Marc Van Campenhoudt
Centre de recherche Termisti
Institut supérieur de traducteurs et interprètes (ISTI)
Bruxelles

LINGUISTIQUE DE CORPUS ET ÉTUDE DES VOCABULAIRES SPÉCIALISÉS

PARIS VIII, 8 JANVIER 2002

1 LA FIN D'UN DÉNI THÉORIQUE

La théorie viennoise de la terminologie n'a jamais accordé une grande attention au dépouillement de textes spécialisés. Eugen Wüster (1979) a clairement affirmé dans l'introduction de l'*Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* - connue en français comme la *théorie générale de la terminologie* - que la terminologie était une science distincte de la linguistique. Parmi les « différences d'attitude fondamentales », il citait l'utilisation des concepts comme points de départ et le refus de la description linguistique au profit de la « norme prescriptive ». Ces deux points de vue expliquent aisément l'absence caractérisée d'intérêt pour la linguistique de corpus, une discipline attachée à la description de faits de langue observés dans l'usage réel, écrit ou oral.

1.1 Une trop lente évolution de la formation

Cet état d'esprit a longtemps marqué la plupart des manuels d'initiation à la terminologie : on y trouve, certes, des développements sur la recherche documentaire et la documentation spécialisée, sur la morphologie des termes dans l'une ou l'autre langue, mais les liens explicites entre ces deux sujets y semblent exceptionnels. L'expression même « linguistique de corpus » n'y trouve guère sa place et ce serait en vain que l'on y rechercherait un exposé sur le balisage des textes, une initiation aux principes de la statistique lexicale et aux méthodes d'identification des figements ou encore une présentation des principales fonctionnalités offertes par les concordanciers[1].

Le premier manuel à intégrer de réels développements sur la linguistique de corpus est le second volume, récemment paru, *du Handbook of Terminology Management,* sous-titré : *Applications Oriented Terminology Management* (Budin & Wright 1997-2001 : 725-844). Ce manuel arrive assurément à point nommé et aborde longuement l'exploitation de corpus en terminologie. S'agissant d'un ouvrage collectif, chacun des cinq chapitres consacrés à cette matière est rédigé par un spécialiste différent, ce qui ne garantit pas toujours - en dépit des efforts - une cohésion didactique idéale. Il reste que les auteurs manifestent une volonté de se situer dans le cadre de la linguistique appliquée et un souci de dispenser un enseignement qui ne dépende pas des spécificités d'un logiciel particulier.

---

1. Formulant cette critique, nous devons reconnaître qu'elle s'applique largement aux manuels de langue française d'initiation à la lexicologie.

1.2 Un hiatus de plus en plus évident

La longue absence de la linguistique de corpus dans les manuels de terminologie ne nous paraît guère surprenante. L'irruption des outils informatiques de dépouillement des corpus sape les bases mêmes de la théorie générale de la terminologie : il semble difficile d'intégrer ces « nouveautés » sans réviser sérieusement la doctrine initiale ! Notre sentiment d'enseignant-chercheur est que l'on attend du jeune terminologue qu'il manie de nouveaux outils logiciels, les extracteurs de candidats termes, sans qu'il ait été préalablement formé aux approches linguistiques qui les sous-tendent. Alors que les cadres théoriques de référence - Wüster, l'école de Vienne - semblent demeurer incontournables, le terrain de l'ingénierie linguistique appliquée aux langues spécialisées a depuis plus d'une décennie été investi par des universitaires venus des sciences du langage. Formés aux différentes disciplines de la linguistique appliquée, voire à la programmation informatique, ils ont appliqué aux langues spécialisées les méthodes de recherche de la linguistique descriptive. Leurs travaux débouchent aujourd'hui sur la commercialisation de logiciels - plus ou moins performants - dont l'usage n'était pas envisagé par l'approche viennoise. La suite logique est une ferme remise en cause du modèle viennois au nom de la réalité tangible des vocabulaires spécialisés, telle qu'elle peut être observée dans l'usage réel.

1.3 La réappropriation par la linguistique

Si les sciences du langage ont elles-mêmes connu un mouvement de dédain pour les corpus, notamment dans la lignée des travaux de Noam Chomsky, les recherches de linguistique descriptive fondées sur l'observation de l'usage réel connaissent depuis le début des années 90 un net regain d'intérêt dans le cadre du traitement automatique des langues naturelles (Habert *et al.* 1997 : 7-11).

Il est intéressant de noter que les écrits théoriques qui conduisent à reconsidérer l'ancienne vision viennoise au nom de la réalité de l'usage sont souvent le fait d'enseignants-chercheurs qui, comme Jennifer Pearson (1998) ou Rita Temmerman (2000a), forment des étudiants en traduction. Beaucoup d'écoles de traduction ont vu leur cadre professoral se renouveler au profit d'une nouvelle génération issue de la linguistique appliquée. Alors que leurs prédécesseurs formaient le traducteur à compiler des dictionnaires pour alimenter des fiches terminologiques ne comportant guère que des équivalents, la « jeune école » insiste bien davantage sur la nécessité d'observer l'usage réel dans des corpus textuels. La mise en place des autoroutes de l'information rend, il est vrai, cette approche beaucoup plus aisément envisageable que par le passé.

Si l'on peut se féliciter du mouvement de retour vers l'usage réel - en tout cas l'usage écrit -, il convient de souligner que celui-ci comporte toutefois un danger, qui est celui d'un dédain pour la gestion de bases de données terminologiques. Il importe que le mouvement de retour au texte s'accompagne d'une réflexion sur la manière de concevoir des dictionnaires électroniques qui rende compte de toute la richesse de l'usage.

2 DES CORPUS

2.1 Quelques définitions de la notion de corpus

Ambiguïté originelle : recueil d'énoncés ou de textes, échantillon (représentatif ?) ou corpus saturé ?

« Ensemble d'énoncés d'une langue donnée (écrits ou oraux enregistrés) qui ont été recueillis pour constituer une base d'observation permettant d'entreprendre la description et l'analyse de la langue en question. » (Arrivé, Gadet et Galmiche 1986.)

« Si l'on veut étudier tels ou tels phénomènes dans une langue naturelle, il faut recueillir les *corpus* correspondants. Ce sont des ensembles d'énoncés (écrits ou oraux, selon les besoins) que le linguiste pose comme un échantillon représentatif de faits de parole (au sens saussurien) des locuteurs de cette langue. » (Chiss, Filliolet et Maingueneau 1993 : 61.)

« A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as sample of language » (Sinclair 1994a : 2.)

CORPUS [13c: from Latin *corpus* body. The plural is usually corpora]. (1) A collection of texts, especially if complete and self-contained: *the corpus of Anglo-Saxon verse*. (2) Plural also *corpuses*. In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analysed by means of *tagging* (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. *Corpus linguistics* studies data in any such corpus ... T.McA. (*The Oxford Companion to the English Language*, ed. McArthur & McArthur 1992)

« La réunion d'un grand nombre de textes indexés constitue une **corpus**, à l'intérieur duquel on se propose d'étudier et de quantifier certains faits lexicaux, syntaxiques, etc. Le corpus comprend en général des divisions ou **sous-corpus**, qui la plupart du temps ont une unité propre (chronologique, stylistique, etc. » (MULLER 1973 : 16.)

2.2 Critiques à l'encontre de la linguistique de corpus

- Peu de corpus oraux disponibles
- Mythe du corpus objectif et homogène ; négligence des facteurs extralinguistiques : conditions de production, incidence du locuteur, contexte, variables sociales, régionales, économiques…

➔ Nécessité de s'interroger sur les critères de constitution des corpus

Certains linguistes, comme Chomsky, ont jadis rejeté l'utilisation des corpus au profit de l'intuition du linguiste, ce qui rappelle l'idée saussurienne du locuteur homogène dans une société homogène. Loin de l'empirisme dont font preuve certains, l'observation d'énoncés attestés dans un environnement textuel participe en réalité d'une démarche scientifique qui s'avère rigoureuse dès lors qu'elle prend en considération les paramètres de l'expérimentation et en relativise les résultats.

L'irruption de l'informatique dans la majorité des centres de recherche en linguistique et l'accessibilité récente d'un grand nombre de textes électroniques via Internet explique sans doute la grande percée actuelle de la linguistique de corpus.

