CHAPITRE VI

D'UNE LANGUE A L'AUTRE : LES CORPUS ALIGNES

1. DEFINITION ET EXEMPLES

On appelle textes alignés (ou bi-textes) des couples de textes dont l'un est une traduction de l'autre et pour lesquels il existe un système de mise en relation entre segments du texte de « grain équivalent » : sections, paragraphes, phrases. On parle également de corpus bilingues.

Des occurrences de *guerre froide* ou *cold war* sont fournies par le Hansard aligné, c'est-à-dire les débats du Parlement canadien où la version en anglais est mise en correspondance avec la version française[1]. Voici quatre exemples de contextes alignés, où, à chaque fois, le texte source est anglais :

```
That is what is called leadership ,     | Voilà en quoi consiste le leadership .
not sticking one 's head in the sand    | Il faut éviter de faire l' autruche ,
, not looking through the rear - view   | de regarder en arrière et d' éprouver
mirror , not having some nostalgia      | une certaine nostalgie de l' ancienne
for the old cold war but saying it is   | guerre froide . Il faut plutôt se dire
time to make some change .              | que le moment est venu d'apporter des
                                        | changements.

This happened in 1990 , and now she     | C' était en 1990 . Aujourd'hui , elle
says : `` I do not understand why all   | dit qu' elle ne comprend pas pourquoi
of a sudden you are now saying we       | tout à coup nous trouvons à redire à
have a problem with the program '' ,    | ce programme . Mis à part le fait que
```

[1] Les contextes ont été fournis par L. Langlois (Dictionnaire canadien bilingue - Université d'Ottawa) utilisant sous licence TransSearch qui permet des concordances sur des textes alignés. TransSearch a été développé au CITI (Centre d'Innovations en Technologie de l'Information - Laval, Canada), devenu le RALI (Laboratoire de Recherche Appliquée en Linguistique Informatique). Cf. (Simard *et al.*, 1992).

quite apart from the fact that the geostrategic situation has changed tremendously in the period we are talking about . The cold war was pretty cold in 1990 .	la situation géostratégique a terriblement changé depuis , la guerre froide était plus que froide en 1990 .
I also want to acknowledge the staff reductions indicated by CSIS in the counterintelligence area . They are probably a function of the reduction in cold war intelligence battles that went on for many years .	Pour terminer , je voudrais parler de la réduction des effectifs mentionnée par le SCRS dans le secteur du contre - espionnage , réduction qui est peut - être attribuable à l' apaisement de la guerre froide .
It is not so easy to keep them in the cold dawn of post - war budgeting .	Il est moins facile de les tenir après la guerre , à l' époque froide des contrôles budgétaires .

On perçoit sur ces exemples, dont le second remotive les constituants de l'expression toute faite, les difficultés de la mise en correspondance (une phrase anglaise d'un côté, deux phrases françaises de l'autre dans l'exemple 2, l'inverse dans l'exemple 3). Le troisième exemple manifeste par exemple des décalages entre les deux versions (*intelligence battles that went for many years* est sans équivalent dans la version française). Le quatrième est une métaphore filée à partir de l'expression toute faite.

Ce bi-texte manifeste des types de contextes nouveaux par rapport à ceux examinés par Barkema (chapitre II) :

- *cold war* {nom}, où *cold war* est le modifieur du nom :

  cold war attack helicopters / hélicoptères d' assaut bons pour la Guerre froide

  cold war style helicopters / hélicoptères rappelant l'époque de la guerre froide

  cold war helicopter program / programme d' achat d' hélicoptères digne de la guerre froide

  the EH-101 cold war helicopters / hélicoptères EH-101 conçus pour la guerre froide

  cold war helicopters / hélicoptères de la guerre froide

  Ces contextes récurrents sont appuyés par la paraphrase suivante : helicopters to fight the cold war / hélicoptères destinés à la guerre froide ;

- des contextes qui précisent les parties prenantes du conflit larvé :

  the Moscow - Washington cold war / La guerre froide entre Moscou et Washington

  helicopters for the cold war with the Soviet Union / hélicoptères pour faire la guerre froide avec l' Union soviétique

  The cold war between the two blocs / cette guerre froide - là entre les deux Blocs

- *post cold war* {nom}, où le nom en question renvoie à une dimension temporelle, modifié par le syntagme *post cold war* :

  the post cold war environment / le climat d' après - guerre froide

  in a post - industrial , post - cold war world environment / en cette période postindustrielle et d' après - guerre froide

  In a post - industrial , post cold war environment / À l' ère postindustrielle , la guerre froide étant chose du passé

the post cold war era / dans l' ère de l' après - guerre froide

post cold war world / depuis la fin de la guerre froide

the post - cold - war situation / l' après - guerre froide

La version utilisée du Hansard aligné, qui correspond à trois ans de débats, représente 21,6 millions de mots anglais et 24,1 millions de mots français. Elle comprend 5 993 occurrences de *guerre*, 384 de *froide*, 5 977 de *war* et 673 de *cold*. Pour un volume globalement équivalent au corpus de Birmingham utilisé par Barkema, on rencontre près de trois fois plus d'occurrences de *cold war ou guerre froide* (314 occurrences). On ne trouve aucune occurrence de *guerres froides* ni de *cold wars*. On ne trouve qu'un seul exemple de discontinuité entre les deux composants de l'expression : c'est l'exemple 4 ci-dessus. Ces constats confirment l'analyse de Barkema sur la rigidité de l'expression. Dans 8 cas d'ailleurs, la traduction de *cold war* se fait par *Guerre froide*, la majuscule soulignant le fonctionnement comme un tout indécomposable.

2. UTILISATION DES TEXTES ALIGNES

Le recours aux textes alignés constitue par certains côtés une riposte aux limites rencontrées dans l'automatisation de la traduction automatique. Le point de départ n'est pas une formalisation de deux langues et de leur mise en correspondance, mais la réutilisation des traductions existantes produites par des traducteurs humains.

Les textes alignés fournissent un appui critique à la traduction. Cet appui peut consister à vérifier qu'il n'y a pas d'omissions dans la traduction. On en a précisément relevé une dans l'exemple 3 de la section 1. Un autre problème est celui des faux-amis partiels (Isabelle et Warwick-Amstrong, 1993, p. 302) : *Max fut arrêté par le FBI -> Max was arrested by the FBI* versus *Max arrêta le moteur -/-> Max arrested the engine, -> Max stopped the engine*. Disposer de contextes alignés permet de vérifier l'adéquation de la traduction qu'on se propose d'utiliser. Il importe alors de pouvoir filtrer les contextes sur des expressions des deux langues à la fois.

Les textes alignés servent de ressource pour les termes dont la traduction « homologuée » dans la langue-cible ne correspond pas forcément à une traduction mot à mot. Le Hansard aligné montre que les traducteurs utilisent généralement *droit compensateur* pour *countervail*, et parfois *droit compensatoire* (Isabelle, 1992). En langue générale, les textes alignés donnent accès à « la bonne expression » que le traducteur ne trouvera pas forcément dans un dictionnaire ou à des solutions auxquelles il n'avait pas pensé mais qui le satisfont et qui lui permettent de varier son expression. Voici quelques équivalences trouvées dans le Hansard pour l'expression *cartes sur table* (*ibid*.) :

Il a mis cartes sur table | He has put his facts on the table

Mettez-donc les cartes sur table | Put your cards on the table

Si c'est le cas, mettons cartes sur table [...] | If that is the case, let us get it on the table [...]

Peut-il jouer cartes sur table ? | Will he come clean with the Canadian people ?

Il devrait jouer cartes sur table avec les Canadiens | It should present Canadians with the straight goods.

Les techniques actuelles d'alignement poussent à vouloir exploiter le « trésor » que constituent les traductions déjà existantes. P. Isabelle (*ibid*.) indique : « Au Canada seulement, bon an mal an, le volume de traductions atteint au moins un demi-milliard de mots. [...] La masse des traductions produites chaque année contient infiniment plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables. » L'objectif est alors de chercher s'il n'existe pas déjà une solution au problème de traduction rencontré, dans les traductions existantes, plutôt que d'en inventer une de toutes pièces. Les bi-concordanciers comme TransSearch permettent de telles recherches.

Les corpus alignés permettent de repérer des néologismes et la traduction qui en est donné. Ils viennent aussi remédier aux inévitables lacunes des dictionnaires. Gale et Church (1991) montrent par exemple que dans les corpus qu'ils avaient alignés, *en jeu* servait souvent de traduction à *at risk*, alors qu'un dictionnaire comme le *Robert et Collins* ne mentionne pas cette équivalence.

3. METHODES D'ALIGNEMENT

L'objectif est, selon P. Isabelle et S. Warwick-Amstrong (1993, p. 288) « la reconstitution automatique des correspondances traductionnelles qui unissent les segments d'un texte source et ceux de sa traduction. » Cet objectif est moins ambitieux que ceux qu'implique une traduction automatique : « Par opposition à la compétence active mise en jeu par les systèmes de traduction automatique, la recherche de correspondances dans les traductions préexistantes suppose seulement une compétence passive qui, en principe, devrait être moins difficile à atteindre » (*ibid.*, p. 289). La nature même de l'objectif conduit à des méthodes différentes. On part de « l'équivalence traductionnelle » qui est au contraire le résultat final escompté de la traduction automatique.

L'alignement peut s'effectuer aux différents niveaux de structuration de l'énoncé : des sections du texte aux mots en passant par les paragraphes et les phrases. C'est ce que P. Isabelle et S. Warwick-Amstrong (*ibid*.) nomment la « résolution » de l'alignement. Les correspondances deviennent de plus en plus difficiles à établir lorsqu'on diminue la taille des entités rapprochées. Les grandes sections d'un document sont général en relation bijective entre les deux versions. C'est encore souvent le cas pour les paragraphes. Les phrases font déjà exception. Une phrase dans une langue peut se traduire par deux phrases, voire plus dans l'autre, nous en

avons vu des exemples. L'ordre des propositions ou des phrases peut varier. En deça de la proposition, la variation de l'ordre des mots ainsi que le remplacement d'un mot dans une langue par une périphrase ou une expression polylexicale dans l'autre constituent des obstacles plus évidents encore à l'alignement.

