TALN 2004, Fès, 19–21 avril 2004

Traduction, traduction de mots, traduction de phrases

Eric Wehrli
LATL - Département de linguistique
Université de Genève
CH-1211 Genève 4
Eric.Wehrli@lettres.unige.ch

Résumé - Abstract

Une des conséquences du développement d'Internet et de la globalisation des échanges est le nombre considérable d'individus amenés à consulter des documents en ligne dans une langue autre que la leur. Après avoir montré que ni la traduction automatique, ni les aides terminologiques en ligne ne constituent une réponse pleinement adéquate à ce nouveau besoin, cet article présente un système d'aide à la lecture en langue étrangère basé sur un analyseur syntaxique puissant. Pour un mot sélectionné par l'usager, ce système analyse la phrase entière, de manière (i) à choisir la lecture du mot sélectionné la mieux adaptée au contexte morpho-syntaxique et (ii) à identifier une éventuelle expression idiomatique ou une collocation dont le mot serait un élément. Une démonstration de ce système, baptisé TWiC (*Translation of words in context "Traduction de mots en contexte"*), pourra être présentée.

As a consequence of globalisation and the development of the Internet, an increasing number of people are struggling with on-line documents in languages other than their own. In this paper, we will first argue that neither machine translation nor existing on-line terminology tools constitute an adequate answer to this problem, and then present a new system conceived as a foreign-language reading assistant, based on a powerful syntactic parser. Given a word selected by the user, this system carries out a syntactic analysis of the whole sentence in order (i) to select the most appropriate reading of the selected word, given the morpho-syntactic context, and (ii) to identify a possible idiom or collocation the selected word might be part of. A demo of this system, dubbed TWiC (Translation of words in context) will be available.

Mots-clefs – Keywords

Traduction automatique, aide à la traduction, analyse syntaxique, collocations
Machine translation, Translation aids, syntactic parsing, collocations

1 Introduction

L'expression de *village global*, d'abord introduite par Marshall McLuhan dans son ouvrage *The Gutenberg Galaxy* pour faire référence aux conséquences globales de l'émergence des médias électroniques, est couramment utilisée aujourd'hui comme métaphore pour décrire le réseau Internet et ses millions de nœuds répartis à travers le monde, auxquels on peut accéder pratiquement immédiatement d'un simple clic de souris. Ce village, cependant, a une propriété peu commune et pour tout dire quelque peu inattendue : son caractère multi-culturel et multilingue, qui évoque davantage l'image de la Tour de Babel que celle du village traditionnel.

En effet, selon GlobalReach (www.glreach.com), la part des utilisateurs non-anglophones d'Internet a passé de plus de 45% en 2001 à moins de 36% deux ans plus tard. Comme, par ailleurs, près de 70% des pages disponibles sur Internet sont en langue anglaise, on peut raisonnablement conclure qu'un très grand nombre d'utilisateurs non-anglophones d'Internet consultent régulièrement des pages en langue anglaise. En fait, il n'est sans doute pas exagéré de penser que jamais dans la monde autant de gens n'ont eu recours à des informations dans une langue autre que la leur.

C'est bien à ces utilisateurs – que nous sommes pratiquement tous à un degré ou à un autre – que nous nous intéressons dans cet article, qui naviguent régulièrement sur des sites anglophones (ou germanophones, ou autres). Possédant une maîtrise de l'anglais (respectivement de l'allemand, ou autre) suffisante pour comprendre ce qu'ils lisent, ces utilisateurs font pourtant face, régulièrement ou occasionnellement, à des problèmes terminologiques : un mot (ou une expression) leur échappe, susceptible de nuire à une bonne compréhension du texte.

Dans la première partie de cet article, nous montrerons comment ni la traduction automatique, ni les outils actuels de terminologie en ligne ne constituent une solution pleinement adéquate au problème décrit ci-dessus. Au risque d'une formule tout à la fois elliptique et facile, nous conclurons que la première en fait trop, la seconde trop peu.

Dans la deuxième partie de cet article, nous présenterons TWiC, un outil spécifiquement développé pour répondre au problème ci-dessus. Alliant un analyseur syntaxique "profond" à une base terminologique bilingue riche, TWiC utilise l'analyseur pour identifier le mieux possible l'unité lexicale sélectionnée par l'utilisateur, de manière à limiter autant que possible le bruit issu associé à une recherche terminologique. Nous montrerons aussi, comment, grâce à l'analyse syntaxique, TWiC peut identifier des expressions à mots multiples même lorsque les constituants de ces dernières ne sont pas adjacents les uns aux autres.