3 QUELQUES CRITÈRES DE LIMITATION ET SÉLECTION DU CORPUS EN LANGUE SPÉCIALISÉE (PEARSON 1998, BOWKER 1998)

Les critères se recoupent très souvent. Leur énoncé peut donc paraître quelque peu redondant.

3.1 Taille

- Taille mythique : environ un million de mots en langue spécialisée (LSp), beaucoup plus en langue générale (LG).

- Nécessité de toujours faire évoluer et donc d'agrandir le corpus.

- La représentativité est plus importante que la taille. Même dans un grand corpus, on doit pouvoir isoler des sous-corpus en fonction de critères précis.

- La taille dépend d'abord des matériaux disponibles (sous-domaine très pointu, déficit lexical dans une langue) et des objectifs de la recherche.

3.2 Domaine et sous-domaine

- Cerner l'appartenance des textes au domaine et aux écoles au sein du domaine en dialogue avec des experts.

- Plus le sujet est pointu, plus la sélection est restrictive, plus la taille du corpus sera limitée.

3.3 Genre

Le genre du texte est largement lié au public auquel le texte est destiné : articles ou contributions destinées aux experts, aux spécialistes, aux étudiants du domaine, à une large vulgarisation ?

3.4 Échantillonnage ?

Il importe d'être très prudent avant de travailler sur des extraits. Un texte doit normalement être considéré dans sa totalité : il est fréquent que l'auteur clarifie son vocabulaire au fur et à mesure qu'il progresse. Il serait donc dommage d'en perdre une partie.

3.5 Diffusion

On préfère normalement travailler avec des textes destinés à être publiés, même s'ils sont seulement diffusés auprès d'un public restreint. Ce critère garantit un minimum de qualité rédactionnelle des textes et en valide la pertinence auprès des utilisateurs du corpus.

3.6 Auteurs

Il importe que les auteurs retenus soient reconnus comme experts par leurs pairs. Il convient de vérifier leur réputation. Un diplôme et un emploi dans le domaine de spécialité constituent également des critères importants.

3.7 Technicité

Le degré de technicité dépend des compétences de l'auteur et du public qu'il vise :

- technique : écrit par une spécialiste pour des spécialistes
- semi-technique : écrit par un spécialiste pour un public spécifique (initié, en cours d'initiation ou non initié)

3.8 Public

- Même degré d'expertise que l'auteur

- Degré d'expertise inférieur à celui de l'auteur : professionnels initiés du même secteur, étudiants de la même discipline, néophytes

Chaque situation peut avoir son intérêt : le jargon se recherchera plutôt dans des textes destinés aux experts, les contextes définitoires dans les autres catégories.

3.9 Objectifs visés

- « Stipulatifs » : normalisation, réglementation, cahier des charges…

- Didactiques

3.10 Cadre, contexte

- Institutionnel

- Éducatif, universitaire

3.11 Langues considérées

Une langue peut-être déficitaire pour un domaine concerné.

Les textes traduits doivent normalement être écartés d'office (problématique de l'alignement). Seul un critère relatif à la qualité du traducteur (expert du domaine, membre du service de traduction d'un organisme de référence pour le domaine) peut justifier de rares exceptions.

3.12 Conserver la mémoire des critères

Tout corpus doit être accompagné de documents permettant de garder la mémoire de son mode de constitution et de balisage. Voir les directives de la *Text Encoding Initiative* en la matière.

4 INTERNET ET LA DISPONIBILITÉ DES CORPUS

Il est aujourd'hui aisé de trouver des corpus d'excellente qualité sur la toile, sinon de considérer la toile comme un vaste corpus… Les progrès techniques semblent même parfois prendre les chercheurs de cours, rendant très vite caduque toute réflexion en la matière. Songeons aux écrits d'il y a à peine six ans concernant la reconnaissance optique…

L'un des exposés les plus systématiques concernant la typologie et la constitution des corpus spécialisés est celui proposé par Jennifer Pearson (1998 : 41-65). Sa pérennité est assurée par l'absence de prise en compte d'Internet, l'auteur présentant des corpus déjà constitués dans le cadre de projets précédents.

Malgré la nature labile des sites Internet et la rapide évolution des supports, il nous paraît intéressant de tenter de fournir ici un aperçu de la disponibilité des corpus sur la toile :

4.1 Recherche de corpus préconstitués

On trouve sur la toile un nombre croissant de sites mettant à la disposition des linguistes des corpus préconstitués, éventuellement déjà balisés, voire étiquetés, mais qui ne concernent pas nécessairement les langues spécialisées. Selon les cas, l'accès est payant ou non.

> Quelques sites de départ[2] : MULTEXT, ELRA, SILFIDE, UCREL, ELSENET, *The British National Corpus, The Oxford Text Archive, Corpora List Archive in Hypermail,* INTRATEXT…

2. Nous fournissons les adresses de ces sites dans l'*Infoport de la terminologie* (www.termisti.refer.org/infoport.htm).

4.2 Recherche de textes spécialisés pour alimenter un corpus

De manière générale, les sites institutionnels (organismes officiels, organisations non gouvernementales, universités…) permettent de télécharger un nombre croissant de textes de volume plus ou moins important. Les sites des entreprises ne sont pas à négliger non plus : on y trouve, par exemple, de plus en plus fréquemment de la documentation technique (modes d'emploi, notices d'entretien). Quant aux grands groupes financiers (transport, énergie, finance, assurances…), ils diffusent souvent leur documentation publique sur leurs sites, notamment leurs rapports annuels.

À titre d'exemple, une personne qui souhaite constituer un vaste corpus sur la protection de l'environnement maritime contre la pollution pourra télécharger les textes législatifs de l'Union européenne, les archives d'organisations de protection de la nature comme *Greenpeace*, les brochures des groupes pétroliers vantant leur respect du milieu marin ou tel rapport de recherche de l'IFREMER…

L'application des principes de la critique des documents et la lecture de Jennifer Pearson (1998) devraient permettre au chercheur d'assembler en quelques jours une documentation jadis inespérée et dont la qualité rédactionnelle est souvent appréciable.

4.3 Recherche de textes spécialisés multilingues à aligner

L'engouement pour l'alignement de corpus conduit à rechercher des sites susceptibles d'offrir les traductions d'un même texte dans plusieurs langues. Les chances de trouver un tel document sont beaucoup plus importantes sur les sites d'entreprises multinationales ayant « localisé » leurs pages *web* ainsi que sur les sites d'organismes officiels qui doivent respecter des obligations légales de multilinguisme. Pour ce dernier cas, on citera :

- anglais-français : institutions dépendant du gouvernement fédéral canadien ;
- français-néerlandais : institutions dépendant du gouvernement fédéral belge ;
- français-allemand-italien : institutions dépendant de la chancellerie suisse ou de cantons bilingues ;
- langues de l'Union européenne : serveurs dépendant de l'Union européenne ou d'organismes s'y rattachant.

On pourra également visiter les sites des organisations internationales actives dans le domaine de spécialité étudié, pour autant que leurs langues officielles soient bien celles qui sont concernées par la recherche terminologique.

> Quelques exemples : Bibliothèque virtuelle de la F.A.O., Centre de documentation de l'ONU, *Oceans and Law of the Sea Home Page* (ONU), Organisation internationale du travail, Fonds monétaire international, Union internationale des télécommunications, Organisation maritime internationale, Organisation mondiale de la santé…

Le principal problème soulevé par les corpus alignés est l'identification de la version originale (langue source) et la version traduite (langue cible). L'information est rarement disponible et supposera vraisemblablement un contact avec le centre de documentation ou le service de traduction, ce qui peut être l'occasion d'obtenir des textes supplémentaires.

4.4 Réflexion sur les caractéristiques des textes disponibles

Internet servant souvent de vitrine aux institutions, il est désormais possible d'y trouver des textes d'une pertinence relativement élevée. Il ne fait toutefois guère de doute que le souci de valoriser l'image extérieure implique inévitablement une uniformité des textes en termes de lecteur modèle, avec les conséquences que l'on peut imaginer quant au degré de technicité ou au choix des termes.

Bien entendu, certains domaines de fine spécialisation sont beaucoup moins riches en corpus disponibles sur la toile : on y trouvera plus facilement des textes de loi ou des cours de médecine que des textes consacrés au calcul de résistance appliqué aux fibres exotiques. Par ailleurs, la répartition des langues sur Internet laisse deviner qu'il sera beaucoup plus compliqué de trouver des textes pertinents pour certaines d'entre elles. Il n'est donc pas sûr que le réseau apporte une réponse adéquate au déficit de corpus pour de nombreux idiomes de faible expansion.

4.5 Internet comme outil de contact avec le spécialiste

L'expérience montre que la visite des sites d'organismes réputés sérieux permet parfois d'identifier un contact fiable au sein de ceux-ci (service des relations extérieures, de documentation, directeur du laboratoire), qui peut déboucher sur l'envoi d'un ensemble de textes beaucoup plus vaste et un contact avec les experts du domaine. Quand l'on songe aux difficultés que pose l'établissement de relations avec les spécialistes adéquats, il s'agit d'une piste que l'on aurait bien tort de négliger. Ceci est d'autant plus vrai qu'un contact avec les auteurs du texte demeure de toute manière une démarche déontologique nécessaire pour des travaux de recherche terminologique dont les résultats contiendront la diffusion publique de nombreux contextes d'attestation originaux.

4.6 Du bon usage de la toile

Marshall Mc Luhan (1968 : 27) a affirmé que « "le message, c'est le médium" parce que c'est le médium qui façonne le mode et détermine l'échelle de l'activité et des relations des hommes. » Internet n'échappe pas à cette géniale intuition, vieille de près de quarante ans. Un parallélisme affirmant que « le corpus c'est la toile » ne risque-t-il pas de nous conduire à ne considérer que des productions écrites qui sont à la mesure de la profonde mutation des activités humaines engendrées par le triomphe de ce nouveau médium ?

Le linguiste prend habituellement le temps d'assembler patiemment un corpus obtenu auprès de sources fiables (éditeurs, organisations internationales, centres de recherche…), de conserver la mémoire de ses caractéristiques et d'en baliser toutes les informations pertinentes[3]. Aujourd'hui, la tentation peut être grande de faire rapidement son marché sur la toile et d'exploiter aussitôt les textes assemblés.

---

3. Par exemple, selon les normes de la *Text Encoding Initiative* (Sperberg-McQueen & Burnard 1999).

L'idée de considérer Internet comme un vaste corpus est à ce point tentante qu'elle constitue dès à présent une piste de recherche sérieusement explorée. Des systèmes dédiés pourraient permettre, dans un proche avenir, de se constituer aisément un ensemble de textes pertinents, regroupés en documents homogènes, correctement documentés, balisés et formatés (Grabar et Berland 2001).

Plus prosaïquement, la puissance des moteurs de recherche permet des vérifications de l'usage que l'on ne pouvait guère espérer voici quelques années. Les expériences que nous avons menées avec nos étudiants de D.E.S.S. au cours de l'année 2001 montrent tout l'intérêt de cet usage dans au moins deux cas de figure :

- Le dépouillement d'un corpus de textes spécialisés à l'aide d'un concordancier ou d'un extracteur permet de découvrir un certain nombre de termes non répertoriés dans les dictionnaires et bases de données terminologiques. Rechercher ces termes sur la toile permet de confirmer leur usage et de le préciser si nécessaire (aire régionale, registre, institution…). Dans certains cas, les seules pages que l'on retrouve sont précisément celles que l'on avait dépouillées.

- Les moteurs de recherche permettent de vérifier l'emploi d'un terme proposé comme synonyme ou comme équivalent par un dictionnaire ou par une base de données terminologiques. Cela va de l'absence de réponse jusqu'à l'observation que l'usage du terme est restreint à telle aire géographique ou à telle institution.

5 LE TRAVAIL DE PRÉPARATION LINGUISTIQUE DU CORPUS

Le travail de constitution du corpus demeure une tâche importante que les facilités d'accès offertes par Internet risquent bien de nous faire oublier. Les textes sont souvent mis à disposition sous la forme de fichiers vectorisés (formats *Acrobat* et *PostScript*), ce qui exige de maîtriser les méthodes de récupération adéquates.

5.1 Le découpage en unités lexicales

Pour l'ordinateur, un mot = toute suite de caractères entre deux espaces blancs.

[blanc]*Suivez-moi,*[blanc]*jeune-homme,*[blanc]*s'écria-t-elle.* [blanc] = 3 mots

Il importe donc de s'interroger sur les problèmes de découpage des unités lexicales (« parsage ») pour les langues considérées. Soit l'on compte sur des options du logiciel pour résoudre ces problèmes, soit l'on applique un prétraitement du corpus.

5.1.1 PONCTUATION

La plupart des logiciels résolvent aujourd'hui correctement les problèmes de découpage liés à la ponctuation.

5.1.2 DIACRITIQUES

Au départ, l'informatique a été conçue pour l'anglais : voici 15 ans, le traitement des accents, cédilles et autres caractères absents de l'alphabet anglais posait encore de nombreux problèmes de tris, aujourd'hui résolus. De nombreux logiciels demandent cependant que l'on spécifie le classement alphabétique.

Ainsi en français : AaÀàÂâ, Bb, CcÇç, Dd, EeÈèÉéÊêËë, etc.

5.1.3 ALGORITHMES DE DÉCOUPAGE DES FORMES

Les problèmes peuvent être très simples, dans le cas d'une langue à faible flexion comme l'anglais, ou très compliqués, dans le cas d'une langue ayant une tendance agglutinante comme l'allemand ou le néerlandais.

Exercice : réfléchissez aux règles de parsage des mots en français sur la base des exemples suivants. Expliquez à un informaticien comment il pourrait s'y prendre pour découper à coup sûr les mots unis par un trait d'union ou une apostrophe.

Dit-il.
Penses-tu ?
Répète-le.
Dis-moi.
Va-t'en.
Laisse-le-moi.
Soyez-en sûr.
Cette personne-là.
Du papier-peint jauni.
Un papier peint en rose.
Suivez-moi, jeune homme !
Son chapeau portait un suivez-moi-jeune-homme.
Le poisson-chat.
Il s'abaisse.
La grand'place.
La grand-place.
Mais qu'en dira-t-on à l'ISTI ?
Je me moque du qu'en-dira-t-on.

5.1.4 PROPRETÉ DE LA MISE EN PAGE

La mise en page peut affecter la logique de contenu du texte autant que son intégrité lexicale ou grammaticale. S'agissant d'un fichier récupéré par reconnaissance optique, par copier-coller ou par conversion de format, il convient de s'assurer de la continuité des lignes, des phrases, des paragraphes. Pour garantir un bon traitement par les outils d'ingénierie linguistique, on veillera particulièrement aux points suivants :

- la disparition des césures coupant les mots en fin de ligne ;

- les fins de ligne ne peuvent pas être confondues avec des marques de fin de paragraphe ;

- la disparition des mises en retrait jouant sur la tabulation ;

- la disparition des marques de fin de paragraphe abusives servant de sauts de ligne.

De manière générale, le fichier doit être aussi propre qu'un fichier de traitement de texte utilisé intelligemment, à l'aide de feuilles de style (modèle de document dans *Word*).

5.2 Sauvegarder correctement un texte depuis Internet

Dans tous les cas, le fichier TXT obtenu doit être inspecté systématiquement, les problèmes de conversion de caractères n'étant pas rares.

Du format HTML au format texte

1. Enregistrer le fichier au format HTML (et non pas le format TXT, car les fins de ligne deviennent des fins de paragraphe)
2. Ouvrir le fichier HTML à partir du traitement de texte
3. Sauvegarder au format du traitement de texte (p.ex. *.txt, *.rtf ou *.doc en *Word*)

Du format PDF *(Acrobat Reader)* au traitement de texte via le format HTML

Par courrier électronique : http://access.adobe.com/francais_1.html
Par un formulaire sur le site d'Adobe : http://access.adobe.com/simple_form.html

Ensuite, comme précédemment, sauvegarder le fichier en HTML, puis le récupérer avec sa mise en page dans le traitement de texte.

Du format PDF *(Acrobat Reader)* au format texte

1. Installer le *plug-in Access 4.05* pour *Acrobat Reader* : http://www.adobe.com/support/downloads/detail.jsp?hexID=5efe
2. Dans *Acrobat Reader*, utiliser le menu « Fichier - Export document to text »

Du format *PostScrip*t au format texte

Le format *PostScript* est très utilisé par les communautés scientifiques utilisant les systèmes UNIX et LINUX.

1. Pour ouvrir un fichier *PostScript* sous *Windows*, installer le logiciel *GhostScript* et son interface graphique *GSview* : http://www.ghostscript.com/
2. Ouvrir le fichier *PostScript* à l'aide de *Gsview*
3. Utiliser le menu « Fichier - Convert… » et choisir le type *PDFwrite*
4. Sauvegarder le fichier avec une extension *.PDF
5. Ensuite ouvrir le fichier PDF avec *Acrobat Reader* et le sauvegarder au format TXT conformément au point précédent.

Exercice : récupérer et nettoyer un fichier de chaque format à l'adresse www.gutenberg.eu.org/pub/gut/publications/cahiers.html#Cahier24

5.3 La préparation structurelle du corpus

5.3.1 ASSURER LA PÉRENNITÉ DU TEXTE

On a cru que l'informatique était la clé d'un archivage à long terme des documents. On s'aperçoit aujourd'hui combien il peut être difficile de lire sur son ordinateur un fichier de traitement de texte vieux d'à peine dix ans, alors que les bibliothèques conservent des manuscrits du Moyen Âge… Pour bien archiver un texte sous un format électronique, on conseille notamment de veiller aux points suivants :

- utiliser un format texte balisé (p.ex. T.E.I.) dont la lecture ne dépende pas d'un logiciel particulier ;
- utiliser un support moderne et faire évoluer le support des archives au fil du temps (bande perforée, bande magnétique, disquette, disquette zip, cédérom… ;
- conserver le fichier dans un lieu protégé (feu, inondation, vol…) ;
- conserver une copie de l'archive dans un lieu séparé ;
- conserver également une impression laser sur papier de qualité : une reconnaissance optique sera toujours possible…

5.3.2 CONSERVER LA MÉMOIRE DU TEXTE

On sait déjà qu'un simple format texte balisé est un excellent format de sauvegarde, peu dépendant des logiciels. Par ailleurs, un corpus textuel doit pouvoir servir à d'autres applications, pour d'autres utilisateurs même si l'on ignore encore aujourd'hui quelles sont les données qui leur sembleront importantes demain. Il importe donc de sauvegarder un maximum d'informations sur le texte d'origine, en sorte qu'elles puissent être exploitées ultérieurement, qu'il s'agisse de données bibliographiques ou d'informations liées à la mise en page du texte (p.ex., une utilisation pertinente des caractères gras et italiques).

5.3.2.1 LA NOTION DE DOCUMENT STRUCTURÉ

Un document écrit, quel que soit son support, peut être défini comme l'alliance indéfectible d'un contenu et d'une structure. Cette définition convient aussi bien à l'écrivain qu'au linguiste. Malheureusement, tous les documents ne constituent pas des documents structurés, au sens où l'entend la gestion électronique des documents[4].

Un format d'échange de traitement de texte comme R.T.F. n'utilise qu'un balisage dit *procédural* indiquant la mise en page (italiques, gras, exposant, justifié…). Le balisage H.T.M.L. utilisé sur Internet ne fait guère autre chose. Dans un cas comme dans l'autre, il ne s'agit pas de documents structurés. En effet, un document structuré utilise un balisage dit *descriptif,* qui identifie la nature de chaque information (titre, exemple, légende, didascalie…). Le balisage descriptif est habituellement de

4. Pour s'initier à cette discipline, on consultera, par exemple, les notes de cours de Serge Stinckwich (1999), de l'Université de Caen.

type logique, c.-à-d. qu'il traduit la structure logique du document (p.ex. un paragraphe est une partie d'un chapitre ou une catégorie grammaticale est une partie de la description syntaxique d'un terme). L'ensemble des balises utilisées est réputé constituer un métalangage. La question de la mise en page du document structuré n'est envisagée que dans un second temps : un texte balisé en X.M.L. peut être aisément mis en page par l'intermédiaire de feuilles de style.

Un balisage doit être bien conçu et il importe que le contenu des balises corresponde effectivement à ce qui a été étiqueté. À ce propos, on ne saurait assez souligner l'importance d'une formation aux documents structurés dans le cadre des études de langues et de lettres. Cette notion de document structuré constitue une évolution importante dans la conception des textes et couvre des aspects aussi bien rédactionnels - sinon littéraires - que linguistiques ou cognitifs : structure des idées, mémoire des textes, exploitation systématique des connaissances…

5.3.2.2 LA *TEXT ENCODING INITIATIVE* (T.E.I.)

On trouvera un texte d'initiation à la T.E.I. à l'adresse http://www.fas.umontreal.ca/EBSI/cursus/vol1no2/beaudry.html.

Voici un début typique de document T.E.I. permettant de conserver la mémoire d'un texte.

```
<teiHeader>
       <fileDesc>
         <titleStmt>
             <title>Thomas Paine: Common sense, a
                 machine-readable transcript</title>
             <respStmt><resp>compiled by</resp>
                  <name>Jon K Adams</name>
             </respStmt>
         </titleStmt>
          <publicationStmt>
             <distributor>Oxford Text Archive</distributor>
          </publicationStmt>
          <sourceDesc>
             <bibl>The complete writings of Thomas Paine,
                collected and edited by Phillip S. Foner
                (New York, Citadel Press, 1945)
             </bibl>
           </sourceDesc>
        </fileDesc>
     <teiHeader>
```

Examinons à présent un échantillon de quelques balises T.E.I. :

<bibl>	contient une citation bibliographique structurée de façon informelle, et dont les sous-champs peuvent ou non être balisés explicitement ;
<div1>...<div7>	contient une subdivision de niveau un à sept des parties liminaires du corps ou des annexes d'un texte ;
<h>	marque un mot ou une expression comme étant graphiquement distincte du texte avoisinant, sans aucune interprétation quant à la raison de cette mise en valeur ;
<title>	contient le titre d'une œuvre, que ce soit un article, livre, journal, ou une série, y compris tout sous-titre ou titre alternatif ;

<p> marque les paragraphes écrits en prose.

5.3.2.3 QUELLES INFORMATIONS PERTINENTES POUR LA RECHERCHE TERMINOLOGIQUE

Il est difficile de dresser un inventaire de toutes les catégories d'informations qui pourraient se révéler pertinentes dans le cadre d'un dépouillement terminologique à l'aide d'un concordancier. Sachant que les contextes d'attestation pertinents seront retenus dans la fiche terminologique, il semble toutefois utile de retenir par ordre d'importance :

- l'auteur ou la collectivité-auteur
- le titre du texte d'origine ;
- la page dans le texte ;
- une éventuelle autre localisation plus ou moins fine (le volume, le chapitre, la ligne dans la page) ;
- l'usage d'un caractère d'imprimerie discriminant (gras, italiques) ;
- la présence de l'attestation dans une partie du texte réservée à des explications (note, glossaire).

6 LES OUTILS DE DÉPOUILLEMENT DES CORPUS

6.1 Historique et tendances

Avant la fin des années 80, il n'était guère courant d'utiliser un ordinateur personnel à des fins de recherche lexicale. C'est alors que sont apparus des logiciels comme *Micro-OCP, Sato* et *Tact*[5], permettant à chacun d'élaborer ses propres concordances sous DOS. Ces produits ont été ensuite concurrencés, avec plus ou moins de succès, par une série de nouveaux logiciels conçus pour les versions successives de *Windows* : *WordCruncher, WordSmith, MononConc…* Des logiciels dédiés spécifiquement à l'extraction de candidats-termes sont apparus plus récemment : le premier fut sans aucun doute *Adepte-Nomino* (dont le prototype *Termino* fonctionnait dès la fin des années 80), suivi beaucoup plus tard par des produits comme *Xerox Multilingual Suite* (2000) ou *ExtraTerm* de Trados (2001).

L'informatique personnelle n'évolue pas toujours, loin s'en faut, vers des produits toujours plus épurés. Avec *Windows,* l'apparence est devenue plus flatteuse, mais les fonctionnalités ne sont pas nécessairement plus évoluées. La tendance est à nos yeux double, comme celle qui peut être souvent observée sur le marché des produits électroniques :

- Activation par défaut de procédures automatisées, censées répondre aux besoins les plus immédiats de l'utilisateur lambda et à son désir de ne pas devoir manipuler une série de réglages élémentaires, propres à la discipline. En linguistique de corpus, il importe de pouvoir jouer sur des paramètres comme le balisage, la longueur des microcontextes, les critères de tri alphabétique, la

5. Notre *Infoport de la terminologie* (www.termisti.refer.org/infof.htm) propose des liens vers la plupart des logiciels cités.

fréquence, les listes d'exclusion, l'étendue des collocations…

- L'utilisateur averti qui souhaite adapter le produit à ses besoins est confronté à une surenchère de paramètres de réglage dont l'influence sur les résultats ne peut que difficilement être évaluée par qui ne dispose pas de tout son temps pour expérimenter longuement toutes les fonctionnalités du logiciel.

Le progrès des logiciels tient compte de l'intérêt que représente l'univers d'Internet pour la linguistique de corpus. Deux tendances manifestes en témoignent aujourd'hui :

- Permettre un dépouillement immédiat de corpus au format H.T.M.L., soit en fournissant l'adresse du site Internet (*MonoConc*), soit en le sauvegardant préalablement sur son disque (*WordSmith*).

- Proposer des services en ligne fonctionnant par l'intermédiaire du courrier électronique (*Intratext*) ou directement sur un site dédié (*Sato Internet*).

6.2 Concordanciers, extracteurs et extracteurs-aligneurs

Les progrès de l'ingénierie de la langue ont donné naissance à des logiciels de plus en plus sophistiqués. Au sein de la famille des logiciels dédiés à la linguistique de corpus, on doit désormais distinguer les concordanciers classiques des extracteurs, qui ont pour mission d'extraire les termes spécialisés identifiés dans un texte. Ces outils ne sont pas nécessairement fondés sur une procédure de lemmatisation (analyse grammaticale), loin s'en faut. Par ailleurs, les extracteurs dont on parle le plus dans les communications scientifiques, comme *Lexter* en France, ne sont pas aisément accessibles au public potentiel ni même à l'ensemble de la communauté des chercheurs.

Notre ambition n'est point de dresser ici un panorama des logiciels existants ou de produire un tableau comparatif de leurs performances, mais plutôt de revenir sur notre propre expérience de chercheur utilisant certains de ces produits.

6.2.1 LE CONCORDANCIER, TOUJOURS AUSSI PERFORMANT

Notre sentiment profond est que le concordancier reste l'instrument de choix pour le linguiste qui souhaite étudier le vocabulaire d'un corpus textuel. Ses options de tri et ses outils statistiques sont bien connus depuis les débuts de la linguistique computationnelle. Il constitue à nos yeux un passage obligé de la formation au troisième cycle.

Des outils comme *OCP, Tact, Sato* ou *WordSmith* permettent un certain nombre de manipulations élémentaires visant à l'observation des fréquences et des figements. La recherche des termes est effectuée de manière complémentaire en fonction de calculs statistiques et de l'observation des microcontextes.

- Calculs statistiques : particulièrement l'information mutuelle (mesure de l'association) et le T-Score (mesure du contraste).

- Observation des microcontextes : dans les microcontextes - ou *kwics (key-words in context)* -, le tri est effectué sur les mots qui suivent ou qui précèdent l'expression recherchée, ce qui permet de très aisément observer les collocations et l'allongement des syntagmes figés.

Le bon usage du concordancier exige une grande attention du linguiste qui s'en sert pour traquer systématiquement les figements et les attestations répétitives. De ce fait, cet outil se révèle particulièrement utile pour observer les faits de langue particuliers au sein des textes spécialisés, sans se focaliser sur la seule catégorie nominale. Les concordanciers peuvent paraître vieillis par rapport aux procédures automatiques proposées par les extracteurs de candidats-termes, mais ils ont l'avantage d'offrir - à qui sait les manipuler - une diversité de procédures de recherche. En d'autres termes, le linguiste se retrouve dans la position du pêcheur qui peut utiliser toutes sortes de filets ou d'hameçons pour tenter d'attraper le poisson rare. S'il n'est point trop pressé, s'il maîtrise les techniques, il sera sûr que bien peu de choses lui échapperont.

Le concordancier qui a actuellement notre préférence est *WordSmith Tools.* Il s'agit en fait d'une suite de logiciels servant à l'analyse lexicale des corpus. *WordSmith* suppose un long compagnonnage avant d'en maîtriser tous les paramètres, ce qui rebute quelque peu les étudiants de troisième cycle.

6.2.2 L'EXTRACTEUR DE CANDIDATS-TERMES

À la différence du concordancier, qui sert à bien des applications en linguistique de corpus, l'extracteur est entièrement dédié à la recherche de candidats-termes. Pour identifier des termes, il peut se fonder sur une analyse linguistique (p.ex. la recherche de syntagmes basés sur le modèle « N *de* N ») ou sur des outils statistiques, voire parfois sur une combinaison des deux approches. Il est à noter que l'analyse linguistique passant normalement par une lemmatisation toujours délicate à programmer, rares sont les extracteurs qui ont été conçus pour traiter différentes langues.

À nos yeux, le grand tort d'un grand nombre d'extracteurs est d'être fondé sur l'idée trop simpliste que les termes spécialisés sont des syntagmes nominaux. Les extracteurs ne sont donc généralement pas capables d'identifier un verbe, un adjectif ou un adverbe spécialisé. Même les termes simples sont souvent exclus de l'algorithme de dépistage. Une telle approche est assurément trop réductrice pour un certain nombre de domaines. Ainsi, un dénombrement comparatif de la répartition des termes simples et des termes complexes dans la base de données terminographiques née du projet européen DHYDRO - toutes catégories grammaticales confondues - donne les chiffres suivants :

	français		**anglais**		**espagnol**	
	effectif	*%*	*effectif*	*%*	*effectif*	*%*
Termes simples[6]	1 710	33,3	1 784	35,2	1 673	29,7
Termes complexes	3 422	66,7	3 282	64,8	3 954	70,3
Total	5 132	100	5 066	100	5 627	100

6. Terme simple : pour les trois langues concernées (non agglutinantes), toute entrée constituée d'une suite de caractères n'incluant ni espace, ni trait d'union, ni apostrophe. Les entrées homographes n'ont pas été comptabilisées. Pour tenir compte des langues agglutinantes, il faudrait suivre les critères de la norme ISO 1087-1 (2000), laquelle définit le terme simple comme découlant d'une seule racine, par opposition au terme complexe qui découle de deux racines différentes au moins. Ce critère de distinction est indubitablement plus adéquat, mais plus difficile à mettre en œuvre.

Le résultat est éloquent, puisqu'il apparaît qu'environ un terme sur trois n'est pas susceptible d'être identifié automatiquement par un extracteur dont les algorithmes sont fondés sur la recherche exclusive de syntagmes.

Le seul élément de comparaison dont nous disposons à cette date concerne la Banque de terminologie du Québec (*Grand dictionnaire terminologique*). Selon les chiffres obtenus de l'Office de la langue française (O.L.F.) en juillet 2001, on y décomptait 96 932 termes simples sur un ensemble de 760 245 termes (entrées et sous-entrées), soit 12,75 %[7]. Si l'on considère ce chiffre comme une moyenne fiable sur un très vaste ensemble de termes couvrant un large ensemble de domaines, nos propres chiffres pour le domaine de l'hydrographie apparaissent sensiblement éloignés. Ce constat nous conduit à envisager que la proportion de termes simples varie de manière considérable d'un domaine à l'autre. Dans ce cas, les extracteurs actuels devraient voir leur productivité varier en fonction du domaine traité. Une ventilation par domaine des chiffres reçus de l'O.L.F. serait assurément très utile pour éventuellement relativiser la moyenne fournie par une mesure de l'écart-type.

6.2.3 L'ALIGNEMENT DE CORPUS

Rédigeant le dossier de soumission du projet européen DHYDRO, en février-mars 1998, nous avions déjà proposé d'intégrer au futur logiciel de rédaction de fiches terminologiques un outil d'interrogation de textes alignés. En septembre 1999, dans le cadre de la *8e Université d'automne en terminologie*, nous augurions avec justesse de la prochaine sortie de « produits beaucoup plus performants, capables d'identifier automatiquement des candidats termes dans plusieurs langues. » Nous précisions même que « Interfacés avec un aligneur puissant et un gestionnaire de mémoire de traduction, ils permettront de directement retravailler une traduction et d'alimenter une base de données terminologique. » (Van Campenhoudt 1999 : 125).

Terminé en juin 2000, *Dhydro* était le premier logiciel, avec la *Xerox Multilingual Suite*, à marier les notions de gestion terminographique et d'exploitation de corpus de textes spécialisés. L'année 2002 verra sans doute se banaliser l'usage de logiciels permettant - à l'instar de *Dhydro* - de vérifier systématiquement comment un terme a été traduit dans un texte, que ce soit à des fins de révision ou pour alimenter une base de données multilingues. Cette dernière perspective n'est toutefois pas sans danger, car elle conduit à placer sur un pied d'égalité le vocabulaire d'un texte original et de sa traduction (*cf.* 4.3). Une telle attitude ne semble défendable que par rapport à des traductions officielles réputées pour leur grande fiabilité et servant de point de référence commun. Tel est par exemple le cas de l'un de nos corpus de recherche : *The Law of the Sea - Le droit de la mer* (Nations unies 1982).

Une autre utilité de l'alignement de corpus est permettre de vérifier la cohérence terminologique d'une traduction. Vérifier si un terme a toujours été traduit de manière acceptable est un enjeu important à une époque où les bureaux de traduction se consacrent avant tout à la gestion de projets, ventilant les parties de la masse textuelle à traduire entre différents traducteurs tenus de suivre des consignes éditoriales très précises.

---

7. Nous tenons à remercier tout particulièrement notre collègue Louis-Jean Rousseau qui a aimablement fait effectuer ce dénombrement par les informaticiens de l'O.L.F.

Dans les prochaines années, il conviendra de s' interroger sur les effets positifs ou négatifs qui auront été induits par l'usage systématique des mémoires de traduction à des fins terminographiques. Dès à présent, on peut notamment envisager les effets suivants :

- un moindre recours aux bases de données terminologiques, la mémoire de traduction pouvant servir à vérifier comment un terme se traduit habituellement ;

- une remise en cause du contenu des bases de données terminologiques en fonction de l'usage réel observé dans des alignements fiables ;

- une multiplication des erreurs de traduction dues à l'usage d'alignements utilisant des traductions de mauvaise qualité ;

- une diminution de la synonymie, un terme étant toujours traduit comme il a été initialement traduit dans la mémoire de traduction.

6.3 Méthodes statistiques : quelques notions de lexicométrie (d'après MULLER 1968)

On nomme *candidat-terme* toute suite de caractères identifiée comme susceptible de constituer un terme spécialisé.

L'approche statistique a pour principal mérite de ne dépendre ni des langues ni d'une modélisation linguistique : fondée sur des dénombrements et des calculs, elle ne suppose aucun développement d'ingénierie linguistique.

6.3.1 ÉTENDUE DU CORPUS

La manière de mesurer la longueur d'un texte varie : nombre de pages, nombre de lignes, nombres de mots, nombre de caractères. Si les bureaux de traduction aiment bien les premières mesures, quitte à se fier - sans grande réflexion - aux approximations des outils statistiques, la lexicométrie se fonde essentiellement sur la notion de « mot », mais en distinguant :

6.3.2 FORMES :

- Le nombre total de formes fléchies comptées dans un texte, y compris celles qui sont répétées maintes fois (*token*) ;

- Le nombre total de formes fléchies différentes comptées dans un texte, les répétitions n'étant pas prises en compte (*type*).

6.3.3 LEMMES :

- Le nombre de formes canoniques différentes (entrées de dictionnaire) observées dans un texte.

Dénombrer les lemmes suppose de posséder un bon logiciel de **lemmatisation** pour la langue concernée, c.-à-d. capable de désambiguïser correctement les formes homographes (*affluent, fils, mousse, voile*, etc.).

6.3.4 MESURER LA RICHESSE LEXICALE

$$\text{Type-Token Ration} = \frac{\text{Nombre de formes fléchies différentes}}{\text{Nombre total de formes fléchies}} \times 100$$

L'interprétation de cette mesure doit être prudente, car elle n'est fondée que sur un corpus donné, traitant d'un sujet donné. Elle n'est que la mesure d'une connaissance active et ne rend pas compte de la connaissance d'un vocabulaire par un auteur. Par ailleurs, la richesse mesurée est nécessairement influencée par la longueur du corpus, car plus un texte est long, plus il y a des chances de voir se répéter les mêmes mots.

Il reste que cette mesure est intéressante pour observer des corpus alignés, voire pour approcher des différences de richesses entre écrits plus ou moins spécialisés.

6.3.4.1 AVEC LEMMATISATION : LA RICHESSE LEXICALE

La mesure la plus rigoureuse consiste assurément à diviser le nombre de lemmes par l'étendue de tout le corpus. De manière générale, on peut être surpris au premier abord par les chiffres obtenus : Dans l'ensemble de ses discours radiodiffusés (1958-1965), le général de Gaulle utilise quelque 4 000 lemmes différents, ce qui est énorme (Cotteret & Moreau 1969).

Pour pouvoir comparer la richesse de plusieurs corpus, il convient de ramener la mesure à une proportion entre le nombre de lemmes et l'étendue du corpus, soit le nombre total de formes fléchies.

$$\text{Richesse lexicale} = \frac{\text{nombre de lemmes}}{\text{nombre de formes}}$$

6.3.4.2 SANS LEMMATISATION : LE TYPE-TOKEN RATIO

Souvent, à défaut de prendre le temps de lemmatiser, on se contente d'établir le rapport entre le nombre de formes fléchies différentes (*type*) et le nombre total de formes fléchies (*token*) :

Cette mesure biaise nécessairement les résultats dans les langues qui, comme le français, présentent une forte flexion et où le *type* ne s'assimile donc que difficilement à un lemme. Par contre, cette mesure est très satisfaisante pour une langue telle que l'anglais, où les mots ne varient presque pas.

6.3.5 Mesurer la connexion lexicale

L'étude de la connexion lexicale vise à comparer des lexiques issus de textes différents. Pour mesurer cette liaison, on calcule un indice qui tend vers zéro quand les deux textes n'ont aucun lemme en commun et vers 1 lorsque la connexion est maximale.

Tout d'abord, on établit la matrice des lemmes communs et différents à partir des données concernant la fréquence de chaque lemme dans chaque texte. Les informations ainsi présentées permettent très facilement de calculer la moyenne **M** des lemmes communs aux deux corpus et de mesurer leur indépendance relative **I**.

M = 29/53 = 0,55 $I_{manif\ 1}$ = 11/40 = 0,28 $I_{manif\ 2}$ = 13/42 = 0,31

	Présent dans Manif 1	**Absent dans Manif 1**	**Total**
Présent dans Manif 2	29	13	42
Absent dans Manif 2	11	∅	11
Total	40	13	53

On propose généralement un indice de connexion **C** qui est le rapport du nombre de lemmes communs et de la moyenne géométrique des deux lexiques[8]. Pour l'exemple ci-dessus, C = 0,71.

$$C = \frac{\text{lexique commun}}{\sqrt{(\text{lexique 1 x lexique 2})}}$$

Lorsqu'on recherche la présence de nouveaux termes dans un corpus, on peut donc envisager de rechercher les termes qui ne sont pas présents dans un autre texte, dont on connaît déjà la terminologie, mais on peut aussi rechercher tous les mots qui ne sont pas dans une liste d'exclusion (p.ex. une liste de tous les mots d'un dictionnaire de référence).

Cette méthode ne peut suffire à elle seule, car elle ne prend pas en compte les syntagmes. Par contre, elle est un excellent complément aux méthodes d'extraction qui assimilent les termes à des syntagmes.

6.3.6 L'INFORMATION MUTUELLE (CHURCH & HANKS 1990, D'APRÈS KRAIF 1997)

L'information mutuelle permet de mesurer le degré d'association de cooccurrents. Intuitivement, elle effectue le rapport entre la probabilité de cooccurrence (contiguë ou non) observée et la probabilité théorique dans le cas où les formes sont indépendantes.

Pour deux formes x et y, on note :

$$I(x,y) = \log 2 \ \frac{P(x,y)}{P(x) \times P(y)}$$

8. Valeur de la racine Nième du produit des N termes d'une série statistique.

Où P(x,y) est la probabilité d'observer x et y ensemble, P(x) et P(y) sont les probabilités respectives d'observer x et y indépendamment.

MI(x,y)	Freq(x,y)	Freq(x)	Freq(y)	X	y
10	161	1419	4764	requested	Anonymity
8.2	14	1419	1529	requested	permission
7.8	5	1419	698	requested	asylum
7.3	5	1419	968	requested	copies
7.1	4	1419	935	requested	detailed
6.8	4	1419	1090	requested	background
6.2	9	1419	3744	requested	documents
6.0	5	1419	2519	requested	protection
5.7	6	1419	3498	requested	additional
5.4	4	1419	2928	requested	meetings
5.0	199	1419	190545	requested	by
5.0	9	1419	8983	requested	information

Tableau extrait de Thierry Fontenelle (Euralex) :
Le dictionnaire Robert & Collins *informatisé : un outil pour le traducteur,*
ISTI, conférence de DESS du vendredi 12 janvier 2001.

Cet indicateur n'est pas valide pour des occurrences rares du type *hapax* (dans ce cas il y a surestimation du lien d'association). Bindi *et al.* (1994), d'après leur corpus, regroupent les « bigrammes » observés en trois classes, suivant la valeur obtenue :

- information mutuelle élevée : noms propres, expressions figées empruntées aux langues étrangères, composés et cooccurrents appartenant à une langue technique ou de spécialité.

- information mutuelle moyenne : « mots de tous les jours ». On observe des expressions du type *telefonata anonima*.

- information mutuelle faible : les paires observées reflètent les structures syntaxiques et grammaticales.

6.4 Approches lexico-syntaxiques : l'exemple d'*Adepte-Nomino*

L'extracteur que nous connaissons le mieux est *Adepte-Nomino*, codéveloppé par l'Université du Québec à Montréal et l'Office de la langue française[9]. Fondé sur une lemmatisation suivie d'une analyse des syntagmes, il ne peut logiquement traiter que les textes de langue française. Comme beaucoup d'extracteurs, il est conçu en fonction d'une théorie bien précise et n'offre donc guère la possibilité de réexploiter le corpus en fonction de stratégies de recherche complémentaires : calculs statistiques, recherche sur les affixes, corpus d'exclusion, etc. Son usage doit donc pouvoir être complété par l'emploi d'un concordancier.

Adepte-Nomino a la particularité d'être un outil dédié spécifiquement à la langue française et peut donc offrir le confort de la lemmatisation. Il prend en compte des types déterminés de termes :

6.4.1 UNITÉS COMPLEXES NOMINALES : NOM + EXPANSION

- NOM + NOM : *carte mère*
- NOM + PRÉPOSITION + NOM : *carte à mémoire*
- NOM + PRÉPOSITION + VERBE : *carte à jouer*
- NOM + ADJECTIF : *abonné téléphonique*
- NOM + PARTICIPE PASSÉ : *carte embossée*

6.4.2 UNITÉS COMPLEXES NOMINALES ADDITIONNELLES

Il s'agit d'une liste des expansions construites à l'aide :

- des prépositions *avec, pour, sans* et *sur* : *état sans littoral*
- des déterminants : *accès à la mer,* mais aussi *exercice de la liberté de transit* et *droit de la mer*
- des expansions infinitives (*machine à laver*)

La préposition *et* ne semble pas prise en compte.

6.4.3 PARAMÈTRES COMBINATOIRES

Lorsqu'on est confronté à un figement, on peut s'interroger sur les limites réelles du terme. Ainsi, dans le cas de *système d'extraction minière et de traitement des minéraux,* les termes envisageables sont :

- *système d'extraction minière*
- *système d'extraction*
- *extraction minière*
- *traitement des minéraux*
- *système de traitement des minéraux*
- *système de traitement*

---

9. Le logiciel de base, *Nomino*, a été mis au point par l'équipe de Pierre Plante à l'Université du Québec à Montréal (UQAM). L'O.L.F. a développé *Adepte* avec le soutien financier du RINT, dont les partenaires ont longuement expérimenté le logiciel. Pour un exposé détaillé des fonctionnalités d'*Adepte-Nomino*, on se référera utilement à Perron (1996).

Adepte-Nomino permet de paramétrer utilement les critères de combinaison. Il permet même d'accepter de neutraliser des enchâssements ; par exemple, dans *importation nette de produit de base*, il proposera d'isoler *importation de produit de base.*

Dès lors pour la collocation *incident de navigation maritime en haute mer*, il envisagera :

- *incident de navigation maritime*
- *incident de navigation*
- *navigation maritime en haute mer*
- *navigation maritime*
- *haute mer*

- *incident de navigation en haute mer*
- *incident maritime en haute mer*
- *incident maritime en mer*
- *incident de navigation en mer*
- *incident en haute mer*
- *navigation maritime en mer*
- *navigation en haute mer*
- *navigation en mer*

6.5 Plaidoyer pour une approche pragmatique fondée sur les corpus

6.5.1 LE TERME FACE AU CORPUS : UNE DIVERSITÉ DE MODÈLES FORMELS

Le lieu n'est pas adéquat pour revenir en détail sur l'étude morphologique du terme (on se référera utilement à Kocourek 1991 : 105-183). Il reste que le travail sur corpus dans un cadre multilingue nous a conduit à reconsidérer l'image du terme qui correspondrait typiquement à un syntagme nominal du type SUBSTANTIF + EXPANSION et serait nécessairement figé sur l'axe syntagmatique.

À cet égard, nous apprécions beaucoup la méthodologie de dépouillement du logiciel *Adepte-Nomino*, qui permet de prendre en compte des syntagmes plus ou moins figés échappant aux modèles de construction les plus fréquents, ce que ses concepteurs nomment des *unités complexes nominales additionnelles (UCNA).* Il s'agit d'une liste des expansions construites à l'aide :

Cependant, un rapide relevé des entrées du *Dictionnaire hydrographique* atteste de l'extraordinaire diversité de formation des termes au départ d'une tête nominale.

N + *de* + N : *aire de vent*

N + *de* + déterminant + N : *âge de la lune*

N + *de* + N + de +déterminant + N : *alignement de contrôle des compas magnétiques*

N + *de* + N + N : *calque de situation surface*

N + *à* + N : *alidade à pinnules*

N + *sur* + N : *levé sur balises*

N + *à* + adj +N : *levé à grande échelle*

N + adj : *aberration annuelle*

N + (adj + [N+ Adj]) : *ligne d'égale variation annuelle*

N + adj + *à* + N + adj : *levé topographique au cercle hydrographique*

N + Nom propre : *tour Bilby*

Sans compter les expressions idiomatiques dans lesquelles un substantif joue un rôle essentiel : *à fleur d'eau, à flot, à terre*

Par ailleurs, une recherche dans des textes spécialisés relevant du domaine maritime confirme notre sentiment que dans certains domaines tel celui de la marine, les termes peuvent être de simples suites de caractères relevant de différentes catégories grammaticales :

- substantifs (*guindant, gréement*) ;
- verbes (*empanner, amurer*) ;
- adjectifs (*tribord, archipélagique, transatmosphérique*) ;

Ces termes simples de différentes catégories grammaticales sont eux-mêmes susceptibles de servir de tête à une expansion, p.ex. :

- ADJECTIF + NOM : *bâbord amure, premier substitut* ;
- VERBE + PRÉPOSITION (+ DÉTERMINANT) + NOM : *virer de bord, se maintenir à l'écart* ;
- PRÉPOSITION + NOM : *au vent, sous le vent* ;
- DÉTERMINANT + ADVERBE + NOM : *au plus près.*

Par ailleurs, nous observons régulièrement que certains termes peuvent :

- avoir des homonymes dans la langue courante : *engagement, border, finir* ;

- être des combinaisons de syntagmes : *droit d'accès à la mer et depuis la mer.*

- appartenir à des langues étrangères comme l'anglais *(outrigger)* ou le latin (*Thon noir : Thunnus atlanticus*).

