Chapitre 4

# CESTA : la Campagne d'Évaluation des Systèmes de Traduction Automatique

## 4.1. Introduction

Ce chapitre décrit la Campagne d'Évaluation des Systèmes de Traduction Automatique, CESTA, organisée dans le cadre de l'action Technolangue mise en place par plusieurs départements ministériels français. L'objectif de CESTA était d'organiser deux campagnes de test, la première mesurant la qualité des traductions produites automatiquement par des systèmes – commerciaux ou de recherche – dans un domaine général, la seconde mettant en évidence la capacité des systèmes à s'adapter à un domaine spécifique, en un temps limité. Parallèlement, CESTA visait à étudier la fiabilité des métriques d'évaluation automatiques, en vue d'offrir à la communauté un protocole et des données d'évaluation réutilisables, pour des systèmes traduisant de l'anglais ou de l'arabe vers le français.

Nous présenterons dans un premier temps le contexte de CESTA, ses objectifs et le protocole d'évaluation adopté, et notamment les métriques d'évaluation utilisées. Nous décrirons ensuite les résultats obtenus par les systèmes lors des deux phases de CESTA, d'abord dans le domaine général, puis dans un domaine spécifique, celui de la santé. Nous mettrons également en évidence les observations relevant de la méta-évaluation des métriques automatiques, c'est-à-dire la comparaison des résultats obtenus par celles-ci avec la perception humaine de la qualité, afin de déterminer les métriques automatiques qui reproduisent le mieux les jugements humains. En conclusion du chapitre, nous indiquerons les principales contributions de CESTA à l'avancement de l'évaluation des systèmes de traduction automatique.

---

Chapitre rédigé par Olivier HAMON, Andrei POPESCU-BELIS, Tony HARTLEY, Widad MUSTAFA EL HADI et Martin RAJMAN.

### 4.2. Contexte et objectifs de CESTA

Le projet dont la durée est de trois ans a débuté en janvier 2003 et a pris fin à l'autonome 2006. Il a été financé dans le cadre interministériel de l'action Technolangue (www.technolangue.net). Ce projet, à l'instar des autres projets Technolangue, a été intégré dans la plateforme EVALDA.

Les objectifs de CESTA sont multiples. Le premier objectif de ce projet est de définir un protocole d'évaluation incluant des jugements d'experts et des métriques automatiques pour évaluer des systèmes de traduction automatique aussi bien commercialisés qu'issus de la recherche académique. Un deuxième objectif est de tester les métriques automatiques lorsque la langue cible est le français ; certaines de ces métriques sont déjà bien connues (pour l'anglais tout du moins), alors que d'autres sont plus prospectives, notamment celles fondées sur la syntaxe et la sémantique [RAJ 01, RAJ 02]. La méta-évaluation de ces métriques se fait en comparant leurs résultats à ceux des jugements humains pour déterminer quelle métrique (ou combinaison de métriques) prédit le mieux les jugements humains. Le choix des aspects de la qualité des traductions automatiques s'est porté sur les deux principaux attributs utilisés par la communauté, à savoir l'adéquation (appelée aussi fidélité) et la fluidité (appelée aussi lisibilité ou intelligibilité). Ces attributs sont décrits plus en détail – et accompagnés de beaucoup d'autres – dans le cadre FEMTI pour l'évaluation des systèmes de traduction automatique [HOV 02].

Deux campagnes d'évaluation ont été entreprises dans le cadre du projet CESTA. La première campagne avait comme objectif de définir un protocole pour évaluer les systèmes de traduction automatique sur un corpus d'un domaine général et sans enrichissement terminologique. Le protocole a été adapté et réajusté en fonction des résultats obtenus lors de la première campagne. Nous avons ensuite testé les systèmes sur un corpus d'un domaine de spécialité (la santé), en offrant aux participants la possibilité d'une adaptation de leurs systèmes grâce à des données d'adaptation qui leur étaient fournies. Cela a permis de quantifier l'amélioration de la qualité des résultats et par conséquent l'amélioration de la performance des systèmes, en les comparant aux résultats obtenus sans enrichissement terminologique [MUS 02, 04 ; BAB 04].

### 4.3. Systèmes participants

Les systèmes suivants ont participé aux campagnes CESTA. Les descriptions abrégées qui suivent, par ordre alphabétique, ont été réalisées par les auteurs de ce chapitre en se fondant sur des textes fournis par les développeurs des systèmes. Pour la description officielle et détaillée des systèmes, nous renvoyons le lecteur à la documentation des systèmes, aux publications et aux sites Internet correspondants.

