**Meta**
**Journal des traducteurs**
**Translators' Journal**

META

# L'évaluation des systèmes de traduction automatique

Monique C. Cormier

Volume 37, Number 2, juin 1992

URI: https://id.erudit.org/iderudit/002403ar
DOI: https://doi.org/10.7202/002403ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Cormier, M. C. (1992). L'évaluation des systèmes de traduction automatique. *Meta*, *37*(2), 383–384. https://doi.org/10.7202/002403ar

## L'ÉVALUATION DES SYSTÈMES DE TRADUCTION AUTOMATIQUE

Le Groupe d'étude international sur l'évaluation des systèmes de traduction automatique, dirigé par Margaret King de l'Institut Dalle Molle pour les études sémantiques et cognitives (Université de Genève) et Gudrun Magnusdottir de l'Université de Gothenburg, réunissait une quarantaine de chercheurs, développeurs et utilisateurs, dans un forum organisé en Suisse par Kirsten Falkedal, du 21 au 24 avril 1991. Des questions à l'ordre du jour, nous retiendrons les suivantes.

S'agissant du juger si un système de traduction automatique est valable ou non, que faut-il évaluer : la qualité des traductions qu'il produit, sa rapidité d'exécution, son coût, le temps qu'il exige pour la post-édition, sa convivialité, ou encore son potentiel de développement ? Cette première question en entraîne d'autres : que faut-il utiliser pour tester les systèmes : des corpus réels ou des corpus fabriqués ? Serait-il utile d'élaborer une méthodologie générale d'évaluation ?

### L'ADÉQUATION ET LE POTENTIEL DES SYSTÈMES

Pour évaluer la qualité du produit livré par l'ordinateur, on s'est largement inspiré jusqu'à maintenant des méthodes d'évaluation du célèbre rapport ALPAC[1]. Intelligibilité, fidélité, précision du texte, voilà les critères sur lesquels on demande généralement à un groupe de lecteurs de s'appuyer pour évaluer une traduction-machine avec, comme résultats, des appréciations inégales, voire irréconciliables, qu'explique probablement le caractère subjectif et global des critères retenus.

Cela dit, sans renier la valeur des critères énoncés ci-dessus, on se tourne de plus en plus vers des méthodes d'évaluation qui focalisent sur un nombre limité et précis de problèmes linguistiques intégrés à des corpus fabriqués ou artificiels.

Les corpus artificiels (en anglais *test suites*), qui regroupent des phrases dont la construction syntaxique est représentative des textes d'un client, sont très utiles au développeur qui peut, grâce à eux, vérifier la force réelle de son système. Combien d'unités lexicales un corpus à tester doit-il contenir pour qu'il soit considéré comme représentatif ? En réponse à cette question, beaucoup de chiffres ont été lancés, de 4 000 à 60 000. On s'entend cependant pour dire que l'évaluation d'un système général exigera un corpus plus important que celle d'un système appelé à répondre à des besoins limités. Mais ce n'est pas seulement une question de nombre et il ne faudra jamais négliger la qualité des corpus. Si, à la fin des tests, en toute objectivité, le système arrive à bout des problèmes qui lui ont été posés, c'est qu'il est bon.

Dans les cas d'échec, tout n'est pas perdu, puisque le développeur peut essayer d'enrichir son système de nouvelles règles. La capacité ou l'incapacité de celui-ci de les accepter donnera au développeur des indications précieuses sur le potentiel de développement et, par conséquent, sur la force réelle de son système.

### UNE MÉTHODOLOGIE GÉNÉRALE D'ÉVALUATION

Idéalement, il faudrait pouvoir disposer d'une méthodologie générale d'évaluation des systèmes de traduction automatique dont chacun pourrait profiter au lieu de faire ses évaluations de son côté, avec ses critères propres.

En pratique, ce n'est pas possible pour plusieurs raisons. D'abord, les besoins des éventuels acquéreurs de systèmes diffèrent beaucoup de l'un à l'autre, comme varie également la capacité de répondre à des besoins précis. Ensuite, il ne faut pas oublier qu'une évaluation coûte très cher à une entreprise sans qu'elle ait la certitude de trouver satisfaction.

Pas plus que le système multi-usages parfait, la méthodologie générale d'évaluation n'existe pas. Les besoins et les contraintes des clients sont trop différents. La véritable évaluation doit donc s'appliquer à comparer entre eux les résultats de différents systèmes.

***

On sait maintenant que la traduction automatique est la plus utile pour des applications bien précises, c'est-à-dire pour des sous-domaines bien délimités où le volume de traduction est très important. On la considère donc de plus en plus, non pas comme une fin, mais comme un moyen au service du traducteur. Suivant cette logique, les traducteurs exigent des chercheurs que la machine s'adapte à eux et non le contraire.

Le temps où l'on comparait le rendement de la machine à celui de l'être humain semble bien terminé.

MONIQUE C. CORMIER
*Université de Montréal, Montréal, Canada*

Note

1. Après avoir subventionné pendant dix ans la recherche en traduction automatique et y avoir consacré une vingtaine de millions de dollars, les organismes de financement américains ont voulu faire le point. En 1964, l'Académie nationale des sciences des États-Unis créait un comité, l'*Automatic Language Processing Advisory Committee* ou ALPAC, dont le mandat était d'évaluer les résultats obtenus en traduction automatique et, à la lumière de ces résultats, de formuler des recommandations sur la suite à donner au financement de la recherche dans ce domaine.

   L'une des conclusions de l'ALPAC a été la suivante: «[...] *we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation.*» (ALPAC, 1966: *Languages and Machines. Computers in Translation and Linguistics*, Publication 1416, Washington (DC), National Academy of Sciences, p. 32.) Le comité recommanda alors que les fonds de recherche soient plutôt réorientés vers la recherche fondamentale en linguistique computationnelle.