Pour l'évaluation externe des systèmes de TA par des méthodes fondées sur la tâche

Hervé Blanchon, Christian Boitet

Laboratoire LIG, équipe GETALP, BP 53, 38041 Grenoble Cedex 9
Herve.Blanchon@imag.fr, Christian.Boitet@imag.fr

RÉSUMÉ. *Les méthodes externes d'évaluation de systèmes de TA définissent des mesures de qualité à partir des résultats de TA et de leur usage. Alors que les systèmes opérationnels sont depuis longtemps le plus souvent évalués par des méthodes fondées sur la tâche, les campagnes d'évaluation des dernières années utilisent (parcimonieusement) des méthodes subjectives assez chères fondées sur des jugements humains peu fiables, et (pour la plus grande part) des méthodes basées sur des traductions de référence, impossibles à utiliser lors de l'utilisation réelle d'un système, d'autant moins corrélées aux jugements humains que la qualité augmente, et totalement irréalistes en ce qu'elles forcent à mesurer les progrès sur des corpus fixes, sans cesse retraduits, et non sur de nouveaux textes à traduire pour des besoins réels. Il y a aussi de nombreux biais introduits par le désir de diminuer les coûts, en particulier l'utilisation de corpus parallèles dans le sens inverse de leur production et l'utilisation de juges monolingues au lieu de bilingues. Nous prouvons cela par une analyse de l'histoire de l'évaluation en TA, des méthodes d'évaluation du « courant dominant », et de certaines récentes campagnes d'évaluation. Nous proposons d'abandonner les méthodes fondées sur des traductions de référence en évaluation externe, et de les remplacer par des méthodes strictement fondées sur la tâche, en les réservant à l'évaluation interne.*

ABSTRACT. *External methods for evaluating MT systems define various measures based on MT results and their usage. While operational systems are mostly evaluated since long by task-based methods, evaluation campaigns of the last years use (parsimoniously) quite expensive subjective methods based on unreliable human judgments, and (for the most part) methods based on reference translations, that are impossible to use during the real usage of a system, less correlated with human judgments when quality increases, and totally unrealistic in that they force to measure progress on fixed corpora, endlessly retranslated, and not on new texts to be translated for real needs. There are also numerous biases introduced by the desire to diminish costs, in particular the usage of parallel corpora in the direction opposed to that of their production, and of monolingual rather than bilingual judges. We prove the above by an analysis of the history of MT evaluation, of the « mainstream » evaluation methods, and of certain recent evaluation campaigns. We propose to abandon the reference-based methods in external evaluations, and to replace them by strictly task-based methods, while reserving them for internal evaluations.*

MOTS-CLÉS : *évaluation, traduction Automatique, évaluation d'utilisabilité.*

KEYWORDS : *evaluation, machine translation, usability evaluation.*

1. Introduction

Depuis 2000, l'évaluation des systèmes de traduction automatique (TA) a pris une importance considérable au sein de la communauté des chercheurs en TA. Cela concerne surtout la TA statistique de l'écrit et de l'oral, parce que les méthodes d'évaluation utilisées sont dérivées de celles qui ont fait leurs preuves en reconnaissance de la parole. De nombreuses campagnes d'évaluation compétitives ont été organisées, soit directement par les bailleurs de fonds (DARPA, UE, Académie des Sciences Chinoise), soit à leur incitation et avec leur support (NIST, ELDA/ELRA pour la campagne CESTA), soit par des consortiums dans le cadre de projets coopératifs (CSTAR en 1999, projet NESPOLE! en 2002 et 2003, CSTAR de nouveau avec IWSLT depuis 2004, TC-STAR[1] en 2006).

Il s'agit essentiellement de méthodes fondées sur des « références », c'est-à-dire sur des traductions produites par des experts humains : elles consistent à faire des calculs de distance ou de similarité entre les traductions de référence et les traductions « candidates » produites par TA. On fait certes un peu d'évaluation « subjective » de la « fluidité » et de « l'adéquation » (critères les plus courants), mais hors de tout contexte applicatif réaliste. En particulier, on ne juge que des énoncés isolés, et pas l'ensemble d'un dialogue ou d'un texte. À part l'étude d'ATR, en 1999 sur la TA de dialogues oraux de réservation touristique (Sugaya *et al.*, 2001), les évaluations sont faites uniquement sur les résultats de la TA et pas par rapport à la tâche, les deux principaux types de tâche étant la diffusion de qualité d'une information, et la compréhension suffisante pour atteindre un but précis.

Cette pratique permet de mesurer les progrès de systèmes construits par apprentissage à partir de corpus (TA statistique ou analogique, par exemple), mais seulement sur ces corpus, et seulement par rapport aux mesures utilisées. Or, on sait depuis au moins 2004 que les mesures objectives dominantes (BLEU, NIST…) sont de plus en plus mal corrélées avec les mesures « subjectives » lorsque la qualité « subjective » augmente. Et on sait aussi depuis 1972, que la qualité « subjective » n'est pratiquement pas corrélée avec l'utilité perçue par les utilisateurs (les chercheurs atomistes d'Euratom à Ispra donnaient 18/20 d'utilité à Systran russe-anglais, et les traducteurs experts lui donnaient 2/20 de qualité linguistique). Ce phénomène est aussi décrit dans (Church *et al.*, 1993). (Callison-Burch *et al.*, 2006) donnent des arguments théoriques forts qui montrent que ces mesures ne peuvent pas être satisfaisantes. Enfin, dans le cas de la parole, il est notoire que les transcriptions de monologues ou de dialogues interprétés sont jugées comme des traductions de très mauvaise qualité. Or l'interprétariat est très difficile, et bien payé, car son résultat est très utile : utilité et qualité linguistique ne sont pas bien corrélées.

Cela a plusieurs conséquences dommageables. D'abord, il faut changer de mesure quand un système devient opérationnel. Ensuite, cela exclut des compétitions des systèmes construits de façon « experte », peut-être excellents, mais

[1] Les résultats de la campagne 2006 de TC-STAR sont résumés dans (Mostefa *et al.*, 2006).

non adaptés aux corpus de test (comme MÉTÉO ou MedSLT), ou supposant une certaine interaction homme-machine. Enfin, les chercheurs visent à optimiser leurs systèmes pour les mesures (BLEU, NIST, etc.) utilisées par ces campagnes. Quand on passe à des énoncés du même domaine mais plus longs, ou plus « spontanés » à l'oral, les performances diminuent de façon catastrophique, comme on l'a constaté lors de la campagne IWSLT-06 (résultats très inférieurs à ceux de IWSLT-05).

Comment en est-on arrivé là, et que peut-on proposer pour améliorer la situation ? Dans la première section, nous examinerons les pratiques et études antérieures. Des projets assez importants ont en effet été financés pour travailler spécifiquement sur l'évaluation des systèmes de traitement des langues naturelles et de TA. On mentionnera deux études de la JEIDA au Japon (1992-93), et plus récemment les projets EAGLES et FEMTI en Europe, qui ont produit des recommandations très intéressantes pour des évaluations constructives et raisonnées. Dans la seconde section, nous examinerons précisément les méthodes d'évaluation « dominantes » actuelles, de façon à dégager les raisons théoriques et pratiques qui expliquent l'inadéquation des mesures actuelles. Nous utiliserons l'expérience acquise durant trois campagnes d'évaluation en TA de l'oral (projet NESPOLE!, IWSLT-04, IWSLT-06) pour montrer qu'il y a aussi de nombreux biais introduits par les conditions expérimentales. Dans la dernière section, nous proposerons l'abandon pur et simple des mesures fondées sur les références, au moins en *évaluation externe*, et l'adoption de mesures liées à la tâche, inspirées de celles utilisées dans des systèmes opérationnels, adaptées aux contextes actuels d'utilisation de la TA de l'écrit comme de l'oral, et intégrables à coût marginal nul dans les applications.

2. Évolution des idées en évaluation de la TA de l'écrit puis de l'oral

Après un bref historique, nous analysons plus en détail les méthodes d'*évaluation externe* des systèmes de traduction effectivement pratiquées, puis l'évolution vers la « méta-évaluation » et la « formalisation », qui s'est accompagnée d'un élargissement au TALN en général.

2.1 Faits marquants en évaluation de la TA

L'histoire de l'évaluation en TA remonte à ses débuts. On dit parfois, en plaisantant à peine, qu'elle a donné lieu à plus de publications que la TA elle-même. On a toujours distingué entre *évaluations externes*, jugeant la qualité des résultats sur des critères linguistiques (grammaticalité, fidélité, etc.) ou sur des critères d'usage (productivité, coût), et *évaluations internes*, jugeant la conception des systèmes (architecture linguistique et architecture calculatoire) et leurs perspectives d'amélioration et d'extension à de nouvelles langues, à de nouveaux types de documents, et à de nouvelles tâches (e.g. de l'assimilation à la dissémination).

Fin 1966, le rapport ALPAC (ALPAC, 1966), fondé sur une évaluation contestable et contestée des systèmes de TA d'alors -sauf curieusement la version la plus récente du système le plus proche (le Georgetown Automatic Translation ou GAT) - eut une conséquence importante, l'arrêt presque total du financement de la recherche en TA aux USA pour près de 20 ans (jusqu'à 1985), et par ricochet en Angleterre et au Japon, ce dernier pays y revenant vers 1980, à cause des besoins très importants en automatisation de la traduction liés à l'isolement du japonais, à sa difficulté, et à son importance commerciale grandissante.

La TA opérationnelle continua cependant, et les systèmes russe-anglais GAT puis Systran furent régulièrement évalués (à la Wright-Patterson Air Force Base et à Ispra, EURATOM) selon deux critères, la *qualité linguistique* et l'*utilité pratique*, avec des résultats totalement opposés : lors du séminaire organisé en 1972 à Austin, Texas, on parlait de 2/20 (E) en qualité linguistique et de 18/20 (A) en utilité. En tout cas, le rapport ALPAC jeta les bases des protocoles d'évaluation subjective (reposant sur des jugements humains et non sur des mesures de performance en temps, ni sur des comparaisons automatiques avec des *références*).

Le Canada, l'URSS, la France et l'Allemagne continuèrent leurs recherches malgré le rapport ALPAC, quelques sociétés de TA émergèrent aux USA, les plus connues étant Systran et Logos, et quelques recherches y persistèrent (à Berkeley sur le chinois-anglais, et à Austin grâce à des financements de l'USAF puis de Siemens).

En Europe, les besoins en traduction de la CEE justifièrent l'achat puis le développement de systèmes Systran à partir de 1976. Ce fut le début de l'essor de l'évaluation de la TA en tant que domaine économique, puis scientifique.

Plusieurs consultants et cabinets d'audit (bureau Van Slype, Omnium…) proposèrent de nombreuses mesures non liées à la tâche, mais à la *qualité linguistique* (fidélité, grammaticalité, parfois couverture) des résultats bruts et à la structure interne des systèmes, pour mesurer leur potentiel d'amélioration et le coût de cette amélioration (suites de tests, etc.). Ces mesures font essentiellement appel à des jugements subjectifs par des experts humains.

Les chercheurs commencèrent aussi à s'y intéresser, surtout au Canada, à cause du projet TAUM-météo (1974-75) puis du projet TAUM-aviation (1976-81). En effet, la mise en service opérationnel du système TAUM-météo, le 24 mai 1977, fut accompagnée d'une mesure de la *qualité liée à la tâche*, détaillée plus loin. Cependant, cette mesure ne fut pas « théorisée » à l'époque.

Après la fin du programme Eurotra (1982-92), bien des chercheurs, n'ayant plus de financement pour continuer en TA, se convertirent à l'évaluation de la TA. L'histoire se scinde alors en deux volets divergents. D'une part, depuis que la TA est utilisable et utilisée (sur PC et/ou par réseau) par des traducteurs professionnels indépendants, la mesure du « temps gagné », essentiellement liée à la tâche, est très

utilisée. D'assez nombreuses publications en témoignent[2], mais elle n'est pas à la mode chez les chercheurs, car d'une part elle est trop simple (!), et d'autre part elle ne s'applique par définition pas à des maquettes, prototypes, ou systèmes encore non opérationnels.

De plus, les chercheurs ont raffiné à l'extrême les critères de qualité linguistique, tout en cherchant des méthodes automatiques pour les évaluer, de façon à arriver à des évaluations vues comme « scientifiques » et moins coûteuses, car ne faisant pas appel à des juges humains. Avec l'avènement de la TA statistique (vers 1990), on a cherché des méthodes reposant sur la comparaison automatique des traductions automatiques avec des traductions humaines de référence, par analogie avec ce qu'on faisait (à juste titre et avec succès) pour la reconnaissance de parole. En 2002, IBM proposa la méthode BLEU, bientôt suivie de beaucoup d'autres (NIST en 2002, etc.), dites *objectives*.

Malheureusement, l'espoir initial que ces mesures étaient bien corrélées à la qualité des traductions a été déçu : elles sont assez bien corrélées aux jugements humains de fluidité ou d'adéquation quand les traductions sont mauvaises ou très mauvaises (par rapport à ces critères), mais ne le sont plus du tout quand la qualité augmente, au point où d'excellentes traductions humaines sont jugées très mauvaises, et inférieures à des traductions automatiques quasiment incompréhensibles ! Il y a diverses raisons à cela, bien présentées en 2006 par (Callison-Burch *et al.*, 2006), spécialistes en TA statistique. Seule WNM (Babych *et al.*, 2004a) fait exception, mais elle est très peu utilisée.

C'est pourquoi ces deux courants se rejoignent actuellement, avec la mise en œuvre dans le projet GALE d'une mesure liée à la tâche, HWER. Nous expliquerons plus loin pourquoi, à notre avis, elle est biaisée (pour qu'on puisse essayer de la « prédire » à partir de mesures comme BLEU), et par suite inutilement compliquée, coûteuse, et de plus moins adéquate que le traditionnel « temps gagné » pour mesurer la qualité perçue par les utilisateurs relativement aux tâches principales[3], la compréhension, la communication et la production de traductions de haute qualité par des locuteurs de la langue cible connaissant la langue source.

2.2 Historique des méthodes d'évaluation externe

2.2.1. *ALPAC, the (in-)famous report*[4]

Le premier effort d'évaluation ayant connu une large publicité (ALPAC, 1966) est resté célèbre. (Pankowicz, 1966), en charge du financement de la TA à l'USAF, rédigea un contre-rapport extrêmement percutant. (Hutchins, 2003) en a aussi fait

[2] Voir le site `http ://www.geocities.com/mtpostediting/` de J. Allen.
[3] Nous revoyons le lecteur à la section 4.1 pour une définition plus détaillée des tâches.
[4] Titre de l'article de (Hutchins, 2003) qui fait une analyse point par point du rapport.

une assez bonne synthèse. Le rapport ALPAC est bref, 34 pages, il est complété par 20 annexes qui totalisent 90 pages. En particulier, l'annexe 10 (ALPAC, 1966, pp. 67-75) rend compte de l'évaluation conduite.

Deux critères, largement repris par la suite, furent proposés pour évaluer une traduction, son *intelligibilité*, et sa *fidélité*. Ils sont indépendants, car une traduction peut être intelligible et manquer de fidélité ou d'exactitude (précision), et inversement, elle peut être fidèle et manquer d'intelligibilité, bien que cela ne se produise que si le texte source est peu intelligible, comme dans le cas des brevets.

L'intelligibilité de la traduction était évaluée sans référence à la phrase source. La fidélité était mesurée de façon indirecte : le juge devait d'abord assimiler tous les éléments de sens présents dans la traduction, puis évaluer la phrase originale du point de vue de son informativité par rapport à ce qu'il avait compris de la traduction. Ainsi, la phrase source est très informative par rapport à la traduction lorsque cette dernière n'est pas très fidèle, puisque la phrase source contient plus d'information que la traduction.

2.2.2. *La première évaluation liée à la tâche : TAUM-météo et MÉTÉO (1977)*

Durant les quatre ou cinq premières années d'exploitation de ce système, spécialisé aux bulletins météorologiques (à l'exclusion des situations et des avertissements météo), J. Chandioux améliora les grammaires sémantiques du sous-langage des bulletins météo anglais, la traduction se faisant vers le français. Lors de la postédition, les traducteurs ne voyaient que les phrases refusées par le système. La qualité était mesurée par (100-#opérations)/100 pour 100 mots traduits. Au départ, on en était à 45 % de corrections (le remplacement d'un mot coûtant deux opérations, une suppression et une insertion), soit 55 % de qualité. En 1985, on était arrivé à 85 % de qualité, et les traducteurs avaient demandé et obtenu de réviser les traductions complètes.

J. Chandioux réécrivit alors le système dans le langage GramR (version déterminisée des systèmes-q de Colmerauer). Enfin, il écrivit un système français-anglais. La qualité des deux systèmes monta assez vite à 97 %, toujours pour la même mesure. Depuis 1990 environ, MÉTÉO (Chandioux, 1988) a traduit environ vingt millions de mots par an d'anglais en français, et dix millions de mots par an dans l'autre direction, avec cette qualité.

2.2.3. *Travaux de la JEIDA (1989-92)*

L'impact du rapport ALPAC fut indéniable. En particulier, il conduisit à stopper la recherche en TA non seulement aux USA, mais aussi au Japon. Cependant, le besoin en TA, nié par le rapport ALPAC, était très réel au Japon. Après avoir réussi l'informatisation de leur langue, en particulier grâce aux méthodes de saisie « kana-kanji » ou « romaji-kanji » nécessitant de bons segmenteurs, de gros dictionnaires et même des analyseurs locaux, les firmes informatiques japonaises se remirent à la TA dès 1978. Une trentaine de systèmes commerciaux virent le jour, en sus de

systèmes internes orientés vers la recherche, mais de très grande taille et opérationnels (MU du projet national japonais à l'université de Kyoto, ALT/JE de NTT, un autre à NKH, la télévision japonaise). Cependant, les systèmes commerciaux n'arrivèrent pas à un niveau de rentabilité commerciale satisfaisant. D'où l'intérêt de la JEIDA (Japanese Electronic Industry Development Association) pour l'évaluation.

Les premiers travaux de la JEIDA devaient permettre de répondre à trois questions. Quels sont les changements technologiques et sociaux du marché de la TA depuis le rapport ALPAC ? À la lumière de ces changements, les conclusions du rapport ALPAC sont-elles encore valides aujourd'hui ? Sinon, comment doit-on évaluer l'état actuel et le futur de la traduction automatique ? Sous la direction de H. Nomura, un groupe d'experts commença par étudier les systèmes alors en usage dans différents pays et conduisit des enquêtes sur la demande en traduction au Japon, l'état de l'activité en traduction et l'utilisation des systèmes de TA. Mais le rapport (JEIDA, 1989) n'apporta aucune réponse claire aux questions initiales.

La JEIDA commandita une seconde étude plus précise en 1991-92. Cela faisait dix ans que le premier système commercial de TA avait été mis sur le marché japonais, et environ trente systèmes étaient commercialisés. Le rapport (JEIDA, 1992) présente une méthodologie d'évaluation en trois volets : évaluation économique par les utilisateurs, évaluation technique par les utilisateurs, et évaluation technique par les développeurs. Les mesures associées sont très détaillées (quatorze axes, concernant la qualité linguistique, l'ergonomie, le coût de portage à un nouveau domaine, l'utilité pour les traducteurs, etc.). Pour chaque système commercial disponible au Japon, le rapport donne un diagramme « radar » montrant les quatorze valeurs obtenues.

2.2.4. *Campagnes ARPA (1992-1994) et projet « MT proficiency Scale »*

Une évaluation « à la JEIDA » est évidemment plus coûteuse que la mesure de productivité de TAUM-MÉTÉO, mais elle apporte bien plus de renseignements aux décideurs et aux développeurs. L'ARPA (Advanced Research Projects Agency, actuelle DARPA, agence militaire de financement de la recherche aux USA) envoya une mission d'experts au Japon en 1993 pour savoir où le Japon en était en TA, et était donc parfaitement au courant de la méthodologie de la JEIDA. Pourtant, elle ne l'utilisa pas dans les campagnes qu'elle finança par la suite.

Dans ces campagnes d'évaluation (White *et al.*, 1994), l'objectif fut en effet l'évaluation comparative de systèmes commerciaux et de systèmes de recherche utilisant différentes architectures linguistiques, et traduisant tous de diverses langues source vers l'anglais. Or, on ne peut pas mesurer un système de TA non opérationnel selon la majorité des quatorze axes de la JEIDA. D'autre part, il ne s'agissait pas que de TA, mais aussi d'aide à la traduction (THAM[5]). Enfin, dans le

[5] Traduction Humaine Aidée par la Machine, MAHT en anglais.

cadre du projet « MT proficiency Scale », les seules tâches finalement considérées (Taylor *et al.*, 1998, White *et al.*, 2000) furent la compréhension et ses variantes, en vue du renseignement (civil et militaire)[6]. Dans ce contexte, on utilisa deux critères proches de ceux d'ALPAC, la *compréhensibilité* et la *qualité de la traduction.*

Pour la première évaluation, un ensemble d'articles de presse traitant de fusions et d'acquisitions financières fut collecté, puis traduit par des traducteurs professionnels vers les différentes langues source visées. On fit « retraduire » ces articles en anglais par les systèmes de TA, et on évalua ces traductions inverses.

Ce processus est intrinsèquement inadéquat, car on ne peut pas considérer les textes en langue étrangère comme de vrais textes source, pour plusieurs raisons. La plus simple est que toute traduction augmente la longueur d'un document (de 10 à 20 % selon le couple de langues). Une autre est qu'elle appauvrit la variété du style et du vocabulaire par rapport à ceux du document source.

La *qualité de la traduction* était basée sur une mesure utilisée par le gouvernement américain pour évaluer la compétence de traducteurs humains. Cette mesure s'avéra inopérante à cause de la différence en nature et en quantité des erreurs présentes dans les traductions automatiques et dans les traductions humaines. Elle ne fut pas retenue pour les évaluations suivantes. On la remplaça alors par deux critères : *adéquation* et *fluidité*.

La mesure d'adéquation est en quelque sorte inverse de celle d'ALPAC : il s'agit pour un juge anglophone de déterminer le degré auquel les « unités d'information » présentes dans une traduction professionnelle (et non pas dans le texte source, puisqu'il a justement été produit à partir de l'anglais) peuvent être retrouvées dans la sortie d'un système de TA. Une « unité d'information » est un constituant syntaxique qui contient assez d'informations pour permettre une comparaison.

Fluidité n'est pas synonyme de compréhensibilité : on demande seulement aux (mêmes) juges de déterminer si la traduction est rédigée dans un anglais convenable, même si son sens est incorrect ou obscur[7], et sans voir l'original ou une traduction de référence, pour que l'adéquation ne rentre pas en ligne de compte. En pratique, on juge donc d'abord la fluidité, puis l'adéquation.

La mesure de l'adéquation étant fondée sur la comparaison d'une traduction professionnelle avec une traduction produite par un système, (White *et al.*, 1994) font les deux remarques suivantes, toujours actuelles. La première est que la compétence du traducteur humain responsable de la traduction interfère sans doute avec la validité de la mesure, la seconde dit que la mesure de fluidité repose entièrement sur un jugement subjectif peu fiable, car le jugement humain sur le

---

[6] *Snap judgment, gisting, triage, extraction, filtering, detection,* et rien en *publishing*.

[7] La fluidité peut très bien ne pas être pénalisée, lorsque le sens est obscur comme dans les énoncés suivants : « il casse son cigare » ou « a meal will be served 30 minutes before takeoff » (*before* au lieu de *after*) (trouvé dans les données de l'évaluation IWSLT 2006).

caractère bien formé d'un énoncé peut varier, ce qu'ils avaient constaté en faisant appel à plusieurs juges.

2.3 Méta-évaluation, formalisation, et extension à d'autres applications

On s'est alors mis à « évaluer les méthodes d'évaluation », ce qui a conduit à des efforts de formalisation[8], et d'extension à d'autres applications du TALN. Inversement, des méthodes d'évaluation développées pour la recherche d'information et pour la reconnaissance de parole ont fait leur apparition.

Nous situons l'époque de la maturité du domaine vers les années 1995, en particulier avec le projet EAGLES dont les efforts pour formaliser et organiser l'évaluation des systèmes de Traitement Automatique des Langues Naturelles (TALN) furent poursuivis dans le cadre du projet ISLE.

2.3.1. *Le projet EAGLES (1993-1996)*

Le projet européen EAGLES (Expert Advisory Group on Language Engineering Standards) chercha à créer des standards dans le domaine des industries de la langue. Les domaines concernés dans sa première phase (1993-95) furent les corpus, les lexiques, les formalismes grammaticaux et les méthodes d'évaluation.

La première idée de base fut qu'une stratégie unique d'évaluation ne pouvait pas être développée, même pour un domaine d'application spécifique. Les auteurs du rapport (EAGLES-EWG, 1996) affirmèrent ainsi *qu'il serait plus profitable de mesurer si une traduction est assez bonne pour un besoin spécifique, plutôt que d'essayer de définir une notion, nécessairement trop abstraite, de qualité d'une traduction en général.* C'est dire que la qualité d'un système dépend fortement de son usage et de ses utilisateurs potentiels.

La seconde idée de base fut qu'il était nécessaire et possible de proposer un cadre général qui permettrait de définir rapidement et de façon cohérente des évaluations particulières adaptées à chaque application dans son ou ses contextes bien identifiés. Influencé par les travaux précédents, et en particulier par le standard ISO/IEC 9126 (King, 2003), le groupe de travail sur l'évaluation proposa un « modèle de qualité » pour les systèmes de TALN en général. Ce modèle est organisé en une hiérarchie de propriétés et d'attributs dont les feuilles sont des attributs mesurables auxquels sont associées des méthodes de mesure. Il a été validé sur différentes familles de systèmes, dont les correcteurs grammaticaux (EAGLES-EWG, 1996, Annexe D, p. 116) et les aides aux traducteurs humains (id. Annexe E, p. 136).

---

[8] (Van Slype, 1979), (King, 1996), (EAGLES-EWG, 1999, section 3.3) et (Hovy *et al.*, 2002) en dressent un panorama.

La seconde phase du projet (1995-1996) (EAGLES-EWG, 1999) fut consacrée à la consolidation des résultats concernant les correcteurs grammaticaux [id. section 5.1, p. 60], les correcteurs orthographiques [id. section 5.2, p. 76] et les mémoires de traduction [id. section 5.4, p. 106]), et à la dissémination de la proposition.

2.3.2. *Le projet ISLE et la proposition FEMTI*

Les activités du groupe de travail sur l'évaluation du projet EAGLES se poursuivirent dans le cadre du projet ISLE[9] (International Standards for Language Engineering, 1999-2002) financé par l'Europe et par les USA (NSF). Les efforts portèrent sur l'évaluation des systèmes de traduction automatique et donnèrent lieu à une série d'ateliers[10] et à la proposition FEMTI (Framework for the Evaluation of MT in ISLE) (Hovy *et al.*, 2002).

FEMTI propose :

- une classification des caractéristiques principales définissant un contexte d'usage : type d'utilisateur(s), type de tâche(s), type des données d'entrée ;
- une classification des caractéristiques de qualité des systèmes, dans une hiérarchie organisée en sous-catégories dotées, au niveau des feuilles, d'attributs internes et/ou externes (avec leurs métriques), et dont les niveaux supérieurs coïncident avec les caractéristiques définies dans le standard ISO/IEC 9126 (ISO/IEC, 2001) ;
- une correspondance entre ces deux classifications, permettant de trouver les caractéristiques et sous-caractéristiques de qualité adaptées à chaque contexte d'usage, ainsi que les attributs et métriques associés.

La méthodologie FEMTI est destinée à aider plusieurs catégories de personnes :

- les utilisateurs potentiels de TA peuvent définir les caractéristiques les plus importantes pour eux, et choisir le système le mieux adapté à leurs besoins ;
- les utilisateurs potentiels et les développeurs désirant comparer plusieurs systèmes de TA peuvent choisir les caractéristiques les plus proches de leur situation, et déterminer les mesures d'évaluation et les tests pertinents ;
- les développeurs de systèmes de TA peuvent identifier les besoins des utilisateurs et trouver des niches pour leurs applications.

3. Explication et critique des méthodes utilisées dans les campagnes actuelles

Dans les campagnes actuelles, on distingue deux classes de méthodes d'évaluation. Pour la première classe, l'évaluation est réalisée par des juges humains qui doivent donner des notes (scores) en répondant à des questions. Le résultat de

[9] Site web du projet : http ://www.issco.unige.ch/projects/isle/femti/

[10] Consulter http ://www.issco.unige.ch/projects/isle/ewg.html pour la liste complète.

l'évaluation donne alors une évaluation concrète directement interprétable, par exemple « bon » ou « mauvais ». On parle d'*évaluation subjective* car ce sont des jugements de qualité pour lesquels ont observe des accords imparfaits entre juges et de mauvais accords pour un même juge au cours du temps.

Dans les méthodes de la seconde classe, les notes sont calculées par un programme informatique qui produit un résultat numérique à partir des traductions de référence et des traductions candidates fournies en entrée. On parle alors d'*évaluation objective,* car les mêmes entrées reçoivent toujours les mêmes notes.

3.1 Fluidité et adéquation

En 2003, le NIST a proposé un protocole d'évaluation subjective (évaluation conduite par des juges humains) fondé sur les critères ARPA. Des *juges monolingues* notent des traductions candidates selon deux critères : leur *fluidité* et leur *adéquation* par rapport à une traduction de référence.

3.1.1. *Fluidité*

Les consignes d'évaluation du NIST[11] indiquent que la fluidité est le degré de bonne formation linguistique de l'énoncé du point de vue des règles de l'anglais écrit standard. Un segment est bien formé si :

- il respecte la grammaire usuelle ;
- il contient des mots correctement orthographiés (morphologie) ;
- il respecte l'usage commun du vocabulaire, des titres et des noms ;
- il est intuitivement acceptable ;
- il peut être raisonnablement interprété par un locuteur natif de l'anglais.

La fluidité d'un énoncé s'évalue sur une échelle de cinq valeurs (Table 1).

3.1.2. *Adéquation*

L'adéquation d'un énoncé est définie comme suit : « Après avoir fait l'évaluation de fluidité, le juge peut voir l'une des traductions de référence. En comparant la traduction candidate en langue cible avec la référence, il détermine si la traduction est adéquate. L'adéquation concerne le degré auquel les informations présentes dans l'original[12] sont bien transmises par la traduction. Ainsi, pour l'adéquation, la traduction de référence tient le rôle de la source. »

---

[11] Les instructions proposées par le NIST sont accessibles à l'adresse suivante : `http ://projects.ldc.upenn.edu/TIDES/Translation/TransAssess02.pdf`

[12] Il y a ici abus de langage puisque ce n'est pas la phrase originale en langue source qui est présentée au juge, mais une traduction humaine de référence. La validité de la démarche repose donc sur la présupposition que la traduction de référence est déjà elle-même adéquate.

Les juges répondent, sur une échelle de cinq valeurs, à la question : quelle proportion du sens exprimé dans la référence est aussi présente dans la traduction ?

Note	Fluidité (définition)	Adéquation (réponse)
5	Anglais sans défaut	Tout
4	Bon anglais	Presque tout
3	Anglais non natif	Beaucoup
2	Anglais non standard	Peu
1	Incompréhensible	Aucun

Table 1. *Notes NIST pour la fluidité et l'adéquation*

3.1.3. *Critique par la pratique*

3.1.3.1. *Évaluation subjective dans le cadre du projet NESPOLE!*

Le projet NESPOLE! (Lavie *et al.*, 2006) met en situation de dialogue un agent touristique italophone, et un client américanophone, francophone ou germanophone. La traduction se fait *via* un pivot sémantique appelé IF construit pour la tâche[13]. Les scénarios ont été définis grâce à des collectes de données à partir de situations réelles avec de vrais agents touristiques. Nous avons réalisé deux démonstrateurs : le premier en « tourisme réduit » (2001), le second en « tourisme étendu » (2002).

Nous avons conduit une évaluation subjective des deux démonstrateurs pour mesurer à la fois leurs performances brutes, ainsi que les progrès accomplis entre le premier (Rossato *et al.*, 2002) et le second (Blanchon, 2004 ; Blanchon *et al.*, 2004a). Pour le second démonstrateur, nous avons évalué la préservation du sens[14], pour les paires de langues français-italien (tours de parole d'un client) et italien-français (tours de parole d'un agent de voyage), en utilisant une échelle de notation à quatre valeurs qui permet de définir trois catégories[15], et des juges bilingues recrutés en dernière année d'école de traduction, que nous avons formés pour la tâche[16].

---

[13] La situation est asymétrique : les Italiens traduisent en IF des tours de parole d'agent touristique, les Français, les Allemands et les Américains traitent des tours de parole de clients, tous les partenaires produisant depuis l'IF des tours de parole d'agent et de clients.

[14] L'évaluation n'était pas faite au niveau d'un tour de parole complet, mais au niveau des SDU (Simple Dialogue Unit), une SDU étant un segment de tour de parole pouvant être représenté par une structure IF. Dans les données à évaluer, la segmentation en SDU avait été faite par des « experts » de l'IF.

[15] Les quatre valeurs sont : VERY GOOD (toutes les informations sont présentes et faciles à

Dans le cadre de la tâche d'évaluation proprement dite, les juges disposaient de deux copies d'un même fichier de données témoins. Ils devaient évaluer ce fichier avant et après l'évaluation des fichiers de données proprement dits. Cela nous a permis de mesurer, (1) l'accord entre les juges avant et après l'évaluation effective, et (2) la stabilité du jugement de chacun au cours du temps.

		U	VM	¬VM
juges français	Avant	71	28	1
	Après	73	27	0
juges italiens	Avant	88	15	0
	Après	75	25	0

Table 2. *Accord entre juges (en %)*

		S	~S	¬S
juges français	J1	57	26	17
	J2	81	13	6
	J3	64	17	19
juges italiens	J4	81	13	6
	J5	100	0	0
	J6	57	26	17

Table 3. *Stabilité des juges (en %)*

Pour mesurer l'accord des juges (Table 2), nous distinguons trois catégories : les juges ont le même jugement (unanimité, **U**), les juges s'accordent sur deux jugements qui sont dans la même catégorie (vote majoritaire, **VM**), les juges ne s'accordent pas sur la même catégorie (pas de vote majoritaire, **¬VM**). Pour mesurer la stabilité de chacun des juges dans le temps (Table 3), nous avons aussi trois catégories : le juge fait le même jugement (stabilité, **S**), le juge fait deux jugements différents qui restent dans la même catégorie (semi-stabilité, **~S**), le juge fait deux jugements différents qui ne sont pas dans la même catégorie (non stabilité, **¬S**). Nous avons observé que les juges sont bien plus sévères dans leur seconde évaluation des données témoins, à part le juge 5 (J5) qui est stable.

Nous avons calculé les coefficients Gamma et Kappa pour les deux groupes de juges avant et après la tâche. Les coefficients Gamma sont très proches de 1, ce qui indique que les juges sont cohérents sur l'ordre entre les notes. En ce qui concerne les scores Kappa, les deux groupes sont différents. Dans le groupe italien, les juges J4 et J6 ont un excellent accord avant et après l'évaluation. Ils sont devenus plus sévères de la même manière J5 reste consistant. Dans le groupe français, la situation est moins tranchée : le score Kappa diminue aussi pour le groupe, mais les accords entre juges fluctuent.

---

comprendre), GOOD (toutes les informations importantes sont présentes), BAD (une ou plusieurs informations importantes ont été omises), VERY BAD (les informations importantes sont presque toutes absentes). Les deux premières valeurs sont aussi additionnées pour former la catégorie ACCEPTABLE, au côté de BAD et VERY BAD.

[16] Nous disposions de trois juges français pour la paire italien-français, et de trois juges italiens pour la paire français-italien. Nous leur fournîmes une fiche d'instruction expliquant les objectifs de l'évaluation et la notation, ainsi qu'un fichier d'entraînement préalablement évalué par les experts du projet. Chaque juge l'évalua, puis nous révélâmes les notes attendues, et discutâmes avec les juges des divergences observées afin que, durant l'expérience, ils évaluent de façon conforme aux attentes.

3.1.3.2. *Évaluation subjective dans le cadre de la campagne IWSLT-04*

En 2004, les partenaires du consortium C-STAR III montèrent la première campagne d'évaluation compétitive IWSLT, sur la base du corpus BTEC (Akiba *et al.*, 2004). Les couples de langues étaient japonais-anglais (deux conditions expérimentales) et chinois-anglais (trois conditions). Pour l'évaluation, deux ensembles disjoints de phrases chinoises et japonaises furent proposés aux compétiteurs. L'évaluation subjective suivit le protocole du NIST (fluidité, adéquation). Pour l'évaluation objective, on utilisa les mesures[17] BLEU, NIST, GMT, mWER et mPER, pour lesquelles seize paraphrases en anglais étaient disponibles.

En tant que partenaire du consortium non impliqué dans le développement d'un système de traduction japonais-anglais ou japonais-chinois, nous avons participé à cette évaluation pour garantir une évaluation plus juste des systèmes commerciaux (Blanchon *et al.*, 2004c). Nous avons choisi Systran, car les deux paires de langues étaient disponibles et avaient récemment été améliorées, bien que les efforts de Systran aient plus porté sur les couples anglais-chinois et anglais-japonais. Systran avait mis à notre disposition la dernière version Premium Professional (v5).

Pour l'évaluation subjective, chaque jeu de traductions candidates d'un système a été évalué par trois juges[18] de langue maternelle anglaise. Chacun avait uniquement à sa disposition les instructions standard du protocole NIST. Nous avons recalculé les coefficients Gamma et Kappa pour les deux groupes de juges. Les coefficients Gamma sont supérieurs à 0,6 : les juges sont donc d'accord sur l'ordre relatif des notes.

Avant de calculer les coefficients Kappa, nous avons vérifié l'hypothèse des effectifs marginaux équilibrés pour chaque paire de langues. Pour chacune des deux évaluations en fluidité, il apparaît qu'un des juges utilise la note 2 dans 46 % des cas, et pas les deux autres juges. Pour ces deux évaluations, on ne peut donc pas interpréter la valeur du coefficient Kappa global.

Pour les deux évaluations d'adéquation, nous n'avons pas observé de différence significative pour les effectifs marginaux. Les coefficients Kappa deux à deux sont compris entre 0,19 et 0,38. L'accord entre les juges est donc modéré. Les valeurs du coefficient Kappa de chacun des deux groupes d'évaluateurs sont 0,309 pour la paire chinois-anglais, et 0,318 pour la paire japonais-anglais. L'accord entre les trois juges de chaque groupe est donc modéré.

---

[17] Cf. section 3.2 pour des références à ces mesures.

[18] Pour les campagnes suivantes IWSLT (05, 06), les traductions candidates des différents systèmes ont été mélangées avant d'être soumises à l'évaluation subjective et seuls les résultats d'évaluation consolidés par système étaient disponibles. Nous n'avons donc pas pu répéter notre expérience. Pour les campagnes futures, nous allons conserver les résultats d'évaluation de chaque juge pour pouvoir faire la même expérience.

3.1.3.3. *Observations*

Ces deux expériences montrent que l'on peut probablement avoir une confiance plus grande dans les résultats d'une évaluation subjective dans laquelle les juges ont été formés. Il nous semble aussi très intéressant de fournir aux juges un même jeu de données témoins afin de pouvoir mesurer, d'une part, comment ils se comportent les uns vis-à-vis des autres et, d'autre part, comment chacun se comporte au cours du temps. Nous avons vu avec l'expérience NESPOLE! que les juges deviennent de plus en plus sévères quand le temps passe. Il faudrait donc que l'évaluation objective se déroule dans des conditions qui permettent qu'ils jugent avec une sévérité constante.

Lors de l'évaluation IWSLT-04, l'évaluation respectait complètement le protocole NIST : pour une traduction candidate, on évaluait en séquence sa fluidité, puis son adéquation. Lors des évaluations ISWLT-05 et 06, les juges évaluaient d'abord la fluidité de chaque traduction candidate, puis, pour un énoncé source, ils devaient évaluer en même temps toutes les traductions candidates fournies par les différents système de TA, par rapport à la même traduction de référence. Dans la pratique, nous avons observé en 2005 que les juges cherchaient à ordonner les différentes traductions candidates, tâche très difficile, fatigante et consommatrice de temps. Pour l'évaluation de 2006, les évaluateurs fournis par notre groupe ont respecté le protocole NIST standard et ont mis quatre à cinq fois moins de temps !

3.2 BLEU et autres mesures fondées sur des références

La première méthode d'évaluation objective proposée fondée sur des références fut la distance d'édition sur les mots. Cette méthode, empruntée au domaine de la reconnaissance de la parole, est connue sous le nom de « taux d'erreur en mots » (Word Error Rate, WER). Le WER initial est la classique distance d'édition au niveau des mots (Levenshtein, 1966), normalisée par rapport à la longueur. On la calcule par l'algorithme de (Wagner *et al.*, 1974). De nombreuses variations ont été proposées et implémentées depuis (Leusch *et al.*, 2003, Nießen *et al.*, 2000, Och *et al.*, 2002, Tomás *et al.*, 2003).

Il est intéressant de présenter plus en détail les méthodes d'évaluation objective, fondées sur des références et reposant sur la co-occurrence de n-grammes, récemment proposées et très utilisées.

3.2.1. *BLEU*

La méthode BLEU (BiLingual Evaluation Understudy) a été introduite par (Papineni *et al.*, 2002). Elle mesure la similarité entre une traduction candidate et une ou plusieurs traductions de référence. Cette comparaison s'appuie sur une mesure de précision modifiée pondérée par une éventuelle pénalité. Pour chaque n-gramme ($1 \leq n \leq 4$) d'une traduction candidate, on compte son nombre maximum d'occurrences dans chaque traduction de référence (*max_ref_count*) en le minorant

par son nombre d'occurrences dans la traduction candidate (*count*). On divise ensuite le nombre obtenu ($Count_{clip}$ = min (*count*, *max_ref_count*)) par le nombre de n-grammes de la traduction candidate (Équation 1).

Le score p_n (pour un n-gramme) évalue la co-occurrence de tous les n-grammes au niveau de tout le corpus plutôt qu'énoncé par énoncé, pour éviter de prendre en compte la longueur de chaque phrase. Les intuitions proposées sont (1) que p_1 (co-occurrence de mots) est reliée à l'adéquation, et (2) que p_n, pour n>1 (co-occurrence de suites de n mots), est reliée à la fluidité. Ainsi, plus on retrouve de mots présents dans la/les traduction(s) de référence(s), mieux le sens devrait être transmis ; et plus on retrouve de suites de mots, plus le texte devrait être grammatical.

La troisième intuition est que la traduction candidate ne doit être ni trop longue, ni trop courte. Les traductions candidates plus longues que les traductions de référence sont déjà pénalisées par la mesure de précision. Il faut alors pénaliser les traductions trop courtes qui font augmenter la précision. Une pénalité pour brièveté est alors calculée sur tout le corpus, d'où l'Équation 2, où |c| est la longueur totale du corpus de traductions candidates et |r| est la somme des longueurs des traductions de référence qui sont les plus proches de celle de leur traduction candidate.

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count(n-gram)}$$

Équation 1. *Précision* BLEU

$$BP_{BLEU} = \begin{cases} 1 & \text{si } |c| > |r| \\ \exp(1-|c|/|r|) & \text{si } |c| \leq |r| \end{cases}$$

Équation 2. *Pénalité de brièveté* BLEU

Finalement, le score BLEU d'un corpus de traductions candidates est le produit de sa précision et de sa pénalité de brièveté (Équation 3) ; il est compris entre 0 et 1.

$$\text{BLEU} = BP_{BLEU} \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log p_n\right)$$

Équation 3. *Calcul du score* BLEU

3.2.2. *Autres variations*

(Doddington, 2002) soulève deux problèmes liés au calcul de BLEU :

- BLEU fait une moyenne géométrique de la participation des co-occurrences de n-grammes (même poids 1/N pour chaque n). Ainsi, la participation des longs n-grammes (moins nombreux) est surévaluée par rapport à celle des n-grammes courts (plus nombreux). (id.) propose d'y remédier en utilisant une moyenne arithmétique (utilisant le nombre d'appariements sur chaque classe de n-grammes).
- BLEU donne le même poids à tous les n-grammes, alors qu'il serait préférable d'alourdir les n-grammes les plus informatifs.

Finalement, (id.) introduit une mesure de précision pondérée (w_n) sur les n-grammes pour une traduction candidate par rapport à un ensemble de traductions de référence, et une nouvelle pénalité de brièveté (BP_{NIST}) destinée à réduire l'impact des petites variations dans la longueur de la traduction candidate par rapport aux traductions de référence.

Le calcul du score NIST d'un corpus de traductions candidates est donné par l'Équation 4. Ce score n'est pas borné en théorie, mais, en pratique, on a 0≤NIST≤20.

$$\text{NIST} = BP_{NIST} \cdot \sum_{n=1}^{N} w_n$$

Équation 4. *Calcul du score NIST*

Dans le même esprit, (Babych *et al.*, 2004a) font remarquer que les mots d'un texte ont des poids informationnels différents et que par conséquent leur importance pour la traduction varie. Ainsi, pour évaluer l'adéquation d'une traduction candidate, le choix des équivalents traductionnels de ces mots importants devrait avoir plus de poids que le choix des mots utilisés à finalité structurale et sans équivalent traductionnel précis dans le texte source. Ils proposent alors la mesure WNM (Weighted N-Gram Model), qui tient compte de la saillance (« salience ») de chaque mot dans le calcul de la précision. WNM est calculé en utilisant la mesure classique TF.IDF, et le S-score, défini dans (Babych *et al.*, 2004b).

Beaucoup d'autres mesures d'évaluation objective fondée sur des références ont été proposées. Faute de place, nous renvoyons le lecteur intéressé aux sources suivantes : (Banerjee *et al.*, 2005) pour METEOR, (Lin *et al.*, 2004a) pour ROUGE, (Lin *et al.*, 2004b) pour ORANGE, et à (Soricut *et al.*, 2004) pour une synthèse.

Enfin, (Hamon *et al.*, 2006b, Rajman *et al.*, 2001) proposent des mesures objectives (X-Score et D-Score), non fondées sur des références, qui visent à prévoir le classement de différents systèmes qu'on obtiendrait par les mesures subjectives de fluidité, d'adéquation et d'informativité (celles de la DARPA dans les années 90). Ces mesures ont été essayées dans la campagne CESTA (Hamon *et al.*, 2006a). D'après les auteurs, il faudrait des études plus détaillées pour les valider.

3.2.3. *Critique par la pratique et sa justification théorique*

3.2.3.1. *Évaluation objective dans la campagne IWSLT-04*

Dans le cadre de la campagne d'évaluation IWSLT-04, 4 systèmes (anonymisés `JE_1`, `JE_3`, `JE_4`) ont participé à la condition « japonais-anglais, illimitée ». Systran se classe 4e en évaluation subjective et objective (ligne ***`J_3`***, Table 4) face à trois systèmes statistiques.

Nous avons révisé manuellement les traductions produites par le système afin de produire des traductions acceptables en minimisant le nombre de modifications. Sur les cinq cents traductions candidates, seules cinquante n'ont pas été modifiées (10 %). Ce nouveau jeu de traductions, jugées parfaites, a été évalué avec les

mesures objectives, et il se classe 3e derrière deux systèmes de traduction statistique (ligne ***J_4***, Table 4) donnant assez souvent de très mauvaises traductions !

	BLEU		GMT		NIST		PER		WER	
JE_1	0.6306	1	0.7967	2	10.7201	2	0.2333	1	0.2631	1
JE_3	0.6190	2	0.8243	1	11.2541	1	0.2492	2	0.3056	2
J_4	*0.4691*	*3*	*0.7777*	*3*	*9.9189*	*3*	*0.3236*	*3*	*0.3711*	*3*
JE_4	0.3970	4	0.6722	4	7.8893	4	0.4202	4	0.4857	4
J_3	*0.1320*	*5*	*0.5687*	*5*	*5.6476*	*5*	*0.5978*	*5*	*0.7304*	*5*

Table 4. *Évaluation objective des systèmes japonais-anglais pour IWSLT-2004*

3.2.3.2. *Observations*

Cette expérience montre d'une part que des traductions humaines « parfaites » obtiennent des scores qui ne sont pas excellents. Cela nous a permis de rappeler à la communauté (dès 2004) que les méthodes d'évaluation objective n'évaluent pas la *qualité* des traductions[19] soumises à évaluation, mais la *ressemblance* des traductions candidates avec les traductions de référence. Lorsque la ressemblance est parfaite, la traduction candidate est parfaite (sous réserve que la traduction de référence le soit !). Lorsque la ressemblance n'est pas parfaite, il nous semble que l'on ne peut rien dire du tout, si l'on ne voit pas les traductions produites. Cette expérience montre aussi qu'il est injuste de comparer des systèmes qui ne peuvent être ajustés au corpus d'entraînement à des systèmes construits pour eux.

Les données japonaises peuvent être considérées comme des transcriptions « propres » de tours de parole, extraits de leur contexte dans le domaine du tourisme. Le niveau de langue est plutôt poli.

Certains tours de parole sont incompréhensibles sans contexte (切りますよ。⇓ « it cuts » ?). Lorsque le sujet de la première personne est omis en japonais, il est toujours traduit par « it » (ここで降ります。⇓ « It gets off here. » [20]). Dans les traductions, le pronom ou l'adverbe interrogatif est toujours placé en fin de traduction. L'ordre standard des mots en anglais n'est donc pas respecté (オペラ座はどこですか。⇓ « Is the opera house where? » [21]).

[19] C'est vraiment un problème lorsque l'on communique sur les résultats avec d'autres communautés, comme celle de la recherche d'information, qui ont des mesures précises de la qualité d'un système. Bien sûr, en recherche d'information, les tâches ne sont pas de même qu'en TA et il n'y a pas plusieurs réponses possibles à une question pour un jeu de test donné, mais une seule. En TA, « **LA » bonne traduction** n'existe pas, on peut seulement dire si une traduction est bonne ou pas.

[20] « Je descends ici. »

[21] « Où est l'opéra ? »

Aucune des expressions orales de la vie quotidienne n'est dans les dictionnaires Systran (どういたしまして。⇓ « How doing. » [22]). Les requêtes et les invitations ne sont pas toujours bien traduites (一緒に行きましょう。⇓ « It will go together. » [23]).

Quand la valence du verbe pour deux expressions en japonais et en anglais est différente, la traduction est presque toujours mauvaise (寒気がする。⇓ « Chill does. » [24]). Enfin, l'aspect des prédicats japonais n'est pas rendu correctement en anglais (航空券を家に忘れて しまいました。⇓ « The air ticket was forgotten in the house. » [25]).

3.2.3.3. *Vers une réfutation théorique de la valeur des scores BLEU*

Le premier problème est que les nombres produits par les techniques d'évaluation objective ne sont pas directement liés à la qualité de la traduction. Beaucoup de travaux ont été conduits en vue de corréler des résultats d'évaluations objectives avec des résultats d'évaluations subjectives, mais les résultats sont souvent inconsistants. BLEU est supposé bien corrélé avec des évaluations humaines de qualité de traduction, NIST a été proposé pour donner des résultats meilleurs pour des raisons théoriques, mais BLEU et NIST donnent des résultats contradictoires. Ainsi, si la corrélation avec des jugements humains est une mesure de qualité de ces techniques, alors NIST ne peut être meilleur que BLEU… or la corrélation est trop faible pour vouloir dire quelque chose.

(Culy *et al.*, 2003) développaient déjà le même type d'argument. Plus récemment, (Callison-Burch *et al.*, 2006) ont montré que (1) BLEU n'est pas suffisant pour mettre en évidence une vraie amélioration dans la qualité de la traduction, et (2) qu'il n'est pas nécessaire d'améliorer le score BLEU pour obtenir un meilleur jugement de la qualité de la traduction par des évaluateurs humains. Ces phénomènes sont liés au fait que BLEU ne contraint pas l'ordre dans lequel les n-grammes s'apparient, ni l'appariement des n-grammes sur des références multiples. Ces remarques s'appliquent aussi aux travaux et propositions de Lin, Soricut et Banerjee. À la lumière de ces conclusions, il faut donc peut-être réexaminer des travaux dans lesquels des propositions d'amélioration de la qualité des traductions n'ont pas permis d'améliorer le score BLEU.

3.3 Human Translation Error Rate

Dans le cadre du projet GALE[26] (Global Autonomous Language Exploitation), visant la transcription de messages oraux et leur traduction, le NIST, sous

---

[22] « Pas de quoi ! » en réponse à un « Merci ! »
[23] « Allons-y ensemble ! »
[24] « J'ai froid. »
[25] « J'ai oublié mon billet d'avion à la maison. »
[26] `http ://projects.ldc.upenn.edu/gale/`

l'impulsion de Joseph Olive, met en œuvre la mesure HTER (Human-targeted Translation Error Rate) (Przybocki *et al.*, 2006). Comme la classique distance d'édition TER (Translation Error Rate) (Snover *et al.*, 2006), c'est le coût minimum d'une suite d'opérations d'édition permettant de transformer une traduction candidate en une traduction de qualité suffisante pour une certaine tâche. L'originalité est que les opérations d'édition usuelles, à savoir l'insertion, la suppression, et la permutation, s'appliquent non seulement à des mots, mais à des séquences de mots.

3.3.1. *Principe du protocole*

GALE met en œuvre cette mesure dans le protocole suivant. Pour évaluer un système, on part d'un ensemble d'énoncés en langue source, de traductions candidates à évaluer, et d'une ou plusieurs traductions de référence en langue cible, produites par des traducteurs. Des *réviseurs monolingues* éditent les traductions candidates pour les rendre fluides et pour que leur contenu soit conforme à une traduction de référence.

Au lieu de simplement faire postéditer une traduction candidate une fois par un *traducteur humain*, et de mesurer le TER, comme pour MÉTÉO, le NIST a imaginé un protocole beaucoup plus coûteux, sans doute moins adéquat pour mesurer l'utilité, mais minimisant le recours à des traducteurs et/ou évaluateurs bilingues. Le processus d'évaluation se déroule en trois étapes.

3.3.2. *Mise en œuvre*[27]

Lors de la première étape, *trois réviseurs monolingues* postéditent, pour chacun des énoncés, la même traduction candidate (`trads`), en s'appuyant sur une traduction de référence (`refs`), pour la rendre fluide et adéquate. Ils produisent trois *traductions candidates postéditées* (`post-éds`), à partir desquelles on calcule un score TER.

Lors de la deuxième étape, *deux nouveaux réviseurs monolingues* postéditent deux ensembles de *traductions candidates postéditées,* toujours en s'aidant d'une traduction de référence. Le premier ensemble est constitué des traductions postéditées ayant le meilleur score TER (`min étape 1`). Le deuxième est constitué des traductions postéditées ayant le score TER intermédiaire (`med étape 1`). On obtient ainsi deux jeux de *traductions candidates post-postéditées* (`post²-éds`). Un second score TER est alors calculé en utilisant les traductions candidates et les *traductions candidates post-postéditées*.

La troisième étape permet d'obtenir des *traductions postéditées finales* en conservant, pour chaque énoncé, la *traduction candidate post-postéditée* ayant eu le meilleur score TER lors de la seconde étape (`min étape 2`). On peut alors calculer le score HTER final du système.

---

[27] Cf. ***1***

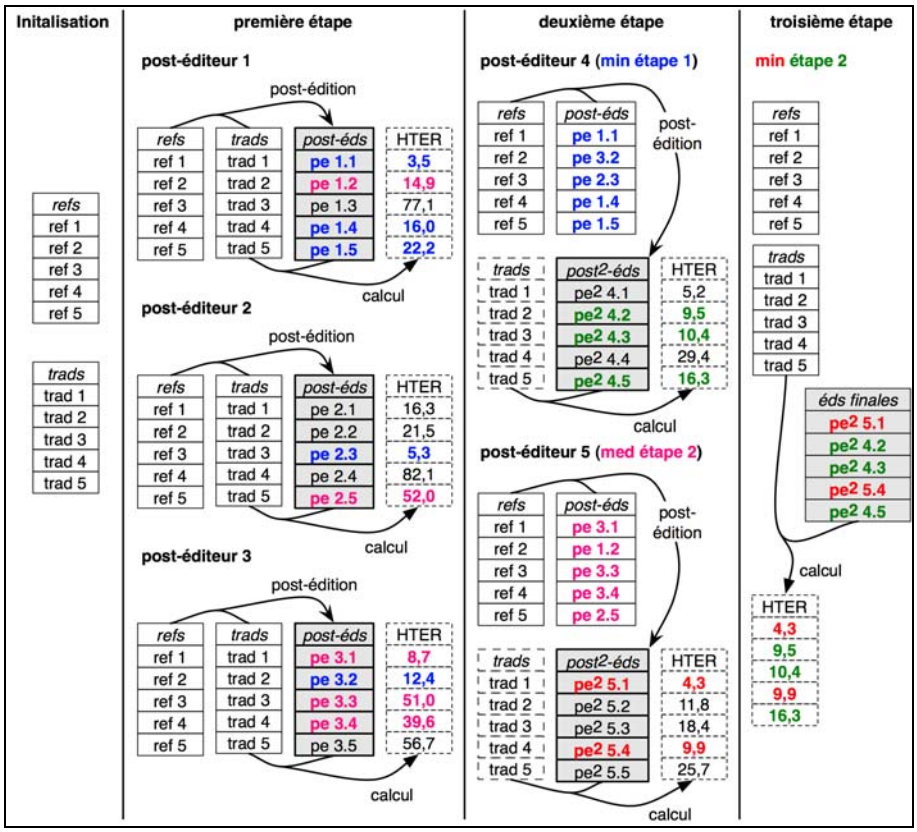

Figure 1. *Mise en œuvre du protocole HTER*[28]

3.3.3. *Commentaires*

La postédition des traductions candidates n'est pas une idée nouvelle, nous la reproposions déjà en 2004 (Blanchon *et al.*, 2004b). Cependant, dans la mise en œuvre du NIST, un ou plusieurs traducteurs bilingues doivent d'abord produire des traductions de référence. Au lieu de faire cela, ils pourraient directement postéditer les traductions candidates, évitant ainsi de faire appel à des réviseurs.

Notre proposition serait donc de faire une mesure bien moins onéreuse, et plus proche d'une vraie mesure d'usage, en demandant à des postéditeurs de produire, au moindre effort, des traductions correctes à partir des traductions candidates. Le travail fourni, estimé en temps ou en opérations d'édition observées, donnerait une meilleure appréciation de la qualité d'usage (utilisabilité ici) du système de TA

[28] D'après S. Roukos (IBM T.J. Watson Research Center), présentation invitée à IWSLT-06.

évalué. Avec cette approche, des traductions humaines de qualité ne nécessiteraient aucun travail de postédition et seraient donc jugées parfaites.

Il faut aussi noter que les différentes étapes du protocole HTER de GALE risquent de poser des problèmes pour l'évaluation de la traduction d'un texte homogène cohérent. En effet, les postéditeurs du premier groupe peuvent faire des choix différents lors de l'édition, par exemple pour le traitement de la référence. Les postéditeurs du second groupe risquent d'avoir à traiter un document incohérent.

Pourquoi un protocole aussi compliqué ? Il semble que ce soit pour essayer d'utiliser BLEU afin de prévoir le plus tôt possible les scores HTER, plus longs à obtenir. En effet, les financements sont de type « Go No Go », et les participants veulent savoir s'ils ont des chances de voir leurs financements se poursuivre.

4. Propositions

Nous nous intéressons ici essentiellement à l'*évaluation externe* des systèmes de TA. Nous avons dit que, dans ce type d'évaluation, on n'analyse pas l'architecture linguistique et calculatoire du système. Cependant, on peut et on doit s'intéresser à la couverture lexicale des systèmes construits de façon « experte ». La taille des dictionnaires est en effet une information toujours donnée par les vendeurs de systèmes de TA. Pour les systèmes construits de façon « empirique », il faut sans doute inclure la taille du corpus parallèle utilisé.

L'*évaluation interne* est bien sûr très intéressante, pour les développeurs ainsi que pour des décideurs intéressés par le potentiel des systèmes, et les méthodes proposées (JEIDA, FEMTI, etc.) sont très bonnes. Mais elles ne fournissent pas de comparaisons très précises et objectives entre les systèmes, et sont en général très coûteuses, donc ne sont pas utilisées dans les campagnes actuelles.

Nous avons analysé ci-dessus les défauts inhérents aux méthodes d'*évaluation externe*, et un certain nombre de biais introduits dans les conditions expérimentales dans le but de diminuer les coûts. Que peut-on donc proposer pour améliorer la pertinence des *évaluations externes* de systèmes de TA et diminuer leur coût ?

Nous proposons d'abord un classement (une « typologie ») simple des mesures externes, avec comme critère principal la présence ou l'absence de liaison avec une tâche réaliste. Dans ce cadre, nous proposons ensuite les mesures liées à la tâche qui nous semblent les plus pertinentes, en indiquant comment diminuer leur coût, ou annuler leur coût en les intégrant à l'utilisation normale des systèmes.

4.1 Un classement simple des mesures externes

4.1.1. *Mesures externes liées à la tâche*

La *diffusion* et la *compréhension* sont les deux tâches principales en TA de l'écrit, alors que pour l'oral la diffusion est remplacée par la *communication*. Le

même système sera en général évalué différemment selon qu'on se place du côté du producteur de l'information, qui projette son « image » et doit le faire avec un haut niveau de qualité, ou du côté du consommateur, qui cherche à comprendre une information ou à communiquer en langue étrangère.

4.1.1.1. *Diffusion (de l'écrit) et communication (à l'oral)*

En *diffusion de l'écrit*, les tâches typiques sont le thème et la version de haute qualité. Par exemple, on veut traduire une documentation technique, ou un appel d'offres européen, en trente ou quarante langues. Ou encore, on veut traduire dans sa langue des brevets ou des contrats rédigés en langue étrangère. La fidélité de la traduction est alors cruciale, le choix des termes est très important, la correction grammaticale est de rigueur, et le style doit être adéquat. Il s'agit donc de produire des traductions de qualité professionnelle.

Un système de TA est alors utilisé par des traducteurs, qui postéditent ses résultats, et éventuellement par des rédacteurs (cas d'un système de TA avec désambiguïsation interactive comme JETS (Maruyama *et al.*, 1990), Taifun ou Tsunami). Les mesures associées à cette tâche sont donc la diminution du coût (en travail humain) et des délais.

En *communication orale*, la tâche typique est d'aider deux personnes à conduire un dialogue bilingue pour accomplir une tâche.

4.1.1.2. *Compréhension*

Les tâches de *compréhension de l'écrit et de l'oral* sont différentes, mais on peut dire qu'il est toujours plus difficile de construire un système très utile pour la compréhension que pour la diffusion ou la communication. C'est le contraire de ce qu'on pensait au début de la TA, mais c'est vrai. En effet, pour la diffusion (écrit), les défauts du système sont largement compensés, en usage, par la compétence bilingue des postéditeurs, ou par la connaissance du domaine. De même, pour la communication (cas de dialogues bilingues), ces défauts sont compensés par la volonté des interlocuteurs de se comprendre.

En *compréhension de l'écrit*, les tâches typiques sont la traduction de pages Web, de journaux, de services de e-commerce, pour que les utilisateurs finals puissent comprendre de l'information en langue étrangère et agir en conséquence. Les mesures adaptées sont soit objectives[29] (nombre d'actes d'achat par page visitée en e-commerce, temps passé par page en lecture de journaux…), soit subjectives (retours des utilisateurs, réponses à des enquêtes de satisfaction…).

En *compréhension orale*, la tâche typique est de suivre un monologue (discours au Parlement, etc.) ou un dialogue en langue étrangère (télévision, renseignements), ou, par exemple, de suivre des nouvelles en arabe à la télévision (démonstration

[29] On peut les mesurer grâce à des cookies.

récente d'IBM). On ne peut pas « corriger un premier jet », il faut traduire à la volée, avec un décalage très faible, même en interprétariat de liaison.

Par conséquent, toute mesure liée à la tâche doit comparer la performance d'un système à celle d'un interprète humain. Il est peut-être possible de faire une telle évaluation sur la transcription écrite du discours source et de sa traduction, mais il est probable que cela introduirait des biais.

Il faut donc juger le taux de compréhension. Pour cela, il existe des mesures objectives (temps passé pour accomplir la tâche, QCM[30] sur le contenu…), comme celle d'ATR (Sugaya *et al.*, 2001), et des mesures subjectives (sentiment de compréhension, jugement de fluidité).

4.1.2. *Mesures externes non liées à la tâche*

4.1.2.1. *Mesures externes liées à des références*

Les *mesures externes liées à des références* ne sont, par nature, liées à aucune tâche réaliste, tout au moins si on utilise les références plusieurs fois, ce qui est leur principal intérêt d'après leurs partisans. En effet, traduire et retraduire un corpus de test muni de références ne correspond à aucun besoin et à aucun marché : une seule bonne traduction suffit. Si un segment déjà traduit doit ultérieurement être traduit, il vaut évidemment mieux le traduire par sa traduction (éventuellement postéditée) stockée dans une mémoire de traduction que le retraduire par TA.

Les *mesures objectives liées à des références* sont les plus utilisées actuellement (NIST, BLEU, METEOR, WER…). La plupart font des calculs sur les n-grammes de mots ($n \leq 4$). Certaines, comme WNM, travaillent à des niveaux linguistiques plus riches.

Les *mesures subjectives liées à des références* sont l'adéquation « à la NIST », utilisée dans les campagnes actuelles, ou l'informativité « à la ALPAC », ou la fidélité « à la JEIDA ». En revanche, la fluidité n'est pas à ranger ici, car on n'a pas besoin de références pour la mesurer.

4.1.2.2. *Mesures externes non liées à des références*

Les *mesures externes objectives non liées à des références* sont d'abord des mesures de coût. Le *coût en temps et en espace* n'est pas du tout le même selon les systèmes, mais, même pour les systèmes statistiques, on arrive maintenant à des temps raisonnables sur des serveurs puissants. On cherche plutôt à voir si les systèmes testés peuvent être utilisés sur PDA, et bientôt sur des mobiles.

Il s'agit ensuite de l'évaluation du *coût de préparation d'un système*. Dans le cas des systèmes statistiques, le coût dominant est celui du corpus parallèle utilisé pour l'apprentissage. Selon K. Knight (session spéciale à CICLING-05), il doit contenir

[30] QCM : question à choix multiple.

au moins cinquante millions de mots dans chaque langue, soit 200 000 pages standard, ou deux millions et demi de phrases de vingt mots. Il a donc coûté entre 200 000 et 250 000 heures de travail humain. En 2006, Ph. Koehn parlait de deux cent millions de mots, soit quatre fois plus. Le tout est de savoir si ce coût doit être imputé ou non, ou en partie seulement, à la construction du système de TA.

Dans le cas des systèmes à dictionnaires et grammaires, le coût dominant est celui des dictionnaires. On peut donc aussi utiliser une *mesure de taille* pour les corpus, dans un cas, pour les dictionnaires, dans l'autre.

On peut ranger ici la mesure HTER non finalisée du projet GALE. Bien qu'elle soit présentée comme liée à la tâche, elle ne l'est pas car la tâche envisagée est tout à fait irréaliste : qui voudrait traduire une documentation en la faisant traduire par trois traducteurs, puis en la faisant postéditer par deux réviseurs travaillant sur deux mélanges différents des traductions obtenues, puis par un dernier réviseur ?

Les *mesures externes subjectives non liées à des références* sont nombreuses. L'évaluation de fluidité en fait toujours partie. L'adéquation peut être mesurée par des tests de compréhension (QCM comme dans le TOEFL), qui ne supposent pas de traduction préalable.

4.2 Proposition : n'utiliser que des mesures liées à la tâche et peu coûteuses pour l'évaluation externe

Nous présentons maintenant quelques propositions, en séparant la TA de l'écrit et la TA de l'oral, puisqu'on a vu plus haut que leurs *évaluations externes* étaient par nature assez différentes.

4.2.1. *Mesures externes objectives pour la TA de l'écrit*

Nous proposons d'abandonner totalement les mesures liées à des traductions de référence, pour les raisons présentées plus haut, et d'utiliser des mesures différentes s'il s'agit de diffusion ou de compréhension.

4.2.1.1. *Tâche de diffusion*

On devrait utiliser uniquement des mesures objectives d'utilisabilité, et aucune mesure subjective.

Mesurer l'utilité serait en un sens idéal, mais on ne peut la mesurer que sur un système opérationnel, dans un contexte réel, alors que l'on veut aussi ou surtout évaluer des systèmes en développement.

On peut mesurer *l'utilisabilité* de systèmes, en développement ou opérationnels, en les intégrant dans un service Web, le même pour tous les systèmes, muni d'un éditeur bilingue. Notre équipe a construit à cet effet la plate-forme TRANSBey (Bey *et al.*, 2006), accessible par le Web, qui permet l'affichage, la traduction (en utilisant

plusieurs outils de traduction), et l'édition (par postédition ou traduction). Diverses mesures sont prévues, en particulier des distances d'édition (entre la « prétraduction » postéditée et la traduction finale, et une mesure du temps passé à chaque sous-tâche (postédition proprement dite, recherche terminologique, ajout au dictionnaire commun, correction de la segmentation ou même de certaines phrases de l'original).

Le premier composant de l'utilisabilité est alors la variation de productivité, que nous définissons ici comme l'efficacité relative des deux contextes de travail, avec et sans système de traduction (ST). Le temps humain est bien connu par tous les donneurs d'ordres en traduction. Pour des textes difficiles, il est de l'ordre de 80 mn/page (60 mn de traduction, 20 mn de révision). Pour des textes formés de phrases très courtes comme celles du BTEC, ou pour de la traduction courante (cas des journaux traduits par Comprendium en Espagne) c'est plutôt 60 mn/page.

$$\text{Efficacité Relative}_{ST} = \frac{\text{Temps}_{Humain}}{\text{Temps}_{ST}}$$

Équation 5. *Efficacité relative d'un système de traduction automatique*

On dira par exemple qu'un système est utilisable dès que son efficacité relative dépasse 2, car cela correspond au gain maximal que l'on peut espérer d'une mémoire de traduction. Pour un certain nombre de paires de langues des systèmes commerciaux actuels, on obtient mieux[31], de 2,5 à 3, et on peut obtenir bien plus, par exemple 12 dans le cas de Comprendium et de la traduction de journaux d'espagnol et catalan et en galicien (5 mn/page de temps humain avec la TA contre 60 mn sans rien et 30 mn avec un outil de mémoire de traduction).

En ce qui concerne les mesures subjectives, comme celles de fluidité et d'adéquation, elles sont ici totalement inutiles, et même contre-productives. Il est en effet parfaitement possible que toutes les phrases d'un certain type soient grammaticalement fausses, ou que certains choix lexicaux soient mauvais, et que la postédition soit très rapide. En particulier, si l'éditeur de traductions est assez puissant, on peut faire des changements globaux, et l'éditeur peut aussi détecter des chaînes déjà transformées et proposer les mêmes transformations.

4.2.1.2. *Tâche de compréhension*

On peut ici utiliser des mesures objectives et subjectives.

S'il s'agit de lire une page Web en langue étrangère, on peut mesurer (*via* des cookies) le temps moyen passé par les internautes sur une page traduite dans leur langue, et le comparer à celui passé sur la page originale par des locuteurs de la langue initiale. Notre hypothèse est que, (1) si ce temps est beaucoup plus court,

[31] Voir le site de J. Allen, `http ://www.geocities.com/mtpostediting/`

c'est que la traduction est peu compréhensible et que les lecteurs ont abandonné ; (2) s'il est bien plus long, c'est qu'on arrive à comprendre, avec des efforts supplémentaires, mais qu'on ne comprend pas tout ; et (3) que, s'il est du même ordre de grandeur, la TA est tout à fait satisfaisante pour cet usage. Cette hypothèse est vraisemblable, mais reste à prouver, par expérimentation.

S'il s'agit d'un service de e-commerce « multilingualisé », on peut compter les « actes commerciaux » (achats, commandes) et le temps passé (par exemple, pour obtenir une réponse satisfaisante d'un service après-vente), et faire le même genre de comparaison.

S'il s'agit de mesurer plus précisément le degré de compréhension, c'est-à-dire, dans ce contexte, l'adéquation, et cela sans références, de façon liée à la tâche, et sans jugements subjectifs humains, la seule solution semble être de s'inspirer des tests de compréhension intégrés aux examens de langues comme le TOEFL, et de demander à des utilisateurs potentiels de répondre à des *QCM de compréhension*. Le calcul de la mesure est automatique, le temps passé à répondre est faible, l'accord entre les utilisateurs est meilleur, dont on peut en faire participer moins, et, d'après R. Mitkov (Mitkov *et al.*, 2006), on arrive maintenant à produire cinq ou six QCM pour une page de façon semi-automatique en 30 mn environ.

Les *mesures subjectives* (il s'agit toujours de mesures liées à la tâche) nous semblent devoir rester assez peu précises, et assez coûteuses. On peut bien sûr demander aux utilisateurs un « retour » par des jugements de qualité, comme on le trouve sur certains sites collectant des évaluations de livres ou de logiciels. Mais que pourrait-on en faire pour améliorer les systèmes ?

Mesurer la fluidité ne serait pas pertinent, et de toutes façons le temps passé, mesurable objectivement, est sans doute très corrélé au manque de fluidité. De façon générale, il vaut mieux éviter les mesures subjectives en compréhension de l'écrit, car les jugements humains varient trop. Ainsi, pendant ACL-03[32], certains exposés présentèrent des exemples de traductions avec leurs scores, et une part importante de l'assistance ne fut pas d'accord avec les jugements de qualité : certains étaient faux, et, dans d'autres cas, le sentiment fut que les traductions, bien que fausses, auraient été parfaitement suffisantes pour que les utilisateurs puissent en comprendre le sens et éventuellement agir en conséquence.

4.2.2. *Mesures externes pour la TA de l'oral*

4.2.2.1. *Tâche de diffusion*

Nous avons dit que toute mesure liée à la tâche devrait comparer la performance d'un système à celle d'un interprète humain.

---

[32] Sessions « Statistical Machine Translation » et « Machine Translation and Chunking », `http ://www.informatik.uni-trier.de/~ley/db/conf/acl/acl2003.html`

Malheureusement, on ne voit pas comment intégrer de mesure objective d'usage dans un système de TA de l'oral visant la diffusion. Par exemple, s'il s'agit de permettre de suivre un discours ou un dialogue en langue étrangère, il n'y a pas d'action de l'utilisateur (pas d'acte d'achat, par exemple), et il est difficile de savoir la proportion d'utilisateurs abandonnant cette activité pour cause d'incompréhension.

La seule évaluation objective possible est à notre avis une mesure par QCM de compréhension, qu'on ne peut pas faire au fur et à mesure de l'usage du système, à cause de la contrainte « temps réel » de la parole. Mais on pourrait la faire de façon différée.

Enfin, évaluer la qualité de diffusion à partir de la transcription écrite du discours source et de sa traduction est techniquement possible, mais coûteux et en fait assez biaisé, donc il faut éviter de le faire. Cela consisterait à mesurer l'effort de « postédition » de la traduction pour arriver à une traduction orale de même qualité que celle produite par un interprète professionnel.

Éditer directement le signal en langue cible ne paraît pas techniquement possible. On devrait donc demander à un interprète professionnel, seul qualifié pour dire ce qu'est une interprétation de qualité, de postéditer la transcription fournie par le système de TA, tout en s'assurant que le résultat oral (produit par synthèse de la parole) est effectivement de la qualité attendue, et « tient » dans le temps disponible. Le biais mentionné vient du fait qu'on s'éloigne ainsi de l'idée de base, qui est de mesurer par rapport à une tâche réaliste.

Enfin, effectuer une mesure subjective de la qualité d'une traduction orale à partir de transcriptions écrites est à exclure. En effet, les transcriptions des traductions d'interprètes professionnels de haut niveau sont souvent jugées comme de mauvaises ou très mauvaises traductions, alors qu'elles sont payées beaucoup plus cher, justement parce que leur qualité « relativement à la tâche » de compréhension en temps réel est très élevée.

4.2.2.2. *Tâche de compréhension*

On peut ici utiliser des mesures objectives et subjectives.

En 1999, ATR a montré la voie en effectuant une comparaison (objective) entre le temps mis à réaliser une tâche de réservation hôtelière avec un interprète humain et avec son système de TA de parole. La conclusion fut que le système testé (ATR-Matrix) avait la même efficacité qu'un Japonais ayant un niveau moyen au TOEFL (Sugaya *et al.*, 2001).

On pourrait généraliser ce type d'évaluation comme suggéré plus haut pour l'écrit, c'est-à-dire en intégrant des systèmes de TA de la parole à tester à un ou plusieurs services Web. C'est d'ailleurs ce qui était prévu au début du projet C-STAR III, avant qu'il ne change de cap pour travailler sur la construction de corpus multilingues et l'évaluation compétitive à la NIST.

La mesure objective par QCM de compréhension est bien sûr possible dans ce cas, mais semble bien plus coûteuse que la précédente, et également moins bien adaptée au cas de systèmes de TA de dialogues oraux finalisés.

4.2.3. *N'utiliser les mesures fondées sur des références qu'en évaluation interne*

Les chercheurs en TA statistique disent que les méthodes d'évaluation liées aux références ont de grandes qualités, en particulier leur caractère objectif, mathématique, et sont très utiles, car elles permettent de mesurer les progrès des systèmes à coût humain nul. Cela est vrai, mais elles présentent des défauts bien plus importants en ce qui concerne l'*évaluation externe*.

D'abord, nous l'avons vu, elles ne mesurent pas bien du tout la « qualité ». Si elles sont assez bien corrélées avec la fluidité, ce n'est pas vrai pour l'adéquation, et encore moins pour l'utilisabilité.

Ensuite, si leur usage est peu coûteux, leur préparation est très coûteuse, puisqu'on doit produire des traductions de référence (plusieurs pour NIST, parfois jusqu'à seize comme dans IWSLT-06).

Enfin, les méthodes utilisant des références ne sont pas utilisables pour évaluer des systèmes en exploitation. Par exemple, Google a mis en service dans ses outils linguistiques quelques systèmes de TA statistique, à côté de systèmes Systran. Pour mesurer la qualité, Google propose aux lecteurs, grâce à une interface simple, de corriger les traductions et de leur renvoyer les traductions de référence. Il est très peu probable d'obtenir ainsi plusieurs traductions de référence pour toutes les phrases d'un très gros corpus. Par contre, cela peut être très utile pour le développement.

Nous proposons donc d'abandonner les mesures liées aux traductions de référence en tant que mesures externes, et, par conséquent, de ne plus les utiliser dans les campagnes d'évaluation compétitives dans lesquelles les seules mesures possibles sont des mesures externes.

En revanche, ces mesures pourraient garder toute leur place pour l'*évaluation interne* et le développement des systèmes de TA statistique, construits pour les optimiser. L'idée de base est en fait l'analogue en TA statistique de la technique des « suites de test » utilisées depuis les débuts de la TA par les développeurs de TA « experte ».

5. Conclusion

Les méthodes externes d'évaluation de systèmes de TA définissent des mesures de qualité à partir de leur fonctionnement observable, et éventuellement de leur couverture potentielle, mesurée par la taille des dictionnaires en TA « experte » et par celle des corpus parallèles en TA « empirique ». Alors que les systèmes opérationnels sont depuis longtemps le plus souvent évalués par des méthodes externes fondées sur la tâche (diffusion ou compréhension à l'écrit, communication

ou compréhension à l'oral), les campagnes d'évaluation des dernières années utilisent (parcimonieusement) des méthodes subjectives assez chères fondées sur des jugements humains peu fiables et (pour la plus grande part) des méthodes basées sur des traductions de référence. Ces méthodes subjectives sont impossibles à utiliser lors de l'utilisation réelle d'un système, d'autant moins corrélées aux jugements humains que la qualité augmente, et totalement irréalistes en ce qu'elles forcent à mesurer les progrès sur des corpus fixes, sans cesse retraduits, et non sur de nouveaux textes à traduire pour des besoins réels. Il y a aussi de nombreux biais introduits par le désir de diminuer les coûts, en particulier l'utilisation de corpus parallèles dans le sens inverse de leur production et l'utilisation de personnes monolingues au lieu de bilingues. Nous avons prouvé cela par une analyse de l'histoire de l'évaluation en TA, des méthodes d'évaluation du « courant dominant », et de certaines campagnes d'évaluation récentes.

Cela nous a conduits à proposer une taxonomie révisée des méthodes d'évaluation, en la centrant sur les principaux types de tâches, qui sont différents à l'écrit et à l'oral. Cette clarification nous a menés a proposer d'abandonner totalement les méthodes d'*évaluation externes* fondées sur des traductions de référence, et de les remplacer par des méthodes strictement fondées sur la tâche, en particulier l'efficacité relative pour la diffusion de l'écrit, la communication orale, et la compréhension de l'écrit et de l'oral. Les mesures fondées sur des références, qu'elles soient objectives comme BLEU, NIST, etc., ou subjectives comme les mesures de fluidité ou d'adéquation actuelles, devraient être réservées à l'*évaluation interne* et au développement des systèmes de TA « empirique » de l'écrit et de l'oral.

6. Bibliographie

Akiba Y., Federico M., Kando N., Nakaiwa H., Paul M., Tsujii J.-I. « Overview of the IWSLT04 Evaluation Campaign », *Proc. IWSLT 2004*, Kyoto, Japan, September 30-October 1, 2004, vol. 1/1, p. 1-12.

ALPAC Language and Machine : Computers in Translation and Linguistics, n° 1416. November 1966. Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Science - National Research Council. 124 p.

Babych B., Hartley A. « Extending BLEU MT Evaluation Method with Frequency Weightings », *Proc. ACL 2004*, Barcelona, Spain, July 21-26, 2004, vol. 1/1, p. 622-629.

Babych B., Hartley A. « Modelling legitimate translation variation for automatic evaluation of MT », *Proc. LREC-2004*, Lisbon, Portugal, May 26-28, 2004, p. 833-836.

Banerjee S., Lavie A. « METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgement », *Proc. ACL-05, Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, USA, June 29, 2005, vol. 1/1, p. 25-32.

Bey Y., Boitet C., Kageura K. « The TRANSBey Prototype : An Online Collaborative Wiki-Based CAT Environment for Volunteer Translators », *Proc. 3rd International Workshop*

on Language Resources for Translation Work, Research & Trainning (LR4Trans-III), Genoa, Italy, May 28, 2006, vol. 1/1, p. 49-54.

Blanchon H. « HLT Modules Scalability within the NESPOLE! Project », *Proc. ICSLP 2004*, Jeju Island, Korea, October 4-8, 2004, 4 p.

Blanchon H., Besacier L. « Traduction de dialogues : mise en perspective des résultats du projet NESPOLE! et pistes pour le domaine », *Proc. TALN 2004*, Fès, Maroc, 19-21 avril 2004, vol. 1/1, p. 55-60.

Blanchon H., Boitet C., Besacier L. « Spoken Dialogue Translation System Evaluation : Results, New Trends, Problems and Proposals », *Proc. IWSLT 2004*, Kyoto, Japan, September 30 - October 1, 2004, vol. 1/1, p. 95-102.

Blanchon H., Boitet C., Brunet-Manquat F., Tomokio M., Hamon A., Hung V. T., Bey Y. « Towards Fairer Evaluation of Commercial MT Systems on Basic Travel Expressions Corpora », *Proc. IWSLT 2004*, Kyoto, Japan, September 30-October 1, 2004, vol. 1/1, p. 21-26.

Callison-Burch C., Osborne M., Kœhn P. « Re-evaluating the Role of BLEU in Machine Translation Research », *Proc. ACL-2006*, Trento, Italy, April 3-7, 2006, vol. 1/1, p. 249-256.

Chandioux J. « 10 ans de METEO (MD) », *Proc. Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires organisé par l'Observatoire Francophone des Industries de la Langue (OFIL)*, Paris, France, mars 1988, vol. 1/1, p. 169-173.

Church K. W., Hovy E. « Good Applications for Crummy Machine Translation », *Machine Translation*. vol. 8, n° 4, 1993, p. 239-258.

Culy C., Riehemann S. Z. « The Limits of N-Gram Translation Evaluation Metrics », *Proc. MT Summit IX*, New Orleans, USA, September 23-27, 2003, vol. 1/1, p. 71-78.

Doddington G. « Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics », *Proc. HLT 2002*, San Diego, California, March 24-27, 2002, vol. 1/1, p. 128-132 (note book proceedings).

EAGLES-EWG EAGLES Evaluation of Natural Language Processing Systems, Final Report EAG-EWG-PR.2, Project LRE-61-100. October, 1996. Center for Sprogteknologi. 287 p.

EAGLES-EWG EAGLES Evaluation of Natural Language Processing Systems, Final Report EAG-II-EWG-PR.2, Project LRE-61-100. April, 1999. Center for Sprogteknologi. 173 p.

Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M., Timimi I. « CESTA : First Conclusions of the Technolangue MT Evaluation Campaign », *Proc. LREC*, Genoa, Italy, 24-26 May 2006, p. 179-184.

Hamon O., Rajman M. « X-score : automatic evaluation of machine translation grammaricality », *Proc. LREC 2006*, Genoa, Italy, May 24-26, 2006, p. 155-160.ss

Hovy E., King M., Popescu-Belis A. « Principles of Context-Based Machine Translation Evaluation », *Machine Translation*. vol. 17, n° 1, 2002, p. 43-75.

Hutchins J. (2003) *ALPAC : the (in)famous report.* in Nirenburg S., Somers H. and Wilks Y. (ed.), Readings in machine translation. The MIT Press. Cambrige, Mass. pp. 131-135.

ISO/IEC Software engineering – Product Quality – Part 1. June, 2001. International Organisation for Standardization & International Electotechnical Commission. 25 p.

JEIDA A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A. July, 1989. Japan Electronic Industry Development Association. 197 p.

JEIDA JEIDA Methodology and Criteria on Machine Translation Evaluation. November, 1992. Japan Electronic Industry Development Association. 129 p.

King M. « Evaluating Natural Language Processing Systems », *Communication of the ACM*. vol. 29, n° 1, 1996, p. 73-79.

King M. « Living up to standards », *Proc. EACL 2003 – Workshop on Evaluation Initiatives in Natural Language Processing : are evaluation methods, metrics and ressources reusable?* Budapest, Hungary, April 14, 2003, vol. 1/1, 8 p.

Lavie A., Pianesi F., Levin L. « The NESPOLE! System for Multilingual Speech Communication over the Internet », *IEEE Transaction on Speech and Audio Processing*. vol. In Press, Corrected Proof, 2006, 10 p.

Leusch G., Ueffing N., Ney H. « A Novel String-toString Distance Measure with Application to Machine Translation Evaluation », *Proc. MT Summit IX*, New Orleans, USA, September 23-27, 2003, vol. 1/1, p. 240-247.

Levenshtein V. I. « Binary codes capable of correcting deletion, insertions and reversals », *Soviet Physics Doklady*. vol. 10, n° 8, 1966, p. 707-710.

Lin C.-Y., Och F. J. « Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics », *Proc. ACL 2004*, Barcelona, Spain, July 21-26, 2004, vol. 1/1, p. 605-612.

Lin C.-Y., Och F. J. « ORANGE : a Method for Evaluating Automatic Evaluation Metrics for Machine Translation », *Proc. COLING 2004*, Geneva, Switzerland, August 23-27, 2004, vol. 1/2, p. 501-507.

Maruyama H., Watanabe H., Ogino S. « An Interactive Japanese Parser for Machine Translation », *Proc. COLING-90*, Helsinki, August 20-25, 1990, vol. 2/3, p. 257-262.

Mitkov R., Ha L. A., Karamanis N. « A computer-aided environment for generating muliple-choice test items », *Natural Language Engineering*. vol. 12, n° 2, 2006, p. 177-194.

Mostefa D., Hamon O., Choukri K. « Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR : Results from the first evaluation campaign », *Proc. LREC*, Genoa, Italy, 24-26 May 2006, p. 149-154.

Nießen S., Och F. J., Leusch G., Ney H. « An Evaluation Tool for Machine Translation : Fast Evaluation for MT Research », *Proc. LREC 2000*, Athens, Greece, 31 May - 2 June 2000, vol. 1/3, p. 39-45.

Och F. J., Ney H. « Discriminative Training and Maximum Entropy Models for Statistical Machine Translation », *Proc. ACL-02*, University of Pennsylvania, Philadelphia, USA, July 7-12, 2002, vol. 1/1, p. 395-302.

Pankowicz Z. Commentary on ALPAC Report, Personal Memorandum. December 1966. RADC, Griffiss Air Force Base. 160 p.

Papineni K., Roukos S., Ward T., Zhu V. « BLEU : a Method for Automatic Evaluation of Machine Translation », *Proc. ACL-02*, Philadelphia, USA, July 7-12, 2002, vol. 1/1, p. 311-318.

Przybocki M., Sanders G., Le A. « Edit Distance : A Metric for Machine Translation Evaluation », *Proc. LREC 2006*, Genoa, Italy, May 24-26, 2006, p. 2038-2043.

Rajman M., Hartley T. « Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores », *Proc. MT Summit VIII*, Santiago de Compostela, Spain, September 18-22, 2001, vol. 1/1, p. 29-34.

Rossato S., Blanchon H., Besacier L. « Speech-to-Speech Translation System Evaluation : Results for French for the NESPOLE! Project First Showcase », *Proc. ICSLP 2002*, Denver, USA, 16-20 September, 2002, 4 p.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J. « A Study of Translation Edit Rate with Targeted Human Annotation », *Proc. AMTA 2006*, Cambridge, MA, USA, August 8-12, 2006, vol. 1/1, p. 223-231.

Soricut R., Brill E. « A unified Framework for Automatic Evaluation using N-Gram Co-Occurrence Statistics », *Proc. ACL 2004*, Barcelona, Spain, July 21-26, 2004, 8 p.

Sugaya F., Yasuda K., Takezawa T., Yamamoto S. « Precise Measurement Method of a Speech Translation System's Capabilities with a Paired Comparison Method between the System and Humans », *Proc. MT Summit VIII*, Santiago de Compostela, Spain, 18-22 September, 2001, vol. 1/1, p. 345-350.

Taylor K., White J. « Predicting What MT Is Good for : User Judgments and Task Performance », *Proc. AMTA'98 (LNAI 1529)*, Langhorne, PA, USA, October 28-31, 1998, vol. 1/1, p. 364-373.

Tomás J., Mas J. À., Casacuberta F. « A Quantitative Method for Machine Translation Evaluation », *Proc. EACL 2003 – Workshop on Evaluation Initiatives in Natural Language Processing : are evaluation methods, metrics and ressources reusable?* Budapest, Hungary, April 14, 2003, vol. 1/1, 8 p.

Van Slype G. Critical Study of Methods for Evaluating the Quality of Machine Translation., Prepared for the Commission for the European Communities. Final Report., n° BR 19142. 30 novembre 1979. Bureau Marcel van Dijk. 187 p.

Wagner C.-K., Fisher M.-J. « The String-to-String Correction Problem », *Journal of the ACM*. vol. 21, n° 1, 1974, 168-173.

White J., Doyon J., Talbott S. « Determining the tolerance of text-handling tasks for MT output », *Proc. LREC-2000*, Athens, Greece, 31 May–2 June 2000, vol. 1/3, 29-32.

White J. S., O'Connell T., O'Mara F. E. « The ARPA MT Evaluation Methodologies : Evolution, Lessons and Further Approaches », *Proc. Technology Partnerships for Crossing the Language Barrier (the First Conference of the Association for Machine Translation in the Americas)*, Columbia, Maryland, USA, October 5-8, 1994, 13 p.