Le système de traduction automatique COMPRENDIUM est actuellement développé par la société Translendium SL, partie de Braintribe Lingua (voir www.translendium.com). Ce système est basé sur des règles linguistiques fournissant une analyse morphologique et syntaxique détaillée du texte source, à partir duquel est dérivée la version finale dans la langue cible. Le système est sensible au contexte, c'est-à-dire que le sens des mots est défini d'après leur fonction dans la phrase, accompagné aussi d'une analyse sémantique. Comprendium est un système de transfert à trois phases: analyse, transfert et génération. Le système analyse d'abord chaque phrase du texte source en utilisant l'information lexicale, puis, après avoir déterminé la catégorie grammaticale des mots, le système en fournit les structures syntaxiques sous-jacentes. Pendant la phase de transfert, ces structures sont transformées dans les structures équivalentes de la langue cible. Enfin, la phase de génération accomplit les transformations de surface restantes. La version évaluée était Comprendium Translator Desktop Power 2.0 accompagné des modules d'extraction terminologique et de création de corpus alignés et de listes de mots non traduits par le système.

Le système MLTS, en version français/arabe (société CIMOS, voir www.cimos.com) a été conçu par une équipe de linguistes, de traducteurs et d'ingénieurs informaticiens. C'est un logiciel de traduction automatique, conçu pour aider les traducteurs et non pour les remplacer dans leur tâche. MLTS est une application de traduction automatique de textes sources appartenant à différentes langues. La première version supporte l'anglais, le français et l'arabe, certaines langues sont en cours d'implémentation (comme le farsi), d'autres langues suivront en fonction de la demande (russe, espagnol, allemand, ou portugais).

Le moteur de traduction du RALI, Université de Montréal, est un système de traduction statistique basé sur les séquences de mots (*phrase-based SMT*) appelé RAMSES [PAT 05 ; LAN 06]. Les modèles de langue utilisés ici ont été entraînés avec l'implémentation *Kneser-Ney-discounting* de la boîte à outils SRILM. L'obtention du transducteur nécessite quant à lui plus d'efforts : un bitexte d'entraînement doit tout d'abord être aligné au niveau des mots. Dans le cadre des évaluations les alignements de Viterbi produits par les modèles IBM 2 de la boîte à outils GIZA++ ont été utilisés. Les paramètres du transducteur sont alors extraits des alignements de mots à l'aide d'un outil développé à l'interne.

Le système REVERSO (société Softissimo, voir www.softissimo.com et www.reverso.net) utilise un ensemble de règles propre à chaque direction de traduction, ce qui fait que la traduction se fait directement d'une langue A à une langue B sans passer par une étape intermédiaire (langue pivot). REVERSO se distingue, de plus, par l'intégration de règles propres à chaque langage à l'intérieur des différents moteurs d'analyse afin d'éviter que la séparation en modules trop délimités ne nuise à la qualité de l'analyse globale.

Le système de traduction arabe/français de l'Université Technologique de Aachen (RWTH) est un système statistique construit à partir d'un grand nombre de corpus, et notamment entraîné sur environ 4.7 million de paires de phrases extraites de la base de données des documents officiels des Nations Unies. Le noyau du système utilise un décodeur novateur, basé sur la recherche d'expressions, qui utilise une combinaison linéaire de plusieurs modèles incluant des modèles lexicaux à base de mots, des modèles de traduction d'expressions, ainsi qu'un modèle de langage. Ces modèles sont obtenus à partir des données d'entraînement d'une manière non supervisée, c'est-à-dire qu'aucune règle n'est encodée de manière explicite à l'intérieur même du système. Le moteur est robuste et capable de traduire en temps réel, traduisant plus de cent mots par seconde. Une description plus détaillée peut être trouvée en [HAS 06].

SDL Enterprise Translation Server (voir www.sdl.com/fr/products-home/) fait également partie des systèmes ayant participé à la campagne CESTA. Fondé sur une architecture de traduction directe, le système est aujourd'hui fourni à travers une version client/serveur qui assure une efficacité de traitement de plus de 3 millions de mots par heure. Le système existe dans sept paires de langues (bidirectionnel) plus trois langues additionnelles (monodirectionnel). Le système offre la capacité de traduction d'e-mails, de messages instantanés, de pages web ