non publiés	251	303 078	5.02	4 466 681	4.54
Conversation	153	610 558	10.13	4 233 962	4.30
Autre anglais parlé	755	427 523	7.09	6 175 896	6.27

Tableau 3 : Le BNC XML par type de texte.

[7] Pour une description plus détaillée de la manière dont les en-têtes furent utilisés dans le BNC, *cf.* Dunlop, 1995.
[8] Ces mots-clefs sont issus du COPAC, catalogue des librairies universitaires et nationales de Grande- Bretagne, pour l'intégralité des matériaux publiés.

2.3 Annotations.

Les mots du *BNC* ont été étiquetés automatiquement avec *CLAWS 4*, étiqueteur développé à l'Université de Lancaster à partir de *CLAWS 1* qui fut à l'origine conçu pour le corpus *LOB* de un million de mots. Le système est décrit de manière détaillée dans Leech (1994), les aspects théoriques et pratiques dans Garside (1997), et une documentation technique complète de son utilisation pour le *BNC* est fournie dans le manuel accompagnant *BNC World Edition* (Leech, 2000).

CLAWS 4 est un système d'étiquetage hybride, mêlant techniques probabilistes et non-probabilistes. Chaque mot du corpus se voit par lui attribué un (parfois deux) code POS (« *part-of-speech* ») résultant de 4 processus distincts :

1. La segmentation du discours en mots (habituellement identifiés par les espaces) et phrases orthographiques (marquées par la ponctuation) ; Les verbes enclitiques (comme « *'ll* » ou « *'s* » pour l'anglais), et les particules négatives (comme « *n't* ») sont ici des exceptions, tout comme certaines formes contractées telles que « *dunno* » (qui a été symbolisée par « *do+n't+know* »).
2. L'attribution initiale du code POS : tous les codes POS qui peuvent être assignés à un *token* sont extraits, soit par l'examen d'un lexique de 50 000 mots, soit par l'application de quelques procédures morphologiques simples. Quand plus d'un code est assigné à un mot, la probabilité relative pour chacun est fournie de la même manière. Les probabilités sont également ajustées sur la base de la position du mot au sein de la phrase.
3. La désambiguïsation ou sélection d'un unique code est ensuite opérée en utilisant l'algorithme d'alignement de Viterbi, qui recherche les probabilités associées à chaque code pour déterminer la meilleure voie à suivre à travers une séquence de codes ambigus, fonctionnement similaire à celui de l'écriture intuitive dont sont équipés de nombreux téléphones portables aujourd'hui. À la fin de cette étape, les différents codes possibles sont classés par probabilité décroissante pour chaque mot, en contexte.
4. L'étiquetage des expressions idiomatiques est un raffinement supplémentaire de la procédure, pour lequel à chaque groupe de mots, avec leurs étiquettes, est apparié un patron idiomatique prédéfini qui ressemble à un réseau d'associations limitées.

En procédant de cette manière, *CLAWS* atteint un taux de précision de plus 95 % (*i.e.* absence d'indétermination) dans l'assignation de codes POS sur chaque mot du corpus. Pour améliorer ceci, l'équipe de Lancaster a davantage développé l'idée essentielle d'« étiquetage d'expression idiomatique », en utilisant un étiqueteur de patron idiomatique capable de prendre en compte des règles contextuelles bien plus sophistiquées, dérivées en partie de procédures semi-automatiques extraites d'un échantillon de textes désambiguïsé manuellement (*cf. BNC User Reference Guide*, 1995/2000).

Les annotations linguistiques du corpus ont été enrichies pour l'édition XML du *BNC* de façon à respecter trois choses :

- les unités multi-lexicales et leurs composants sont étiquetés de façon explicite en utilisant les éléments XML « mw » et « w » ;
- sur la base du C5, jeu d'étiquettes développé pour le logiciel CLAWS, une annotation complementaire et simplifiée des codes de POS est déployée ;
- la lemmatisation est réalisée automatiquement à partir de règles définies manuellement.

Par unités multi-lexicales, nous entendons le fait qu'au moins deux mots orthographiques fonctionnent selon *CLAWS* comme une entité avec une catégorie grammaticale unique. Les exemples les plus communs sont les locutions adverbiales telles que « *of course* » (*bien sûr*)

ou « *in short* » (*en bref*), et les séquences prépositionnelles comme « *in spite of* » (*en dépit de*) ou « *up to* » (*jusqu'à*). Décider si une séquence orthographique est ou n'est pas une unité multi-lexicale exige parfois l'interprétation (par exemple « *in short* » - *en bref*, n'est pas une unité multi-lexicale dans « *in short sharp bursts* » - *de courtes explosions aiguës*) ; de telles situations ont nécessité d'aller au-delà des seules règles idiomatiques.