2 Traduction automatique et outils terminologiques

Le problème qui nous intéresse peut être illustré par l'exemple d'un lecteur francophone lisant la version en ligne d'un magazine en langue anglaise. Ce lecteur, qui a une connaissance de l'anglais suffisante pour effectuer cet exercice, n'est pourtant pas à l'abri d'un déficit terminologique. Pour prendre un exemple concret, imaginons que le terme *hankers* dans la phrase (1) lui soit inconnu.

(1) A segment of the Latin-American left still **hankers** after revolution.
"Un segment de la gauche latine-américaine aspire encore à la révolution".

Si, comme nous, le lecteur soumet cette phrase à quelques-uns des systèmes de traduction disponibles en ligne, voici ce qu'il risque d'obtenir :

(2)a. Un segment de la gauche Latin-Américaine désire toujours après révolution.

b. Un segment du Latino-Américain est parti toujours désire ardemment la révolution.

c. Un segment du Latino-américain gauche calme hankers après révolution la

d. Un segment de gauche toujours des hankers latino-américains après révolution.

Ces exemples, qui illustrent fort bien les insuffisances de la traduction automatique actuelle, montrent du même coup pourquoi la traduction automatique, en l'état, ne peut constituer une réponse adéquate à notre problème[1]. Non seulement la traduction obtenue est souvent peu fiable, grammaticalement ou stylistiquement peu satisfaisante, parfois même incompréhensible, mais même lorsque la traduction de la phrase est satisfaisante, elle oblige le lecteur à parcourir la phrase traduite pour tenter d'identifier le mot (ou l'expression) correspondant au terme source qui faisait problème. Si le mot *hankers* lui pose un problème de compréhension au cours de sa lecture, ce que le lecteur souhaite, c'est une réponse rapide et précise, qui interrompe le moins possible sa lecture et sa concentration.

Les outils de terminologie en ligne (dictionnaires bilingues, Eurodicautom, etc.) peuvent apporter une aide partielle à notre lecteur. Pour un mot sélectionné, ces outils affichent une liste de traductions possibles. Pourtant, en général, ils exigent comme entrée la forme de citation du mot, obligeant du coup notre lecteur à davantage de manipulations qu'il ne le souhaite (de plus, il n'est pas difficile d'imaginer des situations –ou des langues– où le lecteur pourrait être en mal de donner la forme de citation d'un mot qu'il ne connaît pas.). Cependant, le problème le plus sérieux dans l'utilisation d'outils terminologiques en ligne reste celui du bruit. L'ambiguïté et la polysémie du signe linguistique sont telles qu'un bon dictionnaire bilingue est susceptible de contenir plusieurs dizaines de traductions pour un terme donné. Pour prendre un exemple parfaitement banal, si notre lecteur sélectionne le mot *school* dans la phrase (3), il devra sans doute passer en revue de très nombreuses traductions avant d'arriver à celle de *banc* (de poissons).

(3) A **school** of little fishes swam past.
"Un banc de petits poissons passa"

Certains systèmes d'aide terminologique en ligne sont dotés d'une interface morphologique, ce qui dispense l'usager d'entrer la forme de citation pour le mot sélectionné[2]. Le problème, bien sûr, est qu'en dehors de tout contexte la décomposition morphologique surproduit. Ainsi, si notre lecteur sélectionne *rose* dans son article en anglais, il obtiendra une liste de mots comprenant aussi bien les traductions du substantif anglais *rose* que celles du verbe *to rise*, dont *rose* est la forme du prétérit, ainsi que celles de l'adjectif. Autrement dit, si la présence d'un prétraitement morphologique augmente notablement le confort de l'usager en rendant superflue l'entrée de la forme de citation, elle peut entraîner par contre une forte augmentation du bruit,

[1] Voir Hutchins (2003) pour une réflexion sur l'état actuel de la traduction automatique.

[2] C'est le cas par exemple du système commercialisé par Babylon Ltd. Le système COMPASS (Breidt & Feldweg, 1997), développé dans le cadre d'un projet européen au début des années 90, inclut non seulement une analyse morphologique, mais également une recherche d'expressions à mots multiples, basée sur des automates à états finis. Ce système ne semble pourtant pas avoir été développé au-delà du prototype décrit dans la référence ci-dessus, et il ne nous a pas été possible de déterminer le statut actuel de ce système.

due au fait que le système de morphologie s'appliquant hors contexte génère fréquemment des formes de citation non pertinentes par rapport au contexte linguistique du mot sélectionné.

C'est pourtant le traitement des expressions à mots multiples qui constitue le problème le plus sérieux pour les outils terminologiques en ligne par rapport au besoin d'assistance à la lecture en langue étrangère. D'une manière générale, en effet, les expressions à mots multiples ne sont pas traitées de manière adéquate par ces systèmes. Bien qu'un grand nombre de collocations et d'expressions idiomatiques figurent dans tout bon dictionnaire bilingue, elles sont souvent difficiles à localiser, particulièrement lorsque l'usager ne sélectionne qu'un seul des constituants de l'expression.

3 TWiC

3.1 Présentation du système

Le système TWiC (*Translation of Words in Context*) est un traducteur de mots et d'expressions en contexte, qui utilise une analyse syntaxique pour tenter d'identifier au mieux l'unité lexicale sélectionnée et réduire ainsi le bruit lors de la consultation d'un lexique bilingue. TWiC se présente comme un système d'assistance à la lecture de documents en langue étrangère sur Internet[3]. Lorsqu'un utilisateur sélectionne un mot dans un document, TWiC effectue une analyse syntaxique complète et détaillée de la phrase dans laquelle figure le mot. Les contraintes morphologiques et syntaxiques imposées par cette analyse permettent de lever un grand nombre d'ambiguïtés, dont pratiquement toutes les ambiguïtés catégorielles. De plus, la composante morphologique du système fournit la forme de citation de l'unité lexicale sélectionnée. Comme nous le verrons plus bas, une des caractéristiques les plus originales de TWiC est sa capacité à traiter les expressions à mots multiples, que ce soit des mots composés, des expressions idiomatiques ou des collocations[4].

La figure (5) ci-dessous illustre la fenêtre de dialogue lors d'une utilisation de TWiC pour le mot *natural* dans la phrase (4).

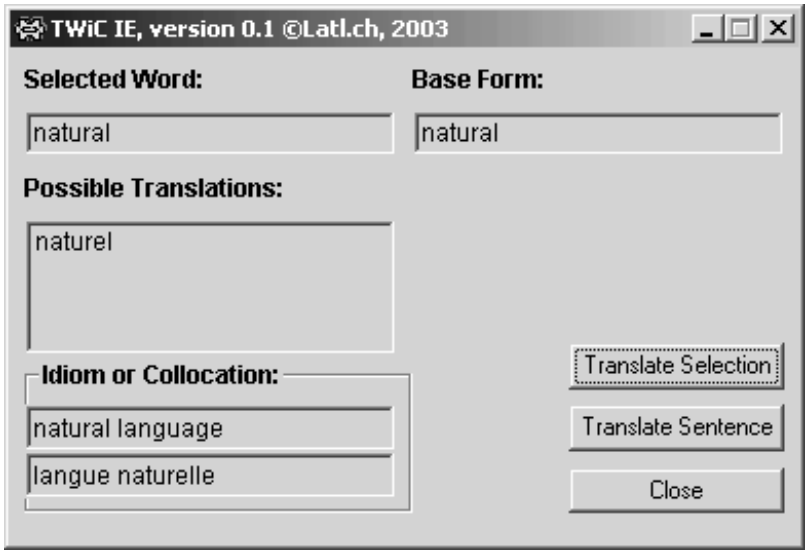
(4) A **natural** language interface was developed.
"Une interface en langue naturelle a été développée"

Plusieurs champs apparaissent dans la fenêtre d'interface TWiC. Dans la partie supérieure, les champs *Selected Word* et *Base Form* correspondent respectivement au *mot sélectionné* par l'usager et à la *forme de citation* calculée par le système. Une liste réduite de traductions est affichée dans le champ *Possible Translations*. Les deux champs consacrés aux expressions à mots multiples (mots composés, expressions idiomatiques et collocations) *Idiom or Collocation* ne sont utilisés que lorsque le système identifie une expression dont le mot sélectionné est un des

[3]Concrètement, il prend la forme d'un "plug-in" pour le butineur Internet Explorer.