Plus prosaïquement, on conviendra que la langue spécialisée utilise tous les procédés de la langue courante :

- dérivation (préfixale, suffixale, parasynthétique, régressive) : *aphotique*
- composition et confixation (antérieure ou postérieure) : *bioluminescence, houlographe*
- néologisme (lexical ou sémantique) : *radome, navarea*
- emprunt (lexical ou sémantique), calque, xénisme : *bedrock, boomer*
- troncation : *gyro* (pour *gyrocompas*)
- abréviation (sigle, acronyme) : *GPS (global positioning system), radar (radio détection and ranging), avurnav. (avis urgent aux navigateurs)*
- mot-valise : *racon (radar beacon)*

6.5.2 IDENTIFIER LE TERME PAR LE COMPORTEMENT DE L'USAGER ?

Qu'est-ce qui fait qu'à la lecture d'un texte plus ou moins spécialisé, telle ou telle suite de caractères est identifiée comme un terme ? L'expérience atteste que les réactions dépendent de la personne et le spécialiste n'est pas nécessairement, loin s'en faut, celui qui en détecte le plus.

S'agissant de concevoir des bases de données terminologiques multilingues, nous sommes tenté de considérer comme terme tout mot ou toute suite de mots identifiés comme potentiellement problématique par le traducteur. La question qui se pose alors pour le concepteur de logiciel – et qu'il ne se pose guère – est : « qu'est-ce qui conduit le traducteur à ouvrir son dictionnaire spécialisé ? » :

- l'expression est inconnue
- l'expression semble avoir un sens inhabituel
- l'expression a un emploi syntaxique inhabituel (genre, construction, valence…)
- la forme de l'expression donne à penser à un emploi spécialisé
- l'expression apparaît souvent avec les mêmes collocations
- l'expression précède ou suit systématiquement un terme connu
- …

On peut aisément imaginer que face à de tels phénomènes, le traducteur scrupuleux s'interrogera sur la nécessité de traduire l'expression de manière particulière ou non. Ainsi, il va se demander comment traduire des passages attesté *n* fois :

- *voilier en route libre derrière*
- *droit d'accès à la mer et depuis la mer*
- *régime juridique des eaux archipélagiques*
- *importation nette de produits de base*
- *système d'extraction minière et de traitement des minéraux*

…et il aura à chaque fois bien raison de vérifier dans des bases de données terminologiques. Là commencera son calvaire sur la route de l'incertitude, car il trouvera peut-être seulement :

- *route libre derrière*
- *accès à la mer*
- *régime juridique*
- *archipélagique*
- *produit de base*
- *extraction minière*
- *traitement des minéraux.*

Une certaine expérience de la linguistique de corpus nous amène à suggérer que le terme mérite d'être défini comme le mot ou la suite de mots qui pose un problème de recherche d'équivalence, de compréhension ou d'usage phraséologique. Cette définition n'est, certes, pas la plus simple à modéliser en ingénierie linguistique, mais elle est celle qui se rapproche sans doute le plus de la réalité quotidienne du traducteur. Ce point de vue rejoint l'opinion nouvelle qui voudrait que la terminologie ne soit pas seulement une discipline consacrée à l'élaboration de glossaires[10]. Elle doit aussi fournir des informations sur les collocations dans les textes spécialisés et sur les métamorphoses de termes qui ne sont pas toujours des syntagmes figés.

---

10. Nous pensons, par exemple à Bourigault et Slodzian (1999), à Cabré (1999) ou à Temmerman (2000a).

BIBLIOGRAPHIE

Arrivé (M.), Gadet (Fr.) et Galmiche (M.), 1986 : *La grammaire aujourd'hui. Guide alphabétique de linguistique française,* Paris, Flammarion.

Bindi (R.) *et alii*, 1994, « Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition », dans *Literary and Linguistic Computing,* IX, 1, 29-46.

Bourigault (D.) et Slodzian (M.), 1999 : « Pour une terminologie textuelle », dans *Terminologies nouvelles,* n° 19, p. 29-32.

Bowker (L.), 1998 : « Exploitation de corpus pour la recherche terminologique ponctuelle », dans Humbley (J.), *Terminotique et documentation, Terminologies nouvelles,* n° 18, juin 1998, p. 22-27.

Budin (G.) et Wright (S.E.), comp., 1997-2001 : *Handbook of Terminology Management*, vol. I : *Basic Aspects of Terminology Management*, vol. II : *Applications Oriented Terminology Management*, Amsterdam et Philadelphia : John Benjamins Publishing.

Cabré (M.T.), 1999 : *La terminología : representación y comunicación. Elementos para un teoría de base communicativa y otros articulos,* Barcelona : Institut Universitari de Lingüística Aplicada.

Chiss (J.-L.), Filliolet (J.) et Maingueneau (D.), 1993 : *Linguistique française. Notions fondamentales, phonétique, lexique. Initiation à la problématique structurale 1,* Paris, Hachette supérieur (Langue, linguistique, communication).

Church (K.W.) & Hanks (P.), 1990 : « Word Association Norms, Mutual Information and Lexicography », dans *Computational Linguistics,* XVI, 1, 22-29.

Cotteret (J.M.) et Moreau (R), 1969 : *Recherches sur le vocabulaire du général de Gaulle : analyse statistique des allocutions radiodiffusées (1958-1965)*, Paris, Colin.

Grabar (N.) et Berland (S.), 2001 : « Construire un corpus Web pour l'acquisition terminologique », dans *Actes des 4es Rencontres terminologie et intelligence artificielle (Nancy, 3-4 mai 2001)*, Nancy, INIST-C.N.R.S., p. 44-54.

Habert (B.), Nazarenko (A.) et Salem (A.), 1997 : *Les linguistiques de corpus*, Paris, Armand Colin (U).

ISO 1087-1, 2000 : *Travaux terminologiques - Vocabulaire - Partie 1: Théorie et application*, Genève : Organisation internationale de normalisation.

Kocourek (R.), 1991 : *La langue française de la technique et de la science. Vers une linguistique de la langue savante,* 2e édit. augmentée, refondue, mise à jour avec une nouvelle bibliographie, Wiesbaden : Oscar Brandstetter Verlag & co.

Kraif (O.), 1997 : *Modèles probabilistes pour le traitement automatique de corpus textuel : perspectives et applications*, Nice, LILLA, http://lilla2.unice.fr/labo_fr/COLL&OUV/OUVRAGES/LILLA/Travaux2/kraif.html.

Mc Luhan (M.), 1968 : *Pour comprendre les médias. Les prolongements technologiques de l'homme,* Paris, Seuil (Points Essais, n° 83).

Muller (Ch.), 1968 : *Initiation à la statistique linguistique,* Paris, Larousse (Langue et langage)..

Muller (Ch.), 1973 : *Initiation aux méthodes de la statistique linguistique,* Paris, Hachette (reprint : Paris, Champion, 1992).

Nations unies, 1982 : *United Nations Convention on the Law of the Sea and related Agreements = Convention des Nations Unies sur le droit de la mer,* New-York, Nations unies, www.un.org/Depts/los/index.htm.

Pearson (J.), 1998 : *Terms in Context*, Amsterdam et Philadelphia, John Benjamins Publishing (Studies in Corpus Linguistics).

Perron (J.), 1996 : « Adepte-Nomino : un outil de veille terminologique », dans *Terminologies nouvelles*, n° 15, juin-décembre 1996, p. 32-47.

Sperberg-McQueen (M.) et Burnard (L.), éd., 1999 : *Guidelines for Electronic Text Encoding and Interchange*, Chicago et Oxford : TEI P3 Text Encoding Initiative (www.tei-c.org).

Stinckwich (S.), 1999 : *Aspects informatiques des bases documentaires hétérogènes et réparties,* Caen : Université de Caen, www.iutc3.unicaen.fr/~stincs/ged/index.html.

Temmerman (R.), 2000a : *Towards New Ways of Terminology Description. The Sociocognitive Approach,* Amsterdam et Philadelphia : John Benjamins Publishing (Terminology and Lexicography Research and Practice).

Temmerman (R.), 2000b : « Une théorie réaliste de la terminologie : le sociocognitivisme », dans Diki-Kidiri (M.), dir., *Terminologie et diversité culturelle, Terminologies nouvelles,* n° 21, p. 58-64.

Wüster (E.), 1979 : *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, (2 volumes), Vienne : Springer (ouvrage lu dans une traduction française non officielle.)

Van Campenhoudt (M.), 1999 : « Terminologie descriptive : petite initiation à l'exploitation de corpus », communication présentée dans le cadre de la 8e Université d'automne en terminologie, dans *En bons termes 1999,* Paris, La Maison du dictionnaire, p. 117-126.

TABLE DES MATIÈRES

_____