Meta
Journal des traducteurs
Translators' Journal

META

L'évaluation des systèmes de traduction automatique dans le cadre d'un service de traduction

Margaret King

Volume 37, Number 4, décembre 1992

Études et recherches en traductique / Studies and Researches in Machine Translation

URI: https://id.erudit.org/iderudit/003314ar
DOI: https://doi.org/10.7202/003314ar

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

King, M. (1992). L'évaluation des systèmes de traduction automatique dans le cadre d'un service de traduction. *Meta*, *37*(4), 817–827. https://doi.org/10.7202/003314ar

Article abstract

This article considers the problem of evaluating a machine translation system from the viewpoint of the head of a large translation service which is considering whether his service could benefit from the introduction of a system. It is suggested that an analysis of the type of work done by the service is an essential first step, and the importance of the context into which the system is to be integrated is emphasized. A methodology for the evaluation is then sketched out, starting with the establishment of criteria relevant to the particular context, and going on to meet those criteria. The data collection techniques are briefly described, and an indication is given of their strengths and weaknesses.

L'ÉVALUATION DES SYSTÈMES DE TRADUCTION AUTOMATIQUE DANS LE CADRE D'UN SERVICE DE TRADUCTION

MARGARET KING
Université de Genève, Genève, Suisse

Résumé

Cet article considère le problème de l'évaluation d'un système de traduction automatique du point de vue de la personne chargée d'un service de traduction important, qui se demande si son service tirerait profit de l'installation d'un système. L'article montre qu'une analyse du type de travail effectué par le service est une première étape essentielle et insiste sur l'importance du contexte dans lequel le système doit s'intégrer. Une méthodologie pour l'évaluation est ensuite esquissée, d'abord par l'établissement de critères particuliers au contexte, et ensuite par les manières de réunir les données servant de base au jugement porté sur la capacité d'un système à répondre à ces critères. Les techniques de collection de données sont brièvement décrites avec une indication de leurs qualités et défauts respectifs.

Abstract

This article considers the problem of evaluating a machine translation system from the viewpoint of the head of a large translation service which is considering whether his service could benefit from the introduction of a system. It is suggested that an analysis of the type of work done by the service is an essential first step, and the importance of the context into which the system is to be integrated is emphasized. A methodology for the evaluation is then sketched out, starting with the establishment of criteria relevant to the particular context, and going on to meet those criteria. The data collection techniques are briefly described, and an indication is given of their strengths and weaknesses.

1. INTRODUCTION

Une méthodologie pour l'évaluation générale des systèmes de traduction automatique, clairement expliquée, établie et communément acceptée n'existe actuellement pas. Jusqu'à très récemment, il n'y a même pas eu beaucoup de discussions dans la littérature, ni de réflexions sur le thème rendues accessibles à la communauté scientifique. Pour la plupart, les résultats des évaluations faites ont été décrits dans des rapports internes, préparés pour celui qui a commandité l'évaluation, et n'ont pas été divulgués à un public plus vaste. Cela implique non seulement que les résultats spécifiques ne sont pas largement connus, mais aussi que la méthodologie suivie reste souvent cachée.

Cette situation commence à changer pour des raisons à la fois économiques et scientifiques. Le coût des systèmes et des évaluations sérieuses ainsi que le besoin d'établir des standards acceptés par toute la communauté — chercheurs, constructeurs, clients et utilisateurs — provoquent un intérêt croissant pour les questions de méthodologie d'évaluation, qui se concrétise par des efforts spécifiques tels que la création d'un groupe de travail international, l'organisation de réunions sur ce sujet par ce groupe et par l'Association Internationale pour la Traduction Automatique, l'intérêt explicite du DARPA aux États-Unis, et la définition de l'évaluation comme thème de recherche dans les programmes de la Communauté Européenne.

Mais nous sommes encore au tout début de nos réflexions et tout reste à proposer et à débattre. Cela implique que l'évaluation reste un sujet complexe et délicat. De plus, l'évaluation concerne toute une variété de personnes: les chercheurs, les bailleurs de fonds, les entrepreneurs, les clients potentiels et actuels, et, peut-être les plus importants de tous, les utilisateurs condamnés, quelquefois contre leur gré, à l'utilisation des systèmes informatisés. Chacun, parmi tout ce monde, a ses propres besoins et ses propres désirs. En conséquence, il n'est pas du tout évident que les critères d'évaluation soient les mêmes pour tout le monde, et que la méthodologie à suivre soit la même.

Il est donc impossible de résumer ici tous les cas, de discuter leurs besoins en matière d'évaluation, de faire des propositions et de les justifier. Nous allons alors délibérément restreindre la problématique, en nous imaginant dans la situation de quelqu'un qui devrait décider de l'introduction d'un système de traduction automatique dans un service de traduction, que nous imaginerons relativement grand. C'est négliger à tort les traducteurs indépendants ou les services plus petits, mais nous espérons qu'ils trouveront quand même quelque chose d'intéressant et d'utile dans cette discussion.

Nous allons aussi nous limiter en laissant de côté tout ce qui concerne les outils informatisés conçus comme aides à la traduction humaine, tels que les postes de travail des traducteurs, les banques de terminologie électroniques, les dictionnaires informatisés et même les vérificateurs d'orthographe ou de style. Cela ne sous-entend pas que de tels outils ne soient pas importants: si l'on considère les vrais besoins d'un service de traduction, il peut même s'avérer que l'introduction d'outils informatisés serait plus intéressante que l'introduction d'un système de traduction proprement dit (c'est-à-dire un système où une partie ou la totalité du processus de traduction est confiée au système). Mais l'espace est limité, et il faut choisir.

Avant d'attaquer l'évaluation elle-même, nous examinerons brièvement deux questions préalables, dont les réponses définiront le contexte qui entoure l'évaluation proprement dite.

2. LES APPLICATIONS D'ABORD

Si l'on se met à la place de notre décideur, une première question s'impose: est-ce qu'il y a vraiment des travaux entrepris au sein du service concerné où l'introduction d'un système de traduction automatique va faciliter le travail? La question est critique, car il ne faut pas s'imaginer que le système à tout faire, qui peut traduire n'importe quel texte sans erreur et avec une qualité de traduction acceptable, existe ou puisse jamais exister.

En gros, les systèmes qui ont déjà fait leurs preuves sont de deux catégories. Il y a des systèmes taillés sur mesure pour traiter un domaine spécifique avec une langue à but spécifique associé à ce domaine. J'hésite à parler d'une «sous-langue», même si ce terme est très courant: parler d'une sous-langue suggère que la sous-langue soit une partie, un sous-ensemble de la langue en question, et qu'il s'agisse par exemple du français, sauf que les phrases subordonnées ne sont pas permises, ou que l'impératif n'est jamais utilisé. La réalité peut être très différente, et la langue à but spécifique peut présenter des caractéristiques que l'on ne trouvera jamais dans la langue «normale», comme l'énoncé suivant, tiré d'un manuel technique soumis à la traduction automatique:

> «Ajouter un nouvel utilisateur se trouvant chaque fois à un niveau d'utilisation inférieur.»

L'expérience a montré la validité et l'utilité de systèmes conçus pour le traitement d'un domaine spécifique: quelques exemples nous fourniront une idée intuitive des types de textes appropriés: TAUM-Météo qui traduit les bulletins météorologiques au Canada

depuis la fin des années 70, TITUS qui traite des résumés d'articles scientifiques en technologie textile depuis le début des années 70, et plus récemment, IFENA qui traite les bulletins d'avalanches suisses.

Si une partie du travail du service implique donc la traduction de textes qui utilisent un vocabulaire limité et un nombre limité de structures syntaxiques, on pourrait envisager, peut-être, l'automatisation de cette partie du travail. Mais il est très peu probable que l'on puisse prendre un des systèmes déjà existants et le modifier pour traiter un nouveau domaine. Sauf dans des cas très rares, les différences entre les deux vocabulaires vont nécessiter un travail considérable au niveau du lexique, et les différences entre les structures syntaxiques vont demander une réécriture des règles de grammaire. Le plus probable est qu'il vaudrait mieux construire une nouvelle description linguistique plutôt que d'essayer de modifier une description construite pour le traitement d'un autre domaine de discours.

Cela ne veut pas dire qu'il n'y a aucun profit à tirer des systèmes déjà existants, et à se demander quel type de système serait approprié. Les trois systèmes cités ont des bases théoriques très différentes : TITUS fait partie de tout un système de gestion des informations, impose un contrôle très strict sur le texte entré dans le système, traduit, pour ainsi dire, ce texte dans une représentation indépendante des langues permises comme langues d'entrée et de sortie, et génère à nouveau le texte à partir de cette représentation. TAUM-Météo est basé sur le transfert, mais se sert d'une utilisation assez limitée de la sémantique sous-jacente au domaine traité. IFENA dépend très étroitement d'une modélisation sémantique du monde des avalanches, mais utilise un module de transfert pour tenir compte des variations syntaxiques entre les deux langues. L'adéquation d'un de ces modèles à une application spécifique est déterminée par un jeu de facteurs comme la faisabilité d'une modélisation sémantique (le monde du domaine de discours est-il sémantiquement clos ?), la possibilité d'imposer un contrôle sur l'entrée (les textes sont-ils rédigés sur place ? les rédacteurs accepteraient-ils un tel contrôle ? si non, est-il envisageable de faire un contrôle informatisé qui sépare les phrases que le système peut traiter des phrases qu'il ne sait pas traduire et envoyer ces dernières aux traducteurs humains ?) ou la possibilité de compenser les faiblesses du système automatique par une intervention humaine (est-ce qu'une révision des traductions produites serait acceptable ? est-ce qu'un système interactif, où le système demande à l'utilisateur de résoudre quelques problèmes d'ambiguïté ou de choix d'équivalents dans la langue cible, serait envisageable ?).

On voit tout de suite que certaines de ces questions sont d'ordre technique et que leur réponse exige des connaissances spécialisées, et que d'autres sont d'ordre pratique, voire psychologique, et que la réponse sera basée sur une connaissance du service de traduction, de ses travaux et de ses traducteurs. Cela est typique de l'évaluation dans la réalité : si on envisage l'intégration d'un système de traduction automatique dans un service de traduction, il ne suffit pas de trouver le bon système du point de vue technique, il faut aussi tenir compte de l'environnement dans lequel le système sera utilisé.

La deuxième classe de systèmes contient les systèmes qui sont plus généraux dans le sens qu'ils n'ont pas été conçus pour un type spécifique de textes. Leurs constructeurs espèrent qu'ils seront capables de fournir une traduction d'une qualité suffisamment élevée et avec un taux d'erreurs acceptable pour toute une gamme de textes.

L'emploi du futur dans la phrase précédente est tout à fait délibéré ; il n'est presque pas pensable que l'on puisse acheter un système, même parmi les plus performants et les mieux établis, et produire des traductions de très bonne qualité tout de suite. Pour comprendre pourquoi, il suffit de remarquer qu'il est très peu probable que le système tel qu'il est vendu sur le marché incorpore déjà toute la terminologie nécessaire aux besoins d'un service de traduction spécifique. Dans la pratique, il est de plus fort probable qu'il y

ait aussi des particularités de la langue au niveau des structures utilisées dont le traitement ne soit pas non plus prévu.

Voilà qui nous ramène à la question des applications ; il est possible qu'il y ait une partie du travail du service où une traduction «rapide mais brute» serait acceptable. Le cas classique d'une telle application est l'utilisation du système GAT, condamné par le rapport ALPAC en 1966 pour ses insuffisances, mais utilisé ensuite pendant presque dix ans par des scientifiques confrontés au problème de devoir se tenir à jour dans une littérature spécialisée écrite dans une langue étrangère : pour leurs besoins, l'important était de pouvoir identifier parmi des centaines d'articles ceux qui étaient important, qui méritaient une étude approfondie et donc, une traduction humaine, le cas échéant.

Si de telles applications n'existent pas dans le service, on se retrouve dans la situation qui vient le plus naturellement à l'esprit des gens quand ils entendent parler de l'évaluation des systèmes de traduction automatique. Un ou plusieurs systèmes existent sur le marché, mais le système parfait n'existe probablement pas. On veut décider d'abord s'il est faisable d'utiliser l'un ou l'autre des systèmes disponibles pour faciliter une partie du travail, s'il est rentable de le faire, et, finalement, si cela est souhaitable.

Après une courte digression sur quelques applications où les réponses à ces questions sont un «non» clair et net, on va consacrer la suite de cet article à l'esquisse d'une méthodologie pour trouver les réponses.

3. LES APPLICATIONS À EXCLURE

Personne n'essayerait d'utiliser un système de traduction automatique pour traduire les œuvres de Racine. Mais il y a d'autres types de textes où il est un peu moins évident que la machine ne sera pas capable de les traiter. On peut distinguer en gros trois catégories : les textes où les mots ou les phrases ambiguës le sont délibérément, les textes où les questions de style sont critiques et les textes où la traduction doit tenir compte non seulement de ce qui est explicitement dit, mais aussi des implications de ce qui est dit.

Nous n'avons pas la place ici pour expliquer le pourquoi de ces trois catégories ; le lecteur intrigué trouvera une réflexion plus approfondie dans King (1987). L'important est de constater que si les traductions entreprises par le service de traduction concernent des textes dans une de ces trois catégories, il ne vaut même pas la peine d'envisager l'évaluation d'un système de traduction automatique.

La grande littérature est exclue par ces trois critères. Les textes à but persuasif, comme les discours politiques ou la publicité, sont aussi exclus par l'exigence que l'ambiguïté ne soit pas voulue. Des traductions où on aimerait en quelque sorte capturer la voix de l'auteur du texte source, même s'il ne s'agit pas de la grande littérature — certaines lettres, par exemple, ou certains textes de vulgarisation — sont exclus par le deuxième critère. Certains textes légaux, où la traduction devrait préserver les inférences aussi bien que le contenu explicite, sont exclus par le troisième.

Ces cas sont donnés à titre d'exemple ; il y en a bien d'autres. Si nous y insistons un peu, c'est parce qu'il y a une pré-évaluation à faire avant qu'on ne pose la première question concernant les systèmes à évaluer, «y a-t-il un système qui traduise entre les langues qui m'intéressent ?»

Cette pré-évaluation est une réflexion sur la situation actuelle, sans laquelle l'évaluation n'a aucun sens.

4. ENCORE DE LA PRÉ-ÉVALUATION

Après avoir décidé qu'il est possible d'appliquer la traduction automatique entre les divers travaux exécutés dans un service de traduction, il faut encore se poser certaines questions.

On a déjà cité la plus évidente : il est loin d'être le cas que des systèmes soient disponibles sur le marché pour tous les couples de langues qu'on trouve couramment dans les services de traduction. S'il est raisonnablement facile de trouver des systèmes traduisant entre l'anglais ou le japonais et une autre langue, il est déjà plus difficile de trouver un système qui traduise entre l'italien et l'allemand, et impossible, à ma connaissance, de trouver un système qui traduise entre le portugais et le grec. La question mérite d'être posée très tôt : si la réponse est négative, l'évaluation est déjà finie.

La deuxième question à se poser concerne les buts de l'introduction d'un système informatisé : que veut-on obtenir exactement ? Les réponses possibles peuvent être très variées, et peuvent avoir une influence critique sur les critères d'évaluation. Si l'on veut que le travail soit accompli plus vite, le temps de calcul sera critique pour les applications où la traduction brute est utilisée, si on prévoit une révision de la traduction brute, le temps de calcul, plus le temps requis pour la révision, devront être pris en compte. Si l'on veut que le travail soit plus rentable, il sera important de comparer le coût de la traduction informatisée avec le coût de la traduction humaine, un calcul loin d'être simple quand on considère les questions d'amortissement, le coût de l'entretien ou des modifications au système nécessaires et d'autres questions de cet ordre. Peut-être envisage-t-on d'utiliser un système informatisé pour accomplir des traductions qui ne seront tout simplement pas faites autrement, et la vitesse de la traduction a beaucoup moins d'importance. Peut-être y a-t-il des textes très ennuyeux, et on veut améliorer le sort des traducteurs humains en leur épargnant le fardeau de les traduire. Il serait impossible de donner ici une liste de toutes les considérations potentiellement pertinentes : dans un cas précis, l'important est de dresser une telle liste avant de commencer l'évaluation, pour qu'on puisse voir clairement quelles sont les considérations primordiales et l'importance relative des autres considérations prises en compte.

Une autre série de questions, d'importance égale, concerne les contraintes sur l'acceptabilité d'un système. Il se peut que le service soit déjà informatisé, par exemple, et un système qui n'est pas compatible avec le matériel et le logiciel déjà en place est automatiquement exclu. Ce type de considération est important, parce qu'il tend à éliminer certains systèmes d'office, épargnant ainsi les peines d'une évaluation détaillée, et présente l'avantage supplémentaire que la décision ne dépend que d'une connaissance du service et du contexte qui l'entoure.

Dans le même ordre d'idées, des considérations économiques fournissent d'autres contraintes. Le coût réel d'un système est parfois difficile à calculer, puisqu'il faut prévoir non seulement le prix initial mais aussi, dans la plupart des cas, le prix des modifications nécessaires pour que le système devienne vraiment utilisable, et le prix d'entretien, sans oublier le coût de la formation des utilisateurs. Mais on peut au moins fixer le montant maximum qu'on serait prêt à payer.

Les critères pratiques et économiques ainsi définis permettent un premier tri des systèmes disponibles. Si un ou plusieurs systèmes survivent à ce premier tri, on est presque prêt à considérer comment évaluer les candidats. Il reste une dernière question à se poser : combien peut-on se permettre d'investir dans l'évaluation même ? Jusqu'ici la plupart du travail était un travail de réflexion, entrepris par la ou les personnes connaissant bien les circonstances du service de traduction. Pour la suite, il sera probablement prudent, sinon nécessaire, de faire appel à un expert technique, qui connaît bien les systèmes de traduction automatique. On verra aussi que certains des tests qui fournissent des données préalables à la formation d'un jugement éclairé exigent des dépenses en temps et en personnel non négligeables. L'évaluation peut alors coûter cher, et il faut bien déterminer d'avance combien on peut y investir — et noter peut-être qu'une erreur dans l'achat d'un système risque de coûter encore plus cher.

5. PREMIÈRE ÉTAPE : L'ÉTABLISSEMENT DES CRITÈRES

Toutes les considérations pratiques et économiques, et les réflexions sur le travail entrepris par le service auront amené à la formulation de quelques critères auxquels un système acceptable devrait satisfaire. En gros, on peut distinguer quatre groupes de critères, ayant trait à :

- la qualité de la traduction produite ;
- les capacités du système par rapport aux caractéristiques spécifiques des textes à traduire ;
- la performance du système ;
- la facilité avec laquelle le système peut être modifié.

Prenons chacun de ces groupes tour à tour.

Sauf dans des cas bien spécifiques, comme quand un système a été conçu et construit pour le traitement d'un domaine limité associé à une langue à but spécifique qui ne permet pas beaucoup de variations dans sa façon d'exprimer les choses, on a déjà remarqué que l'on ne peut s'attendre à ce que le système produise toujours des traductions acceptables. Et même quand on se retrouve dans un des cas spécifiques, il va falloir contrôler la qualité de la traduction. Quand on pense à la qualité d'une traduction, le premier aspect qui vient à l'esprit est sa fidélité au texte de départ. Ensuite vient sa clarté ou son intelligibilité (les deux mots sont utilisés plus ou moins dans le même sens dans la littérature sur l'évaluation), et finalement son style. Évidemment, ces mots ont un sens beaucoup moins riche dans le cas d'un système de TAO que dans le cas d'un traducteur humain ; la fidélité, par exemple, est plutôt limitée à l'absence de contre-sens qu'au choix judicieux d'un équivalent qui reflète toutes les nuances de l'original, et le style est plutôt une question de la capacité de produire le style approprié à un procès verbal ou à un résumé d'un article scientifique, qu'à celle de créer un texte agréable à lire. Mais même en tenant compte de ces limitations, il est très difficile, voire carrément impossible, de donner une définition très nette de ce qu'on veut dire par ces mots. De ce fait, les jugements portant sur ces aspects comportent inévitablement un côté subjectif, comme nous le verrons clairement ci-dessous, quand nous examinerons les tests visant à l'acquisition des données pertinentes.

Le deuxième groupe de critères concerne le comportement du système par rapport à certaines tâches spécifiques, typiquement la traduction de certains types de textes. Ici on peut distinguer l'adéquation du vocabulaire (y compris la terminologie appropriée), la couverture des structures syntaxiques, et le traitement des problèmes sémantiques. Le premier est intuitivement clair. Le système est destiné à traduire des textes dans des domaines spécifiques ; il est important de savoir si le vocabulaire approprié est déjà inclus dans les dictionnaires du système ou s'il va falloir l'ajouter. Il sera peut-être utile d'illustrer les deux autres aspects par un cas concret. Il y a un certain temps, l'ISSCO expérimentait la traduction d'un bulletin d'offres d'emploi (Buchmann *et al.* 1984). Les textes à traduire étaient en allemand, mais une version française peut donner une idée de leur spécificité. Au niveau de la syntaxe, il n'y avait que quatre ou cinq verbes, normalement des verbes vagues, comme «être», «faire», et très peu de phrases complètes. Une annonce typique était, par exemple : «Secrétaire de langue maternelle française avec bonnes connaissances d'allemand. Études secondaires et certificat de capacité fédéral.» Dans l'évaluation d'un système de TAO destiné à traduire des textes pareils, on est très fortement intéressé par sa capacité à traiter des phrases incomplètes, mais pas du tout par son traitement de phrases longues et complexes. Au niveau sémantique, les textes à traduire contenaient beaucoup de noms composés, qui posent un problème notoire pour la traduction automatique. (Comparez les deux analyses possibles de «Kulturgeschichte» comme

Kultur + Geschichte (histoire de la culture) et Kult + Urgeschichte (pré-histoire du culte), exemple tiré de Hutchins et Somers, 1992). Mais, par ailleurs, les syntagmes prépositionnels montraient très peu de variations, et un autre problème notoire ne se présentait donc pas. Tout ceci revient à dire qu'il faut affiner le deuxième groupe de critères en examinant de très près la nature linguistique des textes à traduire.

Quant à la performance du système, on pourrait être intéressé soit par la vitesse de traduction, considérée indépendamment quand il s'agit d'une application du type «rapide et brute», soit par le temps nécessaire pour compléter tout le chemin dès l'arrivée d'un texte à traduire jusqu'au moment où la révision et la mise en page finale soient accomplies. (Il ne sert pas à grand-chose que le système produise cinq cent pages de traduction brute par heure s'il faut trois heures de travail humain pour qu'une de ces pages soit utilisable.) Selon les circonstances, il peut y avoir d'autres critères de performance à ajouter; par exemple, si le système doit tourner sur une machine utilisée pour d'autres programmes, l'espace mémoire requis peut être important. Et presque toujours l'interface utilisateur sera d'une importance considérable. Aucun système ne peut être efficace et performant s'il est doté d'une interface qui rend son utilisation difficile ou désagréable.

Comme nous l'avons déjà remarqué, il est très rare qu'un système disponible sur le marché soit utilisable tel quel, sans modification ou extension pour l'adapter aux besoins spécifiques du service. Pour cette raison, l'aisance avec laquelle le système peut être modifié est normalement très importante. Au minimum, même si l'on peut définir tout le vocabulaire actuellement utilisé, des additions au dictionnaire risquent de se révéler nécessaire avec le temps. Il faudra prévoir son élargissement. Il est rare aussi qu'un système traduise sans faute; la correction des erreurs peut impliquer des changements aux entrées lexicales existantes, quand il s'agit de problèmes de vocabulaire, ou la modification de la description linguistique qui traite des règles de grammaire quand il s'agit de structures syntaxiques, ou encore des problèmes d'interprétation sémantique. Un autre type de modification souhaitable dans certains cas est l'extension du système pour obtenir un traitement des phénomènes non encore traités. Encore une fois, ces modifications peuvent concerner soit le vocabulaire, soit la description syntaxique ou sémantique. Au moment de formuler des critères de jugement du système, on ne peut évidemment pas savoir quel type de modification va être pertinent, ni à quel degré. Mais il est utile quand même d'avoir précisé par avance combien de modifications sont acceptables en principe.

Tout au début de cet article, nous avons remarqué que les besoins peuvent varier énormément selon les applications voulues et le contexte particulier du travail. La définition des critères, et l'établissement de priorités relatives entre les critères choisis est une façon de déterminer les besoins spécifiques.

6. DEUXIÈME ÉTAPE : COLLECTE DE DONNÉES

Une fois qu'on a dressé une liste des critères pertinents pour le cas particulier, et établi leur importance relative — lesquels sont critiques, lesquels sont souhaitables, lesquels sont agréables mais pas strictement nécessaires — on peut procéder à la définition de tests visant la collecte de données qui vont permettre de juger l'adéquation du système par rapport à des critères spécifiques.

On n'entrera pas ici dans une description détaillée des tests connus ou proposés; on se contentera d'une brève esquisse organisée autour des groupes de critères suggérés dans le chapitre précédent.

6.1. LA QUALITÉ

Les tests le plus souvent associés avec les critères touchant la qualité impliquent la définition d'une échelle, où chaque point sur l'échelle est accompagné d'une définition et

d'une note. Pour donner un exemple concret, une des échelles utilisée dans l'évaluation qui a conduit au rapport ALPAC (ALPAC 1966) concerne l'intelligibilité de la traduction. Les notes vont de 9 à 1 (le 9 est le meilleur), et on pourrait traduire quelques-unes des définitions des points sur l'échelle comme suit :

> 9. Parfaitement clair et intelligible. À la lecture, apparaît comme un texte normal ; il n'y a pas de tournures stylistiquement boiteuses ...
> 4. Fait semblant d'être une phrase intelligible, mais en réalité manque plutôt d'intelligibilité. Néanmoins, on peut saisir vaguement l'idée. Le choix des mots, l'ordre des expressions syntaxiques, et / ou le choix d'expressions de remplacement est en général bizarre, et il se peut que des mots importants n'aient pas trouvé une traduction ...
> 1. Sans espoir. Il apparaît qu'on n'arrivera jamais à déceler la pensée exprimée dans la phrase, même après une étude approfondie et beaucoup de réflexion.

Chaque membre d'un groupe de sujets (des personnes parfois choisies pour être typiques des utilisateurs potentiels, parfois choisies selon d'autres critères) attribue ensuite des notes à un échantillon de traductions produites par le système. Le responsable de l'évaluation utilise ces notes pour arriver à un jugement sur la qualité obtenue, vis-à-vis des aspects choisis (l'intelligibilité, dans notre exemple).

On voit tout de suite comment intervient l'élément de subjectivité. Chacun des sujets doit décider quelle note attribuer en partant d'une définition qui permet un degré non négligeable d'interprétation. Cette faiblesse est inévitable, étant donné l'impossibilité de formuler une définition précise de mots comme «fidèle», «intelligible», «clair». Si on essaie de compenser la subjectivité en multipliant le nombre de personnes qui attribuent les notes, l'administration des tests et l'interprétation des résultats risquent de devenir coûteuses. Il faut noter aussi que les compétences de ces personnes peuvent être différentes selon l'aspect de la qualité considérée : on peut difficilement juger de la fidélité d'une traduction si on ne comprend pas l'original, mais la même exigence n'existe pas nécessairement pour juger de sa clarté. Nous n'avons pas la place ici pour discuter les faiblesses et les avantages des méthodes basées sur l'utilisation des échelles ; une discussion plus détaillée se trouve dans (Wilks 1979 et Caspari 1987).

6.2. ADAPTATION AUX TÂCHES SPÉCIFIQUES

Si nous nous tournons maintenant vers les critères qui touchent aux capacités du système à traduire des textes spécifiques, nous trouvons deux méthodes de collecte de données.

La première exige un travail préalable d'analyse des textes pour déterminer quelles sont leurs caractéristiques les plus importantes, soit à cause de leur fréquence dans les textes, soit à cause du fait que leur traduction correcte est critique. Sur cette base, un jeu de tests est établi, c'est-à-dire une série de petits textes artificiellement construits, où chacun est destiné à tester le comportement du système face à un phénomène particulier. Supposons, par exemple, qu'on veuille savoir si le système est capable de traiter correctement un verbe comme *demander*, qui a la particularité qu'on demande *à* quelqu'un *de* faire quelque chose. On simplifie le reste de la phrase au maximum, pour éviter toute interaction avec d'autres phénomènes syntaxiques ou sémantiques, afin de fabriquer une entrée comme *Jean demande à Marie de chanter*, dont on contrôle la traduction.

Les jeux de tests de ce type ont l'avantage d'être très précis et très objectifs, mais ils ont deux désavantages : ils sont longs et complexes à construire, et les tests individuels peuvent devenir très nombreux, ce qui implique un temps considérable pour les soumettre au système et pour l'analyse des résultats. Le travail de définition peut être raccourci s'il existe déjà des jeux de tests développés ailleurs pour les langues concernées ; il y a

quelques initiatives dans cette direction (voir, par exemple, Nerbonne *et al.* 1992), mais malheureusement elles sont encore assez rares.

Si l'on décide d'éviter le travail délicat et fastidieux que la construction d'un jeu de tests exige, ou si l'on estime que la précision presque chirurgicale offerte par de tels tests n'est pas utile dans un contexte donné, une autre façon d'examiner les capacités d'un système par rapport à une tâche particulière est de prendre un échantillon des textes typiques de cette tâche et d'en faire un «test corpus». Le raisonnement sous-jacent à cette approche est l'hypothèse voulant que si l'on prend une quantité de texte suffisamment grande, tous les phénomènes intéressants, pour lesquels on aurait construit des entrées artificielles dans un jeu de tests, vont apparaître naturellement dans le corpus. Le premier désavantage de l'approche vient directement de là; pour être sûr que le raisonnement soit valable, il faut que l'échantillon soit grand, et il n'est pas toujours facile de trouver un corpus assez grand sous une forme lisible par la machine. La quantité de texte nécessaire implique aussi un temps considérable pour l'analyse des résultats; on reviendra sur cette question plus en détail ci-dessous.

Une solution intermédiaire entre la construction d'un jeu de tests et l'utilisation d'un corpus typique du domaine d'intérêt, est de constituer un corpus artificiel, mais tiré des textes à traiter, soigneusement constitué pour refléter les caractéristiques les plus importantes des textes. Même si cette solution peut paraître attrayante parce que les composants du corpus sont plus naturels que les entrées d'un jeu de tests, une étude linguistique des textes est indispensable pour préparer le corpus. En outre, si l'on se propose de faire une analyse détaillée des résultats, celle-ci risque d'être compliquée par le fait qu'une interaction entre différents phénomènes linguistiques à l'intérieur d'une seule phrase ne peut pas être *a priori* exclue.

Dans les trois cas — jeu de tests, corpus naturel et corpus artificiel — si on est intéressé aussi par les possibilités d'amélioration du système, il est fort utile de préparer en même temps une deuxième série d'entrées, contenant des cas parallèles à ceux qui seront soumis au système. Cette deuxième série servira, après une modification éventuelle du système pour corriger les fautes rendues apparentes par les résultats obtenus à partir de la première série, comme moyen de contrôler les conséquences des modifications. Cela est important pour deux raisons; d'abord une erreur peut être représentative de toute une classe d'erreurs potentielles. Le contrôle permet de savoir si seul le cas qui s'était présenté dans la première série d'entrées a été corrigé, ou si toute la classe se comporte correctement après la modification. Deuxièmement, il se peut que la correction d'une erreur provoque de nouvelles erreurs ailleurs; le contrôle permet aussi de s'informer sur ce point.

Revenons maintenant à l'analyse des résultats. On peut envisager une analyse quantitative ou une analyse qualitative. La première possibilité est plus souvent adoptée quand on a utilisé des corpus pour la collecte des données brutes. Une mesure fréquemment utilisée est d'évaluer le temps nécessaire pour qu'une traduction utilisable puisse être produite pour un certain nombre de pages du texte source, c'est-à-dire le temps requis pour parcourir tout le chemin depuis l'entrée du texte à traduire jusqu'à la fin, y compris le processus de traduction, la post-édition ou révision, la mise en page, etc. Quelquefois, faire l'expérience avec le corpus entier coûterait trop cher; dans ce cas, évidemment, il faut s'assurer que les extraits du corpus choisis pour l'expérimentation soient représentatifs, et, en tout cas, il faut prendre des précautions pour minimiser autant que possible le biais subjectif de ceux qui font la révision. Une autre approche qu'on trouve dans la littérature sur l'évaluation est de compter les erreurs. Mais ici on rencontre la difficulté qu'il est parfois difficile de déterminer combien d'erreurs on a trouvé. Par exemple, dans la phrase *Je l'ai vues*, il est clair que l'accord du participe avec le complément d'objet

préposé est incorrect. Mais si la bonne phrase aurait dû être *Je les ai vus*, est-ce qu'on compte une erreur ou deux ? Le même type de difficulté se présente sous une forme légèrement différente si on veut aller encore plus loin et classer les erreurs, suivant l'idée que, intuitivement, certaines erreurs sont plus graves que d'autres. On n'a pas la place ici pour examiner ce problème d'une façon détaillée, mais le même exemple illustre le problème. Peut-être un mauvais traitement de la morphologie est-il à l'origine de l'erreur, mais il se peut également que le problème soit d'ordre sémantique et vienne d'une mauvaise interprétation de la référence du pronom (Van Slype 1979).

6.3. PERFORMANCE

La plupart des tests suggérés jusqu'ici ont exigé des connaissances techniques en linguistique, en linguistique informatique ou en traduction pour leur mise en place. Les tests qui touchent la performance d'un système, sont, par contre, plutôt une question de bon sens, de sensibilisation à l'informatique et aux aspects humains de l'ergonomie.

La vitesse avec laquelle un système produit une traduction brute, par exemple, est déterminée en partie par l'efficacité des logiciels, en partie par les exigences en mémoire relatives à l'espace mémoire disponible, en partie par l'utilisation éventuelle des spécificités d'un type ordinateur ou d'un système de gestion donné.

Sauf dans les cas où l'on envisage d'utiliser la traduction brute comme produit final (l'application que nous avons appelée «rapide et brute» plus haut), la vitesse du processus de traduction n'est pas très intéressante en soi ; c'est plutôt le temps requis pour l'achèvement du produit final qui est important. Des facteurs externes, comme la possibilité de l'utilisation d'aides informatisées pendant la révision, ou le manque d'une bonne mise en page automatique intégrée au système, peuvent être de très grande importance à ce propos. Et c'est ici aussi qu'il faut examiner de près les questions de possibilités d'intégration du système avec un parc informatique déjà existant, si celles-ci sont une condition préalable à l'acceptabilité d'un système.

Il ne faut pas non plus oublier l'importance du confort de l'utilisateur du système : nous avons déjà remarqué qu'un système ne peut pas être efficace si tout le monde déteste l'utiliser. Il n'y a pas de règles établies pour ce qui constitue une bonne interface utilisateur ; le mieux est de faire des expériences, en se rappelant qu'un certain temps de familiarisation est nécessaire et que certaines facilités qui attirent l'utilisateur débutant peuvent devenir frustrantes pour un utilisateur plus expérimenté.

6.4. L'EXTENSIBILITÉ

L'extensibilité a en réalité deux aspects ; la modification du système pour corriger les fautes de traduction produites par le système actuel et l'extension du système pour qu'il traite de nouveaux phénomènes. La facilité avec laquelle le deuxième type d'extension peut être accompli est assez difficile à juger si on n'a pas accès aux bases théoriques et techniques du système, et les compétences nécessaires pour les juger. (Quand il s'agit de l'achat d'un système commercialisé, il est rare que le vendeur accepte de révéler les bases techniques de son système, sauf dans ses très grandes lignes.) Même pour la modification, un accès aux bases technologiques ou au moins aux résultats intermédiaires produits pendant le processus de traduction facilitera l'estimation de sa faisabilité. Mais on peut, dans ce dernier cas, contourner en quelque sorte le problème si on peut établir un accord avec le vendeur selon lequel on établit un autre type de classification d'erreurs, selon leur facilité de réparation. Il sera prudent de contrôler, si possible, le réalisme de cette classification en demandant la réparation d'un échantillon des erreurs ainsi classées. De la même façon, s'il y a des extensions que le vendeur estime ne pas être trop coûteuses, on pourrait envisager de lui demander de les faire, simplement pour voir combien de temps est en réalité nécessaire.

7. CONCLUSION : LA VRAIE ÉVALUATION COMMENCE

En fin de compte, seule la personne qui connaît bien les vrais besoins du service de traduction, l'environnement du travail et les contraintes imposées par le contexte spécifique peut juger si un système de traduction automatisé sera d'une aide précieuse ou un encombrement qui coûte cher. Tout ce qui précède n'a servi qu'à suggérer quelques facteurs à prendre en considération et quelques méthodes connues pour recueillir les informations nécessaires pour qu'une décision puisse être prise en pleine connaissance de cause.

BIBLIOGRAPHIE

ALPAC (1966) : *Language and Machines: Computers in Translation and Linguistics*, publication 1416, Washington, National Academy of Sciences.

BOESEFELDT, Katarina et Pierrette BOUILLON (1992) : «Une représentation sémantique et un système de transfert pour une traduction de haute qualité», *Coling 92*, Nantes.

BOUILLON, Pierrette et Katarina BOESEFELDT (1992) : «Problèmes de traduction automatique dans le sous-langage des bulletins d'avalanches», *Meta*, 37-4, Montréal, Presses de l'Université de Montréal.

BUCHMANN, Beat, WARWICK, Susan et Patrick SHANN (1984) : «Design of a Machine Translation System for a Sublanguage», Stanford, *Coling 84*.

CASPARI, Gabriele (1987) : «Untersuchungen zu Bewertungskriterien für maschinelle erstellte Uebersetzungen», *Unveröffentliche Diplomarbeit*, Universität des Saarlandes.

DUCROT, J. M. (1973) : «Le système TITUS II», *A.R.N.T. Information et documentation 4*, pp. 3-40.

HUTCHINS, John W. (1986) : *Machine Translation: Past, Present, Future*, Chichester, Ellis Horwood, 382 p.

HUTCHINS, John W. and Harold L. SOMERS (1992) : *An Introduction to Machine Translation*, Academic Press.

ISABELLE, Pierre (1987) : «Machine Translation at the TAUM Group», Margaret King (Éd.), *Machine Translation Today*, Edinburgh, Edinburgh University Press.

KING, Margaret (1987) : *A Tutorial on Machine Translation*, ISSCO WP, n° 53.

KING, Margaret (Éd.) (1987a) : *Machine Translation Today*, Edinburgh, Edinburgh University Press.

KING, Margaret and Kirsten FALKEDAL (1990) : «Using Test Suites in the Evaluation of Machine Translation Systems», Helsinki, *Coling 90*.

NERBONNE, John, NETTER, Klaus, DIAGNE, Abdel Kader, KLEIN, Judith and Ludwig DICKMANN (1992) : «A Diagnostic Tool for German Syntax», Manuscript, *Deutsches Forschungszentrum für Künstliche Intelligenz*, GmbH Stuhlsatzenhausweg 3, D-6600 Saarbrücken 11.

VAN SLYPE, Georges (1979) : *Critical Study of Methods for Evaluating the Quality of Machine Translation*, Bruxelles, Bureau Marcel Van Dijk and CCE.

WILKS, Yorick and LATSEC inc. (1979) : *Comparative Translation Quality Analysis*, Final Report, Contract F33657-77-0695, LATSEC inc.