[4]Nous reprenons la partition en trois classes des expressions à mots multiples proposée par Goldman et al. 2001. Brièvement, les *mots composés* sont des éléments de catégorie lexicale (nom, verbe, adjectif, adverbe, etc.), dont toutes les parties sont nécessairement adjacentes les unes des autres. Les *expressions idiomatiques* et les *collocations* sont des expressions de niveau syntagmatique (groupe verbal, groupe nominal, etc.), dont les constituants ne sont pas nécessairement adjacents. Alors que les collocations sont simplement des associations conventionnelles de termes, les expressions idiomatiques se caractérisent par un certain degré de figement au plan syntaxique (p.ex. pas de modification ou d'extraction possible) et/ou par une certaine opacité sémantique.

(5) **Interface graphique de TWiC :**



constituants. L'expression source est alors affichée dans le champ supérieur et sa traduction (expression ou mot simple) apparaît dans le champ inférieur. Dans l'exemple (4), l'usager a sélectionné le mot *natural*. TWiC a identifié ce mot et affiche sa forme de citation *natural*, ainsi qu'une traduction possible *naturel*. Dans cet exemple, TWiC a identifié la collocation *natural language* pour laquelle il propose la traduction *langue naturelle*.

3.2 Architecture du système

Le système TWiC se compose de plusieurs modules, dont les plus importants sont :

- L'**extracteur de phrase**, qui extrait du document HTML consulté la phrase dans laquelle figure le mot sélectionné par l'usager. Cette opération est réalisée sur la base d'indices typographiques (p.ex. ponctuation), ainsi que sur la base de balises de mise en page et de structuration du document.
- L'**identificateur de langue**, qui détermine la langue de la phrase extraite, sur la base d'un système classique de tri-grammes.
- L'**analyseur linguistique**, qui effectue une analyse morpho-syntaxique complète et détaillée de la phrase extraite, de manière à identifier l'unité lexicale correspondant au mot sélectionné et dégager ses rapports syntaxiques avec les autres constituants de la phrase.
- La **base de données bilingue**, qui spécifie les correspondances d'unités lexicales (mots et expressions à mots multiples).
- L'**interface graphique**, qui affiche les résultats de la requête.

C'est sans aucun doute l'analyse linguistique qui constitue l'élément noyau de TWiC, et en fait son originalité. L'analyseur linguistique détermine l'unité lexicale concernée par la requête.

Les contraintes morphologiques puis syntaxiques appliquées par ce dernier ont pour effet de lever en bonne partie les ambiguïtés du mot sélectionné. Ainsi, d'une manière générale, les ambiguïtés de nature catégorielle (est-ce un verbe ou un substantif ?), fréquentes pour un mot isolé, disparaissent presque totalement en contexte, réduisant d'autant le problème du bruit mentionné dans la section précédente. TWiC utilise l'analyseur Fips (cf. Wehrli, 1997), configuré de manière à retourner les informations pertinentes pour cette application. Ces informations sont illustrées dans la figure (7) pour le mot *attempt* (*tentative*) dans la phrase (6) :

(6) They foiled an attempt.
"Ils ont déjoué une tentative"

(7) **Analyse morpho-syntaxique :**

Source word	POS tag	Position	Lexeme number	Expression number
they	PRO-PER-3-PLU	0	111000011	
foiled	VER-PAS-3-PLU	5	111016454	141000136
an	DET-SIN	12	111050002	
attempt	NOU-SIN	15	111005034	- 141000136

Dans cette illustration, la première colonne correspond aux mots de la phrase, la deuxième colonne à l'étiquette morpho-syntaxique correspondant au mot, la troisième colonne donne la position du mot dans la phrase (plus précisément la position du premier caractère du mot sélectionné, par rapport au début de la phrase). La quatrième colonne spécifie le numéro de lexème (numéro d'identification pour la base de données). Enfin, la dernière colonne mentionne, lorsqu'il y a lieu, l'appartenance d'un mot à une expression idiomatique ou à une collocation. Le numéro d'identification de l'expression est donné en regard du mot considéré comme tête de l'expression (c-à-d. le gouverneur syntaxique), alors que le (ou les) autres mots participant à l'expression sont accompagnés du même nombre en valeur négative.

Sur la base de ces renseignements, TWiC peut facilement répondre à la requête d'un usager. Le mot sélectionné par l'usager est repéré dans le fichier d'étiquettes sur la base de la position du curseur. L'analyseur a identifié le lexème le plus approprié étant donné le contexte syntaxique. Ce lexème donne donc la forme de référence, qui sert également de clé de recherche pour les correspondances bilingues. Si l'analyseur a identifié une expression, le numéro d'identification de cette dernière sert à la recherche d'un équivalent cible dans le dictionnaire bilingue.

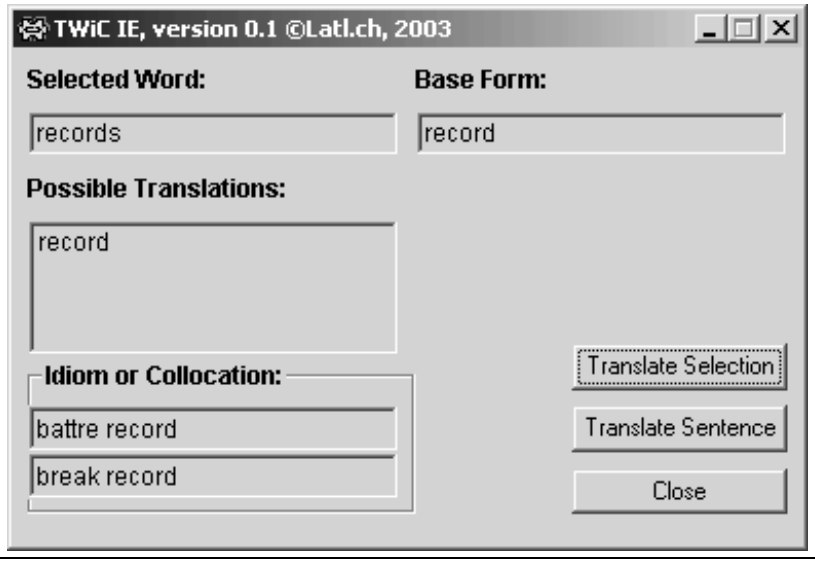
3.3 Traduction d'expressions

Il est important de noter qu'un traitement adéquat des expressions à mots multiples ne peut se satisfaire d'un traitement superficiel de la phrase. Au contraire, il présuppose une analyse complète et particulièrement fine, capable d'identifier et d'interpréter des éléments extraposés, comme par exemple les constituants-*wh* des constructions relatives, les constituants sujets des constructions passives, etc. Ainsi, dans l'exemple (8) ci-dessous, l'identification de la collocation *break-record* (*battre-record*) passe-t-elle par (*i*) la reconnaissance d'une structure relative, dont le pronom (*which*) reprend le substantif *records*, (*ii*) l'interprétation de ce pronom comme sujet de la relative, (*iii*) l'identification d'une structure passive, avec pour corollaire le fait que le sujet apparent est interprété comme objet direct du verbe *break*. C'est bien cette séquence d'événements que l'analyseur construit pour obtenir la structure syntaxique donnée en (8b).

(8)a. Several of the records which were broken during the last Olympic games were at least two-years old.
"Plusieurs des records qui ont été battus lors des derniers Jeux Olympiques dataient d'au moins deux ans."

b. [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ several [$_{\texttt{PP}}$ of [$_{\texttt{DP}}$ records$_i$ [$_{\texttt{CP}}$ which$_i$ [$_{\texttt{TP}}$ [$_{\texttt{DP}}$ e]$_i$ were [$_{\texttt{VP}}$ broken [$_{\texttt{DP}}$ e]$_i$ [$_{\texttt{PP}}$ during the last Olympic games]]]]]] were [$_{\texttt{VP}}$ e [$_{\texttt{FP}}$ at least two-years old]]]]

(9) **Exemple d'une collocation avec extraposition d'un des termes :**



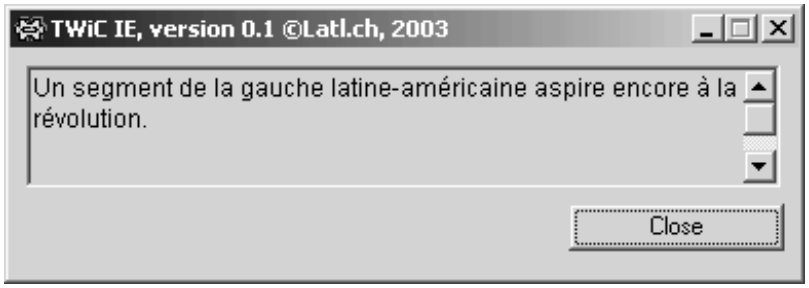
Les avantages procurés par la normalisation des structures qu'effectue un analyseur syntaxique "profond" apparaît clairement dans de tels exemples. Cette normalisation permet d'identifier sans risque d'erreur des collocations (ou des expressions idiomatiques), même lorsque les constituants de ces expressions apparaissent dans un ordre autre que l'ordre canonique et à une certaine distance l'un de l'autre.

3.4 Traduction de phrases

L'objectif principal de TWiC est d'apporter une assistance terminologique contextuelle, comme nous l'avons montré dans les sections ci-dessus. On ne peut exclure, cependant, qu'occasionnellement cette assistance ne s'avère insuffisante pour la compréhension globale d'une phrase, et que le lecteur souhaite dans ce cas une traduction complète de la phrase. C'est pour satisfaire ce besoin que l'interface graphique de TWiC comprend une fonctionnalité *traduction de phrases*, qui permet de déclencher la traduction de la phrase dans laquelle l'usager a sélectionné un mot. L'activation du moteur de traduction est effectuée au moyen de la commande *Translate Sentence ("Traduire phrase")* et ouvre une nouvelle interface graphique, illustrée dans la figure ci-dessous pour la phrase (1), reprise en (10) :

(10) A segment of the Latin-American left still **hankers** after revolution.
"Un segment de la gauche latine-américaine aspire encore à la révolution".

(11) **Exemple d'une traduction de phrase :**



La traduction de phrases proposée par TWiC utilise une version révisée du système Its-2 (cf. Wehrli, 1998), qui reprend l'analyse syntaxique effectuée pour la traduction du mot sélectionné. Même si ce traducteur utilise un moteur syntaxique plus élaboré que ses concurrents commerciaux, il souffre essentiellement des mêmes lacunes que ces derniers. Cette fonctionnalité est donc à utiliser avec beaucoup de réserves.

4 Conclusion

L'assistance terminologique pour la lecture en ligne de documents en langue étrangère est un besoin reconnu, qui va très certainement s'amplifier encore avec la multiplication rapide et l'usage croissant de la documentation en ligne. Un tel système n'entend pas concurrencer la traduction automatique mais bien combler un besoin nouveau conséquence du "village multilingue". L'idée majeure présentée dans cet article est celle de soumettre à une analyse linguistique "profonde" la phrase entière dans laquelle un usager a sélectionné un terme, préalablement à la recherche terminologique. Le recours à une telle analyse présente en effet de nombreux avantages : lemmatisation (analyse / découpage morphologique), levée d'un grand nombre d'ambiguïtés (de nature catégorielle ou morphologique) et surtout capacité à reconnaître l'appartenance du terme sélectionné à une expression figée ou à une collocation, indépendamment de l'ordre relatif ou de l'éloignement des constituants de l'expression. Enfin, l'analyse

linguistique constitue une étape cruciale pour la traduction de la phrase complète, au cas où l'usager le demanderait.

A l'heure actuelle, une version de TWiC existe pour les paires de langues anglais-français et français-anglais, utilisant une base de données bilingue de plus de 50'000 entrées, dont environ 3000 collocations et expressions idiomatiques. Une version allemand-français est en développement.

Remerciements

Plusieurs collaborateurs du LATL ont participé à l'élaboration de composantes du système TWiC, en particulier Violeta Seretan et Luka Nerima, que je tiens à remercier ici. Mes remerciements également à la fondation Boninchi, pour son généreux soutien. L'interface graphique de TWiC pour Internet Explorer a été réalisée en collaboration avec la société Oberon Microsystems SA (http://www.oberon.ch).

Bibliographie

BREIDT, E. & H. FELDWEG (1997), "Accessing Foreign Languages with COMPASS", *Machine Translation* 12, 153-174.

HUTCHINS, J. (2003), "Has machine translation improved? Some historical comparisons", actes de la conférence *MT Summit IX*, New Orleans, 181-188.

GOLDMAN, J.-P., L. NERIMA & E. WEHRLI (2001), "Collocation Extraction Using a Syntactic Parser", *actes de la conférence ACL-2001*, 61-66.

MCLUHAN, M. (1967), *The Gutenberg Galaxy : The Making of Typographic Man*, London, Routledge & Kegan Paul.

WEHRLI, E. (1997), *L'analyse syntaxique des langues naturelles : Problèmes et méthodes*, Paris, Masson.

WEHRLI, E. (1998), "Translating Idioms", *actes de la conférence COLING-98*, Montréal, 1388-1392.

WEHRLI, E. (2003), "Translation of words in context", actes de la conférence *MT Summit IX*, New Orleans, 502-504.