P. Isabelle et S. Warwick-Amstrong (*ibid.*, p. 292) fournissent une définition tout à fait générale de l'alignement :

(T1, T2, Fs, C(Fs(T1), Fs(T2)))

T1 est le texte source, T2 sa traduction. Fs est une fonction de segmentation (cf. chapitres VII et VIII) qui fragmente le texte (il peut s'agit de mots, de phrases, de paragraphes, de sections). C est une fonction de correspondance qui relie l'ensemble des segments produits par Fs sur le texte source, Fs(T1), à l'ensemble des segments fournis par Fs sur le texte cible, Fs(T2).

Deux méthodes sont employées pour l'alignement. La première s'appuie sur l'existence d'une très forte corrélation entre la longueur d'un segment source et celle de sa traduction. La seconde utilise les paires particulières des mots pour mettre en corrélation. D'autres propositions sont des variations sur ces propositions de base ou encore la combinaison des deux approches.

La première méthode utilise donc la « corrélation très forte entre la longueur des segments qui sont mis en correspondance traductionnelle » (*ibid.*, p. 295). Les segments peuvent être mesurés en nombre de mots (Brown et al., 1991) ou en nombre de caractères (Gale et Church, 1991)[2]. Chacun des deux textes est d'abord décomposé en phrases[3]. On se donne un ensemble d'appariements licites (un / zéro, zéro / un, un / un, un / deux, deux / un, etc.). Dans la plupart des cas, on n'autorise pas les liens croisés. On examine alors tous les appariements possibles compatibles avec les appariements retenus comme licites. On calcule un score reflétant la qualité des corrélations des longueurs des segments contenus pour chaque appariement. On retient l'appariement dont le score est le meilleur. Les résultats sont entre 95 et 100 % d'appariements justes. Cette famille de méthodes présente l'avantage de ne pas nécessiter de recours à un dictionnaire. Inversement, l'examen « à gros grain » des corrélations entre les deux textes empêche une resynchronisation quand l'appariement se décale à un endroit donné.

La deuxième méthode prend appui sur les mots apparentés entre deux langues proches (*gouvernement* / *government* par exemple). Il ne s'agit pas d'utiliser un dictionnaire mais de repérer des distances entre chaînes de caractères (par exemple en termes de coût de passage d'une chaîne à l'autre en nombre d'effacements, ajouts et substitutions).

[2] Cf. aussi (Blank, 1995 ; Langé et Gaussier, 1995).
[3] Tâche qui est moins évidente qu'elle n'en a l'air. Que l'on pense aux titres, aux énumérations, aux légendes de tableaux et de figures, aux incises.

4. PROBLEMES ET ENJEUX

P. Isabelle et S. Warwick-Amstrong insistent (*ibid.*, p. 290) sur la « compositionnalité de la traduction » : « la traduction d'une unité textuelle est généralement fonction de la traduction des parties de cette unité, et ce, jusqu'au niveau d'un ensemble fini d'équivalences élémentaires. » C'est effectivement ce principe qui rend possible la démarche d'alignement. Mais en même temps, comme nous l'avons vu, la « résolution » de l'alignement peut être plus ou moins grande : des correspondances des grandes parties du texte et des paragraphes s'accommodent de décalages à un niveau plus fin (c'est le cas du troisième exemple de la section 1, où une partie de la phrase source n'a pas de correspondant traductionnel). Comme l'indiquent P. Isabelle et S. Warwick-Amstrong (*ibid.*, p. 302), un système d'alignement fin permettrait de repérer les erreurs de traduction liés aux faux amis, c'est-à-dire les cas où un mot est traduit par un mot trompeusement proche (comme *eventually* pour *éventuellement*).

Les textes alignés permettent également d'examiner les équivalences entre séquences non compositionnelles : les décalages localisés qu'elles représentent sont contrebalancés par l'alignement des structures plus vastes dans lesquelles elles figurent. Les textes alignés permettent en ce sens une répartition relativement harmonieuse des tâches entre « machine » et traducteur. L'alignement produit un dégrossissage des mises en correspondance. En fonction de la requête qu'il effectue, le traducteur puise dans les réponses et s'appuie sur les blocs alignés pour examiner les parallèles ou les divergences dans le détail. L'alignement produit automatiquement est évidemment limité, mais il est suffisant pour beaucoup de tâches de traductique.

L'alignement, du moins « à gros grains »[4], peut sembler une tâche plus aisée que l'étiquetage ou le parsage. En tout cas, il y a un grand décalage entre la relative simplicité des méthodes employées pour obtenir des textes alignés et la richesse extrême des utilisations de ces corpus bilingues. Ce décalage même est source d'espoir.

[4] Par opposition à un alignement syntagme à syntagme voire mot à mot.