

TERMINOLOGIE DESCRIPTIVE : PETITE INITIATION À L'EXPLOITATION DE CORPUS

Communication présentée dans le cadre de la 8^e *Université d'automne en terminologie*
Université de Rennes II, septembre 1999

1 INTRODUCTION

La terminographie a trop souvent ressemblé à une pratique lexicographique de seconde zone, consistant à accumuler des listes d'équivalents en compilant des dictionnaires aux qualités incertaines. Dans d'autres cas, elle s'est présentée comme une activité d'aménagement et de normalisation linguistique visant à sélectionner le seul « bon terme ». Le résultat a trop souvent pris l'apparence de véritables « fictionnaires » – électroniques ou imprimés – coupés de l'usage réel et ne pouvant que rougir d'une comparaison avec un véritable travail lexicographique.

Exception faite des programmes d'aménagement linguistique, la terminographie sert d'abord et avant tout à fournir au traducteur une information sémantique et lexicale qui lui permette de traduire au mieux des textes spécialisés. Travaillant souvent dans l'urgence, le traducteur est contraint de faire confiance aux glossaires disponibles, sans garantie que les équivalents proposés correspondent à un usage réel des spécialistes. Force est de constater qu'en dépit des recommandations sans cesse réitérées des formateurs, les terminographies ne comportent que très exceptionnellement des informations sur l'emploi du terme dans son contexte...

Sous l'influence de ses « pères fondateurs », la terminologie a longtemps été conçue comme une discipline consacrée à l'étude de la désignation des concepts et elle s'est ainsi démarquée des préoccupations et méthodes de la linguistique. Si un courant « fondamentaliste » demeure actif dans certains pays et dans certains cénacles, les deux dernières décennies ont vu des linguistes – généralement venus de la lexicologie et de la linguistique de corpus – s'intéresser de près aux langues spécialisées et appréhender leur étude à l'aide des instruments propres aux sciences du langage.

Fidèles à la démarche descriptive de leur discipline, ils défendent le point de vue que l'information terminologique doit rendre fidèlement compte d'un usage véritable, observé dans des textes écrits ou des transcriptions d'échanges oraux. Depuis plusieurs décennies, les lexicologues et lexicographes ont recours à l'exploitation de corpus électroniques, qu'ils dépouillent à l'aide de logiciels dénommés **concordanciers**. Si, voici peu, l'accès aux documents électroniques était encore malaisé, l'avènement des autoroutes de l'information permet aujourd'hui de se procurer de vastes ensembles de textes spécialisés. De nombreux organismes nationaux ou internationaux diffusent des textes dans chacune de leurs langues officielles. Cette accessibilité nouvelle a rendu viable la commercialisation de logiciels jadis réservés aux centres de recherche. Dédiés aux traducteurs, ils permettent d'**aligner** un texte source et un texte cible en sorte qu'un **gestionnaire de mémoire de traduction** puisse ensuite identifier tous les passages déjà traduits par le passé.

La conjugaison du concordancier et de l'**aligneur de corpus** offre aujourd'hui une puissance de recherche terminologique considérable. Le bon usage de ces deux outils permet notamment de :

- rechercher les termes spécialisés ;
- identifier l'environnement phraséologique d'un terme et les marques de figement idiomatique ;
- rechercher des contextes définitoires ou encyclopédiques ;
- identifier les relations sémantiques ;
- rechercher la/les traduction(s) d'un terme ;
- vérifier la cohérence terminologique d'un texte spécialisé ;
- vérifier la cohérence des équivalents utilisés dans un texte traduit.

Il paraît donc désormais indispensable d'initier le traducteur – ou le terminologue venu de la traduction – à la puissance potentielle de ces outils.

2 EN GUISE DE PRÉALABLE : UN CORPUS DE QUALITÉ

Il est toujours bon de rappeler ce qui semble aller de soi : tout texte sur support électronique n'est pas un objet d'investigation fiable et pertinent. Les instruments de la critique des documents, bien connus des historiens, s'appliquent ici avec pertinence et on lira avec intérêt le deuxième chapitre de l'ouvrage *Terms in Context* de Jennifer Pearson (1998) pour envisager tous les paramètres qu'il convient de prendre en compte pour constituer un corpus. Pour notre part, nous nous bornerons à énumérer ici quelques questions fondamentales :

- La taille du corpus est-elle suffisante ?
- Le texte n'est-il pas trop ancien, est-ce son ultime version ?
- Les auteurs des textes sont-ils fiables, reconnus, représentatifs de l'usage et de l'état des connaissances ?
- Les fournisseurs du corpus (p.ex. un site internet) ont-ils modifié le texte ?
- Quelles sont les visées du texte : vulgarisation, enseignement, communication entre spécialistes ?
- Le texte relève-t-il clairement d'un ou de plusieurs (sous)-domaine(s) ?
- S'agit-il d'une version originale ou d'une traduction ?
- S'agissant d'aligner des textes : la traduction a-t-elle un statut officiel ?
- etc.

3 EXERCICE PROPOSÉ

Dans le cadre de l'*Université d'automne en terminologie*, nous avons proposé un exercice pratique, fondé sur l'exploitation de textes diffusés sur le réseau internet. Il s'agissait en l'occurrence de télécharger la version originale anglaise et la traduction française du règlement de l'*International Sailing Federation*¹. Ce texte, qui est traduit officiellement dans un grand nombre de langues, inclut une définition juridique précise de ses termes spécialisés les plus fréquents. L'usage d'un concordancier et d'un aligneur devait permettre de vérifier si :

- d'autres termes importants ne requièrent pas une définition, éventuellement présente ailleurs dans le texte ;
- les auteurs et les traducteurs ont veillé à un usage cohérent des termes définis ;
- le contexte d'usage des termes définis est porteur d'informations syntaxiques ou idiomatiques utiles au traducteur.

Nous ne disposons pas ici de la place nécessaire pour répondre à chacune de ces questions et nous bornerons donc à illustrer quelques emplois des logiciels utilisés.

1. www.sailing.org.

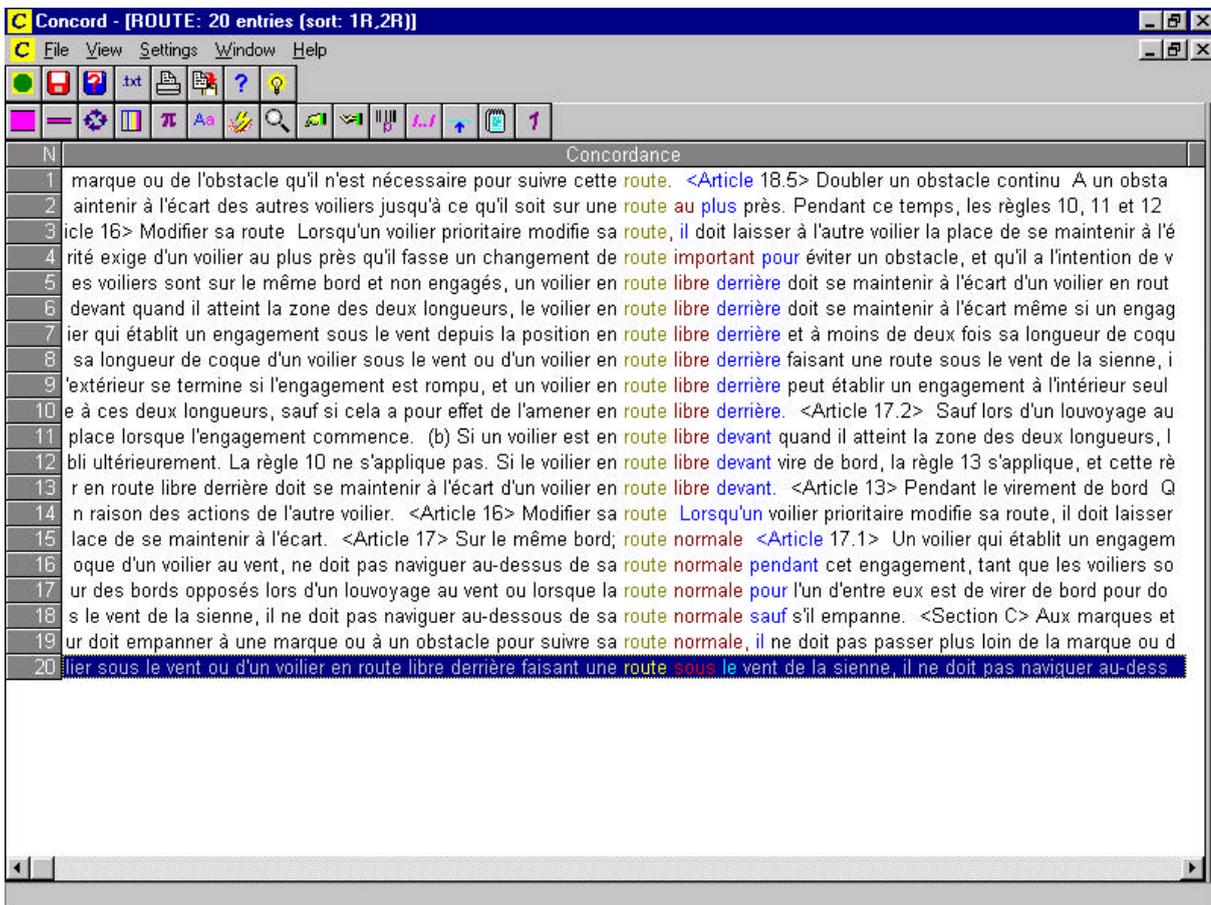
3.1 Dépouillement terminologique à l'aide d'un concordancier

Dans le cadre de l'exercice, nous avons proposé l'usage de *WordSmith Tools*, un progiciel développé par Mike Scott, de l'Université de Liverpool². Ce produit n'est pas le plus élaboré sous l'angle de l'ingénierie linguistique – il ne propose aucun traitement grammatical – mais il a l'avantage de pouvoir s'appliquer indifféremment à un grand nombre de langues, d'offrir de larges possibilités d'exploitation et d'être commercialisé à un coût très raisonnable³.

Nous nous limiterons à illustrer ici quelques potentialités de ce concordancier en matière de recherche des marques de figement, tout en renvoyant le lecteur aux meilleurs manuels de terminologie pour l'étude des critères morphologiques de dépistage des candidats termes⁴.

Le logiciel permet de faire apparaître un mot dans un microcontexte⁵ de n mots, avec des critères de tri élaborés. Par exemple, un tri alphabétique sur les trois mots qui suivent les attestations de *route* dans le corpus considéré fait apparaître les syntagmes *route libre derrière*, *route libre devant* et *route normale* (figure 1).

Figure 1 : tri alphabétique sur les mots qui suivent



2. Un site est consacré au logiciel à l'adresse www.liv.ac.uk/~ms2928/index.htm. On découvrira divers autres concordanciers disponibles sur le marché ainsi que les adresses de sites dédiés à l'exploitation de corpus au départ de notre page www.refer.org/termisti/infof.htm.

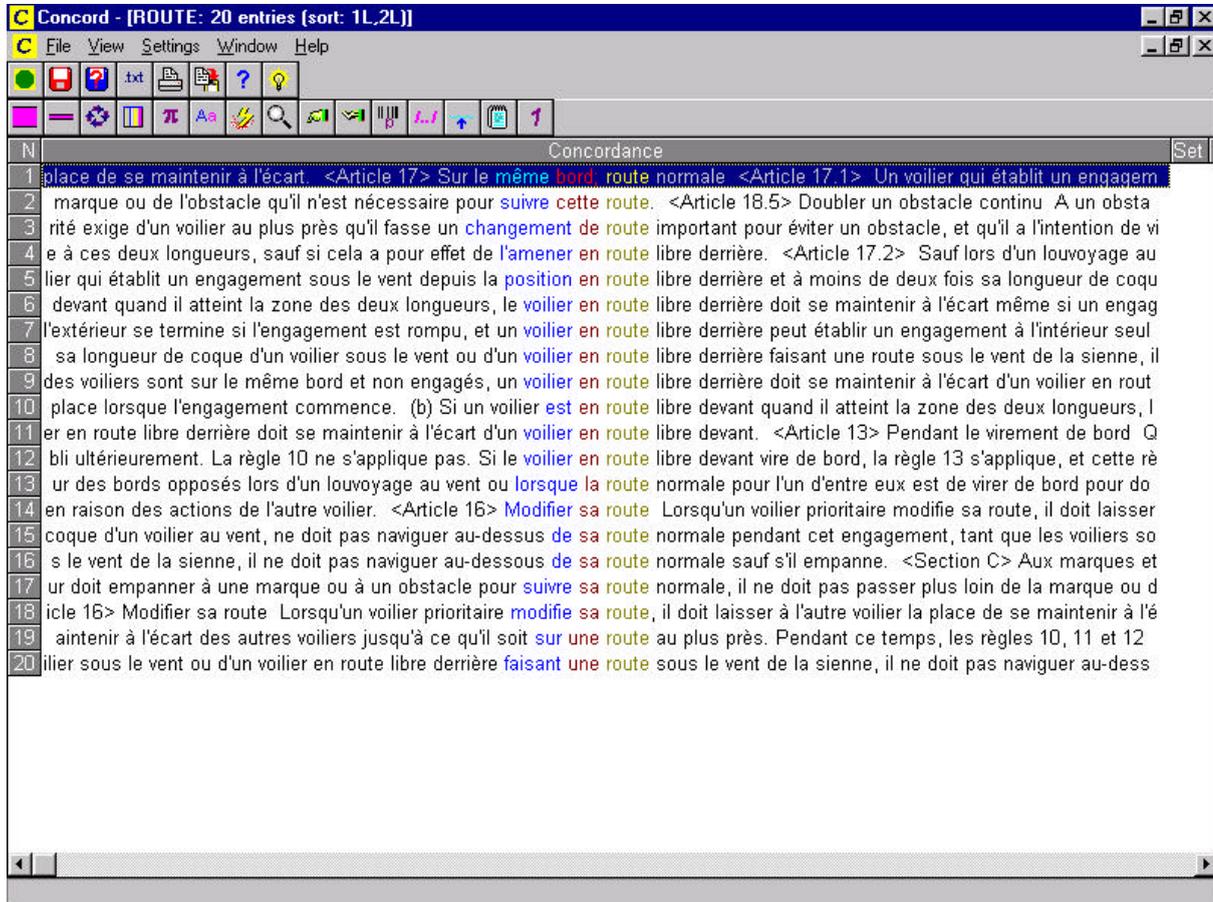
3. Logiciel commercialisé par Oxford University Press : www.liv.ac.uk/~ms2928/index.htm.

4. On trouvera une bibliographie d'introduction à la terminologie à l'adresse www.refer.org/termisti/ref.htm#manuel

5. Souvent dénommé *kwic* (*keywords in context*).

Un tri ultérieur sur les mots qui précèdent *route* (figure 2) permet de découvrir *voilier en route libre derrière* et *voilier en route libre devant*. Le fait que l'on trouve aussi *position en route libre derrière* et « un voilier est en route libre devant » donne à penser que les limites de figement des deux syntagmes sont : *en route libre derrière* et *en route libre devant*.

Figure 2 : tri alphabétique sur les deux mots qui précèdent



L'étude statistique des collocations les plus fréquentes corrobore cette analyse, par ailleurs entièrement confirmée par le chapitre « Définitions » du règlement.

N	cluster	Freq.
1	<i>en route libre</i>	9
2	<i>route libre derrière</i>	7
3	<i>voilier en route</i>	6
4	<i>route libre devant</i>	3
5	<i>sa route normale</i>	3

Les outils statistiques de *WordSmith Tools* permettent, en outre, d'observer l'entourage de chaque mot. On pourra, par exemple, rechercher quelles sont les actions typiques associées à tel concept. Dans la figure 3, les colonnes L1, L2, L... indiquent les mots les plus fréquents à gauche du terme et R1, R2, R... les mots les plus fréquents à droite du terme. Par exemple, pour *marque*, on découvre que les verbes *doubler* et *passer* sont respectivement les deuxième et troisième mots les plus fréquents en 2^e position à gauche du terme, un déterminant (*une* ou *la*) occupant la 1^{re} position. On pourra donc utilement engranger dans la base de données terminologiques les constructions [*doubler* + déterminant + *marque*] ou [*passer* + déterminant + *marque*].

Figure 3 : mots les plus fréquents dans l'environnement immédiat

N	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
1	de	départ	et	à	une		ou		ou	à
2	délimite	doit	pour	doubler	la		de	un	obstac	la
3	en	la	a	passer				l'obstac		de
4	prendre	est	pas	la		comme	à	il		
5	passer	il	termine	d'arrivée			comme	comme	de	n'est
6	le	marque	peut	quand				délimite		un
7	s'appliq	voilier	manqu	ou			manqu		n'est	plus
8	à	plus	avoir	départ		d'arrivée			un	termine
9	la	ou	en	de		manqu				
10		prendre	le				l'obstac	ou		peut
11		s'appliq	s'appliq					le		
12		pas	départ				un			
13		le	de			est		de		
14		de						obstac		
15							il	n'est		avoir
16										
17									termine	voilier
18							délimite			est
19										
20						ou	quand	avoir	marque	
21										quand
22								voilier	plus	
23						a	en		peut	doit
24										
25								doit		prendre
26										pour
27										
28										

Ces quelques exemples ne couvrent qu'une faible partie des potentialités d'exploitation d'un logiciel comme *WordSmith Tools*, qui permet aussi la recherche de traits définitoires ou de relations sémantiques. Nous sommes toutefois persuadé qu'ils suffiront à convaincre toute personne peu formée à la linguistique de corpus de l'intérêt d'un tel outil.

3.2 Utilisation d'un outil d'alignement

L'alignement de textes est une pratique qui a été popularisée auprès des traducteurs par les logiciels de gestion de mémoires de traduction, tel le *Translator's Workbench* de Trados. Ces outils permettent de retrouver toute phrase déjà traduite, ce qui peut s'avérer très intéressant dans des domaines marqués par la répétition. Une mémoire de traduction peut être enrichie par des corpus de textes traduits par le passé : pour ce faire, on essaye d'aligner au mieux les phrases des textes en langue cible et en langue source à l'aide d'un outil nommé **aligneur**.

Aujourd'hui, on s'aperçoit que l'alignement offre un intérêt complémentaire à l'exploitation de concordanciers. Deux grandes utilisations retiendront ici notre attention : la recherche d'équivalents et la vérification de la cohérence terminologique d'une traduction.

3.2.1 RECHERCHE D'ÉQUIVALENTS

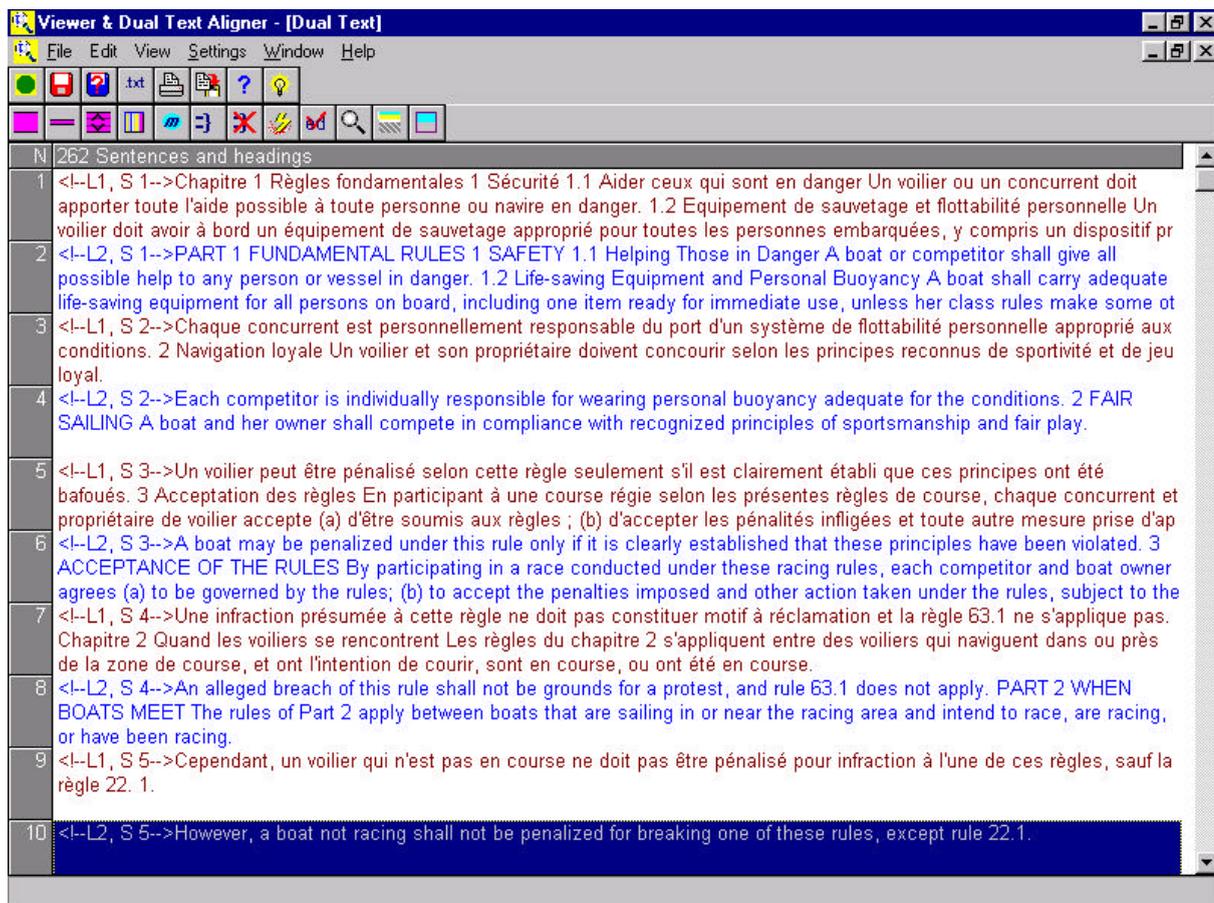
Dans les cas où l'on dispose d'une traduction réputée fiable d'un texte, on peut envisager d'y rechercher comment tel terme a été traduit et faire sienne cette traduction. Une telle démarche requiert toutefois une grande prudence, car d'un point de vue méthodologique, elle ne peut en aucun cas avoir la même valeur que le dépouillement de textes originaux dans la langue considérée. Il s'agit tout au plus d'un technique complémentaire, visant à étudier l'équivalence dans des traductions dignes de foi. À défaut de disposer de dictionnaires fiables, le traducteur sera, lui aussi, tenté de vérifier s'il ne peut pas s'appuyer sur de telles traductions ; tout en s'assurant – bien sûr – que le terme a bien le même signifié dans les contextes identifiés.