Dans l'édition XML du *BNC*, ces unités multi-lexicales ont donc été indiquées à l'aide d'un élément XML supplémentaire (« mw ») indiquant la catégorie grammaticale assignée à toute la séquence. Au sein de l'élément « mw », les mots orthographiques sont également individuellement étiquetés, par un usage similaire à toutes les autres parties du texte de l'élément « w ». L'entité « *of course* » est ainsi étiquetée de la manière suivante :

```
<mw c5="AV0"> <w c5="PRF" hw="of" pos="PREP">of </w> <w c5="NN1" hw="course"
pos="SUBST">course </w> </mw>
```

La catégorie grammaticale des composants de l'unité multi-lexicale est insérée automatiquement, en utilisant un tableau construit à partir du corpus. L'étiquette assignée le plus fréquemment pour un mot est ainsi sélectionnée, puis les aberrations sont manuellement corrigées, même si quelques erreurs peuvent perdurer. Le jeu d'étiquettes simplifié, employé pour cet enrichissement, est structuré à partir de la réduction ordonnées des étiquettes de *CLAWS 5* (dont le jeu d'étiquettes est nommé C5) qui nous a permis de passer de 65 étiquettes à douze classes de mots basiques, tel que le montre le *Tableau 4* ci-contre.

La procédure de lemmatisation adoptée s'inspire du travail présenté dans Beale (1987), et fut raffinée subséquemment par le groupe de Lancaster, expérimenté par un large éventail de projets tels que le programme *JAWS* (Logiciel pour déficient visuel, Fligestone *et al*., 1996) ou le livre *Word frequencies in written and spoken english* (Leech *et al.*, 2001). L'approche générale est d'appliquer un certain nombre de règles morphologiques, combinant des règles simples de dépouillement des suffixes POS avec une liste des exceptions les plus courantes. Ceci fut opéré au moment de la conversion XML, en utilisant le code et un ensemble de fichiers de règles gracieusement fournis par Paul Rayson.

Valeur de l'étiquette simplifiée	Signification	Etiquettes détaillées associées
ADJ	adjectifs	AJ0, AJC, AJS, CRD, DT0, ORD
ADV	adverbes	AV0, AVP, AVQ, XX0
ART	articles	AT0
CONJ	conjonctions	CJC, CJS, CJT
INTERJ	interjections	ITJ
PREP	prépositions	PRF, PRP, TO0
PRON	pronoms	DPS, DTQ, EX0, PNI, PNP, PNQ, PNX
STOP	ponctuation	POS, PUL, PUN, PUQ, PUR
SUBST	substantifs	NN0, NN1, NN2, NP0, ONE, ZZ0, NN1-NP0, NP0-NN1
UNC	Non classifiés, incertains ou mots non lexicaux	UNC, AJ0-AV0, AV0-AJ0, AJ0-NN1, NN1-AJ0, AJ0-VVD, VVD-AJ0, AJ0-VVG, VVG-AJ0, AJ0-VVN, VVN-AJ0, AVP-PRP, PRP-AVP, AVQ-CJS, CJS-AVQ, CJS-PRP, PRP-CJS, CJT-DT0, DT0-CJT, CRD-PNI, PNI-CRD, NN1-VVB, VVB-NN1, NN1-VVG, VVG-NN1, NN2-VVZ, VVZ-NN2
VERB	verbe	VBB, VBD, VBG, VBI, VBN, VBZ, VDB, VDD, VDG, VDI, VDN, VDZ, VHB, VHD, VHG, VHI, VHN, VHZ, VM0, VVB, VVD, VVG, VVI, VVN, VVZ, VVD-VVN, VVN-VVD

Tableau 4 : Le jeu d'étiquettes simplifié

2.4 Encodage.

L'élaboration de la TEI (« *Text Encoding Initiative* ») et le modèle initial de balisage utilisé pour le *BNC* ont été définis à la même période (et de surcroît par les mêmes personnes) ; c'est donc sans surprise que les deux schémas sont très proches malgré quelques différences. Ce modèle a été depuis largement employé et bien documenté, nous ne le discuterons pas ici en détails.

En guide d'indice sur l'ampleur et la nature du balisage dans le *BNC*, voici le début d'un texte écrit typique (A0A) :

```
<wtext type="OTHERPUB">
<div level="1">

<head type="MAIN">
<s n="1">
        <w c5="NP0" hw="camra" pos="SUBST">CAMRA </w>
        <w c5="NN1" hw="fact" pos="SUBST">FACT </w>
        <w c5="NN1" hw="sheet" pos="SUBST">SHEET </w>
        <w c5="NN1" hw="no" pos="SUBST">No </w>
        <w c5="CRD" hw="1" pos="ADJ">1</w>
</s></head>

<head rend="it" type="SUB">
<s n="2">
        <w c5="AVQ" hw="how" pos="ADV">How </w>
        <w c5="NN1" hw="beer" pos="SUBST">beer </w>
        <w c5="VBZ" hw="be" pos="VERB">is </w>
        <w c5="VVN" hw="brew" pos="VERB">brewed</w>
</s></head>

<p>
<s n="3">
    <w c5="NN1" hw="beer" pos="SUBST">Beer </w>
    <w c5="VVZ" hw="seem" pos="VERB">seems </w>
    <w c5="DT0" hw="such" pos="ADJ">such </w>
    <w c5="AT0" hw="a" pos="ART">a </w>
    <w c5="AJ0" hw="simple" pos="ADJ">simple </w>
    <w c5="NN1-VVB" hw="drink" pos="SUBST">drink </w>
    <w c5="CJT" hw="that" pos="CONJ">that </w>
    <w c5="PNP" hw="we" pos="PRON">we </w>
    <w c5="VVB" hw="tend" pos="VERB">tend </w>
    <w c5="TO0" hw="to" pos="PREP">to </w>
    <w c5="VVI" hw="take" pos="VERB">take </w>
    <w c5="PNP" hw="it" pos="PRON">it </w>
    <w c5="PRP" hw="for" pos="PREP">for </w>
    <w c5="VVN" hw="grant" pos="VERB">granted</w>
    <c c5="PUN">.</c>
</s>
…</p>…</div>…</wtext>
```

Ce texte démarre avec un élément « wtext » (pour texte écrit) qui porte l'attribut « OTHERPUB » indiquant le type de texte selon la taxonomie simplifiée évoquée ci-dessus. Il se compose ici de différents prospectus publicitaires, correspondant chacun à un élément « div ». Le prospectus représenté ci-dessus débute par deux intitulés, représentés chacun par un élément XML « head » (pour titre), suivi d'une séquence d'éléments « p » (pour paragraphe).

La segmentation du texte opérée avec CLAWS est préservée par les éléments « s », présents dans l'ensemble du texte et individuellement identifiés par un nombre. Au sein de

chaque « s », tous les mots (ou tokens) identifiés par CLAWS sont indiqués par un élément « w » qui contient l'étiquette spécifique attribué par CLAWS C5 (ex. c5="VN"), l'étiquette simplifiée en dérivant (ex. pos="VERB") et la forme lemmatisée de ce mot (ex. hw="grant").

La portée et la signification de ce système de balisage fait l'objet d'une discussion dans le *User References Guide* accompagnant le corpus.

De nombreuses questions techniques ont été abordées au moment du balisage de la partie orale du corpus. Comme nous l'avons noté plus haut, c'était la première fois qu'un marquage détaillé de discours transcrit était tenté à une telle échelle. La transcription elle-même était réalisée par une équipe non spécialiste de linguistique (mais qui était cependant familière des variations régionales devant être retranscrites – une équipe recrutée dans l'Essex, par exemple, n'était pas sensée transcrire un matériau en provenance d'Irlande du Nord). Les transcripteurs ajoutaient une forme minimale (non SGML) de balisage sur le texte, qui était ensuite normalisée, convertie en format SGML, et validée pour les besoins logiciel spécifiques (*cf*. Burnage, 1993). La structure de balisage permet de mettre en évidence un certain nombre de caractéristiques, notamment les changements de locuteurs et le détail des chevauchements, les mots utilisés tels que le transcripteur les percevaient, les faux départs, troncatures ou hésitations, certaines caractéristiques de l'exécution comme les pauses, les mises en scène, etc. En outre, les informations détaillées concernant les locuteurs et contextes de discours ont bien sûr été enregistrées dans l'en-tête, lorsqu'elles étaient disponibles.

Voici un échantillon d'un texte oral retranscrit :

```
<u who="PS04U">
  <s n="1297">
      <w c5="VBZ" hw="be" pos="VERB">Is </w>
      <w c5="PNP" hw="he" pos="PRON">he </w>
      <w c5="XX0" hw="not" pos="ADV">not </w>
      <w c5="VVG" hw="go" pos="VERB">going </w>
      <w c5="AV0" hw="home" pos="ADV">home </w>
      <w c5="AV0" hw="then"  pos="ADV">then</w>
      <c c5="PUN">?</c>
  </s>
</u>

<u who="PS04Y">
  <s n="1298">
      <w c5="ITJ" hw="no" pos="INTERJ">No  </w>
      <pause dur="8"/>
      <w c5="CJC" hw="and" pos="CONJ">and </w>
      <w c5="UNC" hw="erm" pos="UNC">erm  </w>
      <pause dur="7"/>
      <w c5="PNP" hw="i" pos="PRON">I</w>
      <w c5="VBB" hw="be" pos="VERB">'m </w>
      <w c5="VVG" hw="leave" pos="VERB">leaving </w>
      <w c5="AT0" hw="a" pos="ART">a </w>
      <w c5="NN1" hw="turkey" pos="SUBST">turkey </w>
      <w c5="PRP" hw="in" pos="PREP">in </w>
      <w c5="AT0" hw="the" pos="ART">the </w>
      <w c5="NN1" hw="freezer" pos="SUBST">freezer</w>
      <c c5="PUN">,</c>
      <trunc>
      <w c5="UNC" hw="an" pos="UNC">an </w></trunc>
      <w c5="NP0" hw="paul" pos="SUBST">Paul </w>
      <w c5="VBZ" hw="be" pos="VERB">is </w>
      <w c5="AV0" hw="quite" pos="ADV">quite </w>
      <w c5="AJ0" hw="good" pos="ADJ">good </w>
      <w c5="PRP" hw="at" pos="PREP">at </w>
      <w c5="VVG-NN1" hw="cook" pos="VERB">cooking  </w>
```

```
      <pause/><align with="KBFLC07U"/>
      <w c5="AJ0-NN1" hw="standard" pos="ADJ">standard </w>
      <w c5="NN1" hw="cooking" pos="SUBST">cooking</w>
      <c c5="PUN">.</c>
    </s>
  </u>
```

Les mots et phrases sont étiquetés de la même manière que pour les textes écrits, tel que décrit ci-dessus. Les phrases sont cependant ici groupées par déclaration (ou flux de discours). Chaque déclaration est balisée par un élément « u » (pour « utterance » / paroles), qui marque donc une partie continue de discours, et qui comporte en son sein un attribut « who » dont la valeur donne la clef pour accéder à des informations plus détaillées concernant le locuteur, enregistrées dans l'en-tête TEI du texte de la retranscription (ex. u who="PS04Y"). Les éléments « pause » et « event » (événement) sont aussi utilisés pour indiquer toute caractéristique paralinguistique du discours ainsi retranscrit.

Comme le montre l'exemple ci-dessus, l'intention du transcripteur était de fournir une version du discours plus proche de l'écrit, que du signal audio immédiat. L'indication orthographiée des différentes pauses fut ainsi normalisée (« *erm* », « *mmm* »), de même que celle de l'intonation interrogative que des caractères conventionnels de ponctuation servent à marquer comme questions de manière constante. Pour une étude plus détaillée de la rationalité sous-jacente à ces pratiques, ainsi que d'autres aspects de la transcription de discours, voir Crowdy (1994).

2.5 Logiciels et distribution.

En 1994, la manière de diffuser, à des fins non-lucratives, un corpus de la taille du *BNC* n'allait pas de soi. L'envergure des données excluait d'office l'usage d'options à bas coût telles que les protocoles de communication anonymes (ftp). Notre décision initiale fut de distribuer une version compressée du corpus, tenant sur un ensemble de trois CDs, avec un logiciel d'exploitation simple qui pouvait être installé par un personnel raisonnablement qualifié, afin de permettre à chaque individu participant au réseau d'obtenir une copie locale du corpus. Le développement d'un tel système fut entrepris la dernière année du projet grâce à des fonds supplémentaires de la *British Library*. Le logiciel qui est maintenant fourni avec le corpus est connu sous le nom de *XAIRA* (*XML Aware Indexing and Retrieval Architecture*) et dérive de l'outil initial nommé *SARA* (*SGML Aware Retrieval Application*). *XAIRA* a été développé, grâce à des fonds de la fondation Andrew Mellon, dans le but de fournir un outil *open source* au format XML, destiné à l'étude de corpus de langue de toute taille (pour plus de détails, http://www.xaira.org).

Il était évident depuis le départ que l'accès au *BNC* ne devait pas nécessiter l'usage d'un logiciel spécifique – c'était, après tout, ce que sous-tendait l'usage de la norme internationale SGML pour encoder le corpus original, plutôt que de créer un système taillé pour un outil logiciel spécifique.

Le *BNC* précède l'ère du *World Wide Web*[9]. Cependant, au cours de sa première année de diffusion, il est apparu que le Web serait la voie de mise à disposition idéale, ne serait-ce que pour permettre aux chercheurs non Européens d'y avoir accès, ce que les restrictions de licences leur empêchaient jusque-là. À cette fin, La *British Library* a généreusement offert au projet un serveur, et une interface Web du corpus fut alors développée. Ce service, toujours

[9] La locution « *World Wide Web* » n'apparaît en effet qu'à deux reprises dans le corpus et toutes deux dans un bref échange, se dé**r**oulant dans une liste de discussion par courriels en janvier 1994, au sujet de la promotion d'un club de football, The Leeds United Football Club. Les plus fréquents collocats du mot « *web* » dans le corpus sont « *spider* » (*araignée*), « *tangled* » (*enchevêtré*), « *complex* » (*complexe*) et « *seamless* » (*sans faille*). De ce point de vue, au moins, le *BNC* ne peut décidément pas être considéré comme un miroir de la langue anglaise actuelle.

disponible à l'adresse http://sara.natcorp.ox.ac.uk permet à quiconque de réaliser des recherches élémentaires sur le corpus, avec des options de visualisation restreintes. Les personnes souhaitant davantage de possibilités et des fonctions plus sophistiquées peuvent également télécharger une copie de l'interface client du programme *SARA* qui permet d'accéder au même serveur : est uniquement demandée une modeste contribution financière qui permet de proroger une première période d'essai gratuite.

Pour compléter ce service, et répondre à une forte demande d'aide dans l'utilisation du *BNC* de la part d'enseignants, un tutoriel détaillé fut élaboré (Aston, 1999) afin d'initier tout à chacun aux diverses possibilités et fonctionnalités du logiciel, grâce à des exercices orientés sur la linguistique. Le service en ligne demeure encore très populaire et reçoit chaque mois plusieurs milliers de requêtes.

3 LES RÉVISIONS SUCCESSIVES DU *BNC*

Comme nous l'avons déjà indiqué ci-dessus, le *BNC* n'a jamais été conçu pour être un corpus de suivi (monitor corpus). Néanmoins, il fut l'objet de deux révisions majeures depuis sa première apparition, et une troisième pourrait être jugée nécessaire dans le futur si la demande reste aussi forte pour cet instantané détaillé et unique de la langue anglaise. Dans cette partie, nous parlerons de certains des changements apportés au corpus, et des principes généraux délimitant le champ des possibles.

La deuxième édition du *BNC*, aussi connue sous le nom de *BNC World Edition* a été publiée en décembre 1999, cinq années après la première apparition du *BNC*. Afin que le corpus puisse être diffusé au niveau mondial, seul un petit nombre de textes (moins de cinquante), pour lesquels l'obtention de droits d'usage internationaux s'est avérée impossible, ont été enlevés du corpus

Bien que souhaitable, l'envergure du *BNC* excluait toute relecture complète des épreuves. Essayer de corriger les erreurs dans le *BNC* reviendrait à balayer entièrement le sable d'une plage, comme cela a été imaginé par le Morse et le Charpentier de *Alice in Wonderland* :

> *"If seven maids with seven mops / Swept it for half a year / Do you suppose," the Walrus said, / "That they would get it clear?" / "I doubt it," said the Carpenter, / And shed a heavy tear.* [10]

Dans un certain sens, toute transcription de texte oral est inévitablement approximative. Même pour les textes écrits, décider de ce qui est ou non une erreur n'est pas toujours évident : des mots mal orthographiés peuvent apparaître dans le matériau publié, et l'on pourrait en conséquence s'attendre à ce qu'ils apparaissent tel quel dans un corpus. Quand des corrections ont eu lieu dans la phase d'élaboration du corpus, les erreurs furent parfois l'objet d'une indication en balise, de manière à préserver l'erreur initiale et sa correction. Ceci nous fournit au moins des indications sur le type d'erreurs qu'il est possible de rencontrer. Il est cependant impossible d'expertiser sérieusement l'importance de telles erreurs, ni même d'en repérer l'origine exacte du fait de la variété des traitements imposés aux textes sources. En principe, il est impossible de distinguer une erreur résultant (par exemple) d'un logiciel de reconnaissance optique de caractère (OCR) non approprié, d'une erreur déjà présente dans le texte original, à moins bien sûr de procéder à une expertise minutieuse et comparative des épreuves et des textes sources[11]; quant à l'utilisation de correcteurs orthographiques automatiques, elle n'aura pour conséquence que d'assombrir un peu plus une eau déjà trouble.

[10] Poème de Lewis Caroll (le Morse et le Charpentier marchent au bord de la plage : « "Si sept femmes de chambre, de sept balais armées / Le balayaient durant, tout entière, une année, / Supposes-tu, s'enquit ingénument le Morse, / Qu'elles viendraient à bout de ce tas de poussière ?" / "J'en doutes", répondit le Charpentier, / Tout en laissant couler quelques larmes amères. »)

[11] Un tel travail reste cependant possible puisque l'intégralité des matériaux textuels originels est conservée au service informatique de Université d'Oxford (OUCS).

Et pourtant, un certain type de correction automatique était possible et fut appliqué au *BNC World Edition*. En grande partie grâce à la fiabilité de l'échantillonnage du *BNC*, les règles utilisée par *CLAWS* ont pu être notablement perfectionnées, ce qui a permis de réduire de manière significative le taux d'erreurs et le degré d'indétermination des étiquettes POS de cette édition du *BNC*. Ce travail, effectué à Lancaster avec des fonds du Ministère Britannique de la Recherche en Ingénierie et Sciences Physiques (Financement de recherche n° GR/F 99847), est décrit en détail dans le manuel fournit avec le corpus (Leech, 2000). Ce manuel estime que, à la suite de cette opération de traitement automatique, le taux d'erreur pour tous les mots a été réduit à environ 1,15%, tandis que la proportion de codes ambigus est réduite à approximativement 3, 75%[12].

Dans le même temps, un certain nombre d'erreurs systématiques furent identifiées. Cette gamme de doublons ou erreurs de balisages a servi de base à une vérification complète des données démographiques associées aux locuteurs de chaque texte oral, partie du corpus dans laquelle nombres d'erreurs furent identifiées pour la première version du *BNC*.

La troisième édition du *BNC*, connue sous le nom de *BNC XML Edition*, est parue en mars 2007. À notre grand étonnement, alors que les technologies sur lesquelles il reposait avaient changé quasiment du tout au tout, la demande pour le *BNC* n'a montré aucun signe de faiblesse au cours des cinq années qui suivirent la sortie du *BNC World Edition* : le SGML avait laissé place au XML comme langage de codage, tandis que l'arrivée de l'Unicode et du *World Wide Web* résolvaient nombre de problèmes soulevés dans les années 1990, notamment en rendant possible une autre organisation logicielle. Dans les années 1990, les pratiques usuelles de développement de logiciel conduisaient à élaborer des applications monolithiques, sophistiquées mais idiosyncrasiques ; dans les années 2000, les concepteurs universitaires de logiciels peuvent se reposer sur quelques centaines d'utilitaires de coopérations à petites échelles et d'outils développés pour des objectifs divers sur des interfaces normalisées.

Pour l'édition XML du *BNC*, la documentation concernant de la structure d'annotation, le schéma utilisé pour la valider et le balisage lui-même ont été entièrement revus. Les objectifs premiers de cette révision étaient :

- réduire la complexité du balisage, en particulier en supprimant les comportements indiqués de manière incohérente ou rare ;
- améliorer l'utilisation du corpus avec des outils XML génériques en utilisant seulement les caractéristiques normées du XML ;
- accroître la conformité de la structure d'annotation utilisée avec les normes internationales telles que la TEI.

La conversion du corpus de SGML à XML fut réalisée de manière quasi-automatique ; mais avec des balises plus malléable et accessibles, il apparaissait plus aisé d'aborder certains des moins fréquents ou des plus excentriques aspects du balisage du *BNC*.

Le *BNC World Edition* indiquait par exemple une correction éditoriale du corpus de deux manières différentes, quand cette indication était présente. Aucune tentative d'harmonisation ne fut engagée vis-à-vis de la description des caractéristiques extra-linguistiques des textes oraux ou de celle des codes employés pour caractériser des traits saillants dans les textes écrits. Nous avons tenté d'aborder ces inconvénients par une homogénéisation de ces descriptions, en recodant par exemple « *baby cries* » (*pleurs de bébé*), « *baby screams* » (*cris de bébé*), « *baby noise* » (*bruits de bébé*), etc. en un unique « *baby cries* » (*pleurs de bébé*). Pour prendre un autre exemple, le pays Bahreïn (« *Barhain* » en anglais) figure 174 fois dans le *BNC* ; en 9 occurrences, il est orthographié Bahrein. Sur ces 9 orthographes erronées, seules deux sont balisées comme erreur, de façons différentes qui plus est. Ce genre de

[12] Ces estimations ont été obtenues après l'examen manuel d'un échantillon de 50 000 mots prélevé sur le corpus, ainsi que le précise le manuel cité.

pseudo-précision contribue simplement à plonger l'utilisateur dans la confusion. Nous avons aussi tenté de simplifier plusieurs aspects du balisage, en particulier la façon dont étaient représentés les discours simultanés.

À chaque nouvelle révision du BNC, l'environnement type d'utilisation du corpus était profondément modifié. À l'époque du *BNC World*, il semblait évident que le corpus devait maintenant pouvoir être installé à bas coût, pour un usage personnel, sur un simple ordinateur équipé d'un système d'exploitation *Windows* quel qu'il soit : nos licences et notre politique de distribution ont été corrigées pour rendre ceci possible. Avec le *BNC XML*, cette évolution se confirme, mais en modifiant le logiciel permettant d'accéder au corpus, nous avons également tenté de prendre en compte le mode de distribution qui caractérise l'informatique basée sur le Web. La dernière version de *XAIRA* a été conçue pour supporter les prestations du Web et ainsi, cet accès au corpus peut facilement s'intégrer à d'autres applications basées sur le Web. Les contraintes des licences limitent certes le domaine des possibles, mais on peut désormais de manière sûre et sans plus d'effort obtenir très rapidement un aperçu du corpus sur d'autres applications basées sur le Web. Ceci poursuit l'orientation initiée par le développement du service en ligne du *BNC*, par une disponibilité accrue du corpus pour une communauté d'utilisateurs toujours plus large.

4 LES LECONS DE L'EXPÉRIENCE DU *BNC*

Tout le monde sait qu'il faut faire une étude de marché avant de commencer la distribution d'un projet, en particulier quand on atteint le niveau d'investissement initial requis par le *BNC*. Mais, comme tant d'autres choses que tout le monde est censé savoir, cette sagesse populaire s'avère finalement trompeuse dans le cas du *BNC*. Quand les partenaires initiaux du projet discutèrent du marché potentiel des copies du *BNC*, il semblait évident qu'il ne pouvait qu'être restreint en taille. Dans le milieu des années 1990, il était manifeste que seule une communauté de chercheurs spécialisés en traitement automatique de la langue naturelle, TAL, et, bien entendu, les départements de R&D des entrepreneurs engagés dans le TAL ou la lexicographie pouvaient être, au moins, intéressés par une collection de cent millions de mots d'anglais, sous une forme alors encore dénommée de « lisible par ordinateur » (*machine-readable*). L'évidence de ce modèle se reflétait dans les options retenues pour le cadre juridique de diffusion du corpus, comme pour les méthodes de distribution : la licence que tout acquéreur potentiel devait signer (en deux exemplaires) mentionne par exemple des « groupes de recherche » et s'oppose farouchement à tout contrôle d'un usage en réseau du corpus à l'intérieur d'une institution – mais nulle part n'y est envisagée la possibilité d'un achat pour un usage individuel ou avec un groupe d'étudiants.

Les faits nous ont cependant rapidement démontré que le marché en question était bien plus important et de nature très différente. Les principaux utilisateurs du *BNC* se révélèrent être des personnes travaillant en linguistique appliquée et non dans les linguistiques computationnelles, en particulier des individus que concernaient l'apprentissage et l'enseignement de la langue. Leurs compétences informatiques étaient moindres que prévues, mais leur enthousiasme sans commune mesure plus vaste. Ils comprenaient non seulement des linguistes computationnels ou des chercheurs en TAL, mais également des historiens de la culture, et même des apprenants en langue.

Rétrospectivement, le projet *BNC* a connu les mêmes angles morts technologiques que d'autres projets de la même époque. Bizarrement, nous ne nous attendions pas du tout au succès rencontré par la révolution du XML ! Nous avons donc perdu énormément de temps en conversion de formats et compromis. De la même façon, parce que nous n'avions pas prévu que les ordinateurs indépendants, fonctionnant à 1 Ghz avec des disques durs de 20 Gigabytes, deviendraient la norme de l'équipement des foyers, nous n'avons pas anticiper le fait qu'un jour il serait possible de stocker, avec leurs retranscriptions, les versions audio

numériques des textes que nous transcrivions. De ce fait, nous ne nous sommes jamais demandé s'il serait utile d'essayer d'obtenir des droits pour distribuer ces versions audio. De même nos efforts de développement logiciel se sont focalisés sur le développement d'une application client/serveur, un système basé sur l'hypothèse que l'utilisation du *BNC* serait caractérisée par une unique ressource informatique partagée permettant plusieurs points d'entrée, plutôt que par une massive reproduction sur des machines individuelles.

Quelles autres opportunités n'avons-nous pas saisies ? Dans la conception initiale, il y a clairement un perceptible déplacement de la notion de « représentativité » à l'idée du *BNC* comme d'une source de corpus de spécialités. D'un échantillon de l'ensemble de la langue, le *BNC* s'est rapidement mis en place comme dépositaire de la variété de la langue. C'était, de façon rétrospective, un repositionnement assez sensible : il est difficile d'envisager la compilation de matériaux plus divers que le *BNC*. Un rapide coup d'œil à des listes de discussion se rapportant aux corpus montre que les questions les plus fréquentes sont des questions du type « *Je suis en train de chercher un ensemble de textes de type X* » (j'ai noté pour les questions les plus récentes que ce X concerne des interactions docteur / patient, les débats juridiques, les controverses, le flirt…) ; dans presque tous les cas, la réponse à une telle question est « i*l y a ça oui, quelque part dans le BNC, mais c'est à vous de la trouver…* ». La version XML du corpus rend ce genre de démarches plus aisé, en assurant un meilleur accès à une série de méta-données qui peuvent être recherchées en même temps que le contenu textuel lui-même. Nous pouvons par exemple repérer les débuts de romans, catégorisés comme « fiction », en cherchant un échantillon de textes intégrant des sections d'ouverture ; ou nous pouvons sélectionner des sessions de formations en recherchant ces mots dans le titre des textes catégorisés comme discours en contexte normé.

De façon très claire, la conception du *BNC* est totalement passée à côté de l'opportunité de monter un formidable corpus de suivi, l'un de ceux qui aurait pu saisir l'écoulement du fleuve de la langue et évoluer au fil du temps. Ce serait particulièrement déprimant si les linguistes de ce siècle continuaient à étudier la langue des années 1990 durant autant de temps que, ceux qui les ont précédés, ont eux-mêmes étudié la langue des années 1960. Néanmoins, quand bien même cela serait évidemment intéressant, les chances restent très maigres d'obtenir les fonds nécessaires pour construire une série de corpus semblables au *BNC*, à intervalles réguliers, disons toutes les décennies. En revanche, nous pouvons déjà prévoir que dans un futur proche, nous aurons à disposition différents corpus de langue d'une toute autre échelle. Comment gérer au mieux la diversité et l'imprévisibilité du Web pour en faire notre source future d'information linguistique est une autre, et tout à fait différente, histoire.

5 BIBLIOGRAPHIE

Aston G. and Burnard L. (1998). *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh : Edinburgh University Press.

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

Atkins B.T.S. & Zampolli A. (1994). *Computational Approaches to the Lexicon*, Oxford : Oxford University Press.

Burnage G. & Dunlop D. (1992). « Encoding the British National Corpus », in : Aarts *et al.* (eds). *English language corpora: design, analysis and exploitation*, Amsterdam : Rodopi, p 79-95

Burnard L. (ed) (1995). *Users' Reference Guide for the British National Corpus, version 1.0*. Oxford : Oxford University Computing Services.

Burnard L. (1999). « Using SGML for linguistic analysis : the case of the BNC », in : *Markup languages theory and practice*, vol. I.2, Cambridge, Mass: MIT Press, p. 31-51 (également publié en 2001 dans *Maschinelle Verarbeitung altdeitscher Texte*, vol. V, Tuebingen : Max Niemeyer, p. 53-72).

Carpenter L., Shaw S. & Prescott A. (eds) (1998). *Towards the Digital Library: the British Library's Initiatives for Access programme*, London : British Library.
Clear J. H. (1993). « The British National Corpus », in : Delany P. & Landow G. (ed). *The Digital Word : text-based computing in the humanities*, Cambridge (Mass) : MIT Press, p. 163-187.
Crowdy S. (1994). « Spoken Corpus Transcription », *Literary & Linguistic Computing*, vol. 9:1, p. 25-28.
Crowdy S. (1995). « The BNC spoken corpus », in : Leech G., Myers G. & Thomas J. (eds). S*poken English on computer: transcription, mark-up and application*, Harlow : Longman, p. 224-235.
Dunlop D. (1995). « Practical considerations in the use of TEI headers in large corpora », in : Ide N. & Veronis J. (eds). *Text Encoding Initiative: background and context*, Dordrecht : Kluwer, p. 85-98.
Garside R. (1995). « Grammatical tagging of the spoken part of the British National Corpus: a progress report », in : Leech G., Myers G. & Thomas J. (eds). *Spoken English on computer: transcription, mark-up and application*, Harlow : Longman, p. 161-167.
Garside R. (1996). « The robust tagging of unrestricted text : the BNC experience », in : Thomas J. & Short M. (eds). *Using corpora for language research: studies in the honour of Geoffrey Leech*, Harlow : Longman, p. 167-180.
Garside R., Leech G. & McEnery T. (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London : Longman, chapters 7-9.
Ide N., Priest-Dorman G. & Véronis J. (1996) *Corpus Encoding Standard*. [disponible sur http://www.cs.vassar.edu/CES/]
Lee D. (2001) « Genres, registers, text types and styles : clarifying the concepts and navigating a path through the BNC Jungle », *Language Learning and Technology*, vol. 5.3 [disponible sur http://llt.msu.edu/]
Leech G., Garside R. & Bryant M. (1994). « CLAWS4: The tagging of the British National Corpus », in : *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto : COLING 94, p. 622-628.
Leech G & Smith N. (2000) *Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging*, Lancaster : UCREL [disponible sous forme numérique comme élément du *BNC World Edition*].
Lehmann H., Schneider P. & Hoffmann S. (1999). « BNCweb », in : Kirk J. (ed). *Corpora galore: analysis and techniques in describing English*, Amsterdam : Rodopi, p. 259-266.
Renouf A. (1986) « Corpus development at Birmingham University », in : Aarts J. & Meijs W. (eds). *Corpus Linguistics II : New Studies in the Analysis and Exploitation of Computer Corpora*, Amsterdam : Rodopi, p. 7-23
Sinclair J. McH. (1987). *Looking Up*. London : Collins.
Sperberg-McQueen C.M. & Burnard L. (1994). *Guidelines for electronic text encoding and interchange* (TEI P3), Chicago et Oxford : ACH-ALLC-ACL Text Encoding Initiative.
Zampolli A., Calzolari N. & Palmer M. (eds) (1994). « Current Issues », in : *Computational Linguistics: In Honour of Don Walker* (*Linguistica Computazionale* IX-X), Pisa : Giardin