3.2.2 VÉRIFIER LA COHÉRENCE TERMINOLOGIQUE D'UNE TRADUCTION

Lorsque l'on révise une traduction spécialisée, on doit notamment s'interroger sur la cohérence de la terminologie utilisée. Si l'on a pu identifier les termes du texte source, il devient très aisé d'utiliser l'alignement pour vérifier si chaque terme a toujours été traduit de manière pertinente dans le texte cible.

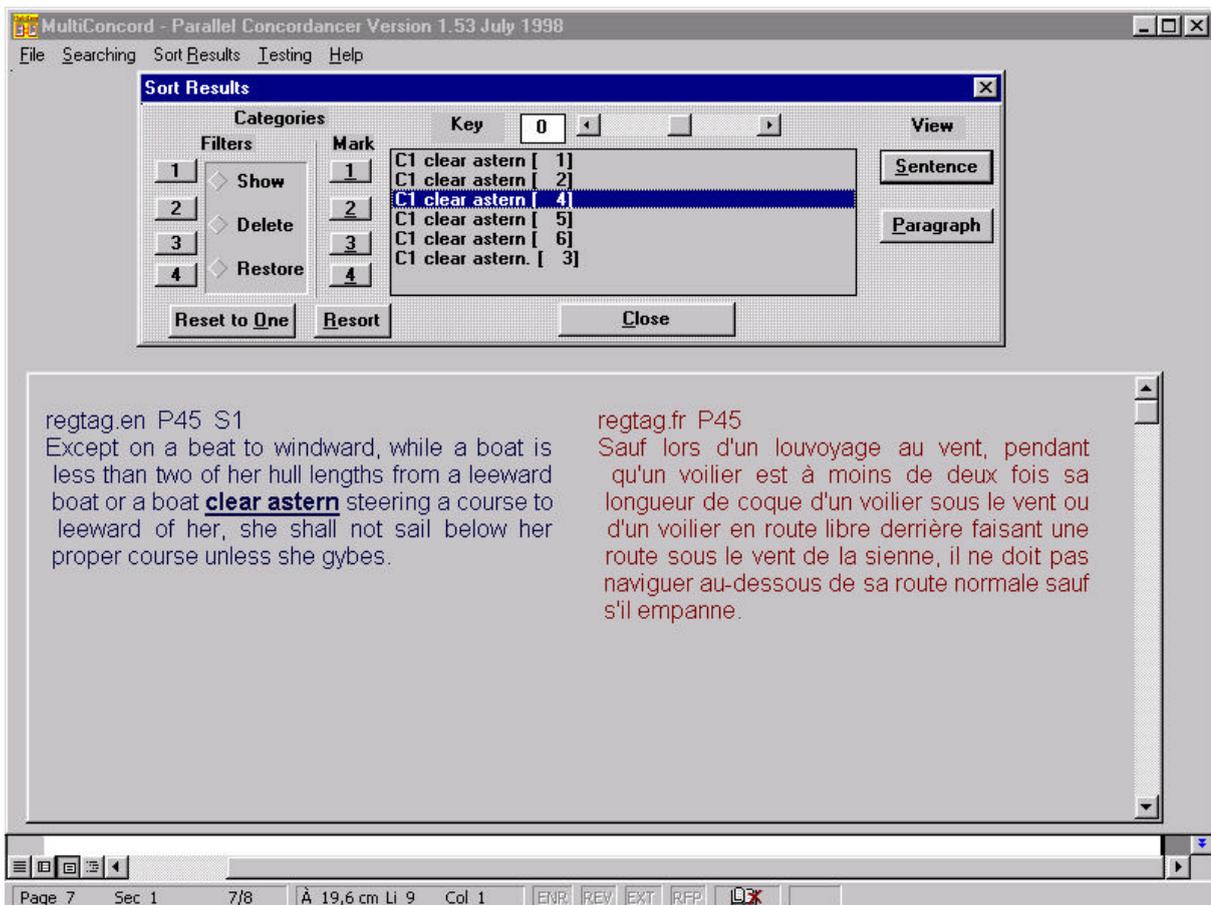
Avant tout, il convient de procéder à l'alignement. À défaut de posséder un coûteux aligneur de la dernière génération, on pourra utiliser *WordSmith*, qui inclut un aligneur certes élémentaire, mais d'une efficacité suffisante pour un usage de nature lexicale (figure 4).

Figure 4 : alignement du texte source et du texte cible



Pour l'exploitation de l'alignement, nous utiliserons un autre logiciel, nommé *MultiConcord*. Cet outil développé par l'Université de Birmingham⁶ est sans doute rudimentaire et nullement dédié à la seule recherche terminologique, mais il est d'un coût modique et suffit amplement à nos visées. Après avoir indiqué quels étaient les textes alignés, on peut consulter tous les contextes où est attesté tel ou tel terme. On découvre ainsi que dans le corpus considéré, l'expression *clear astern* est toujours traduite en français par *en route libre derrière* (figure 5). La procédure de recherche est d'une simplicité déconcertante, au point que l'on peut se demander pourquoi elle tarde à se banaliser.

Figure 5 : utilisation de *MultiConcord*



CE QUE L'AVENIR NOUS RÉSERVE

De nombreux logiciels d'exploitation des corpus textuels sont disponibles sur internet à un prix très abordable. Peu automatisés, ils ont le mérite de former l'étudiant aux arcanes de la recherche de candidats termes et d'équivalents. Il ne fait toutefois guère de doute que seront commercialisés dans les prochains mois des produits beaucoup plus performants, capables d'identifier automatiquement des candidats termes dans plusieurs langues. 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. Plus que jamais, les bureaux de traduction auront besoin de disposer d'une personne-ressource capable de tutoyer ce genre de produit.

6. <http://sun1.bham.ac.uk:80/johnstf/lingua.htm>.

À l'heure où le couple linguiste-informaticien fait nettement progresser les industries de la langue, le dictionnaire spécialisé multilingue ne peut plus conserver aujourd'hui le statut de seule source de référence accessible. Du fait de ces progrès, la terminologie apparaît de plus en plus comme une étude du terme dans son contexte, largement fondée sur les méthodes de la linguistique de corpus. Ce changement de perspectives fait aujourd'hui l'objet d'affirmations fortes de la part d'un certain nombre de linguistes (Bourigault et Slodzian 1999). Il semble logique que l'enseignement de la traduction spécialisée évolue en conséquence, comme le suggèrent déjà, à juste titre, les études méthodologiques menées par Lynne Bowker (1998).

*Marc Van Campenhoudt,
Centre de recherche Termisti,
Institut supérieur de traducteurs et interprètes,
Bruxelles.*

BIBLIOGRAPHIE

Bourigault (Didier) et Slodzian (Monique), 1999 : « Pour une terminologie textuelle », dans Enguehard (Chantal) et Condamines (Anne), éd. : *Terminologie et intelligence artificielle, actes des 3^{es} rencontres « Terminologie et intelligence artificielle » (Nantes, 10 et 11 mai 1999)*, *Terminologies nouvelles*, n°19, décembre 1998 – juin 1999, p. 29-32.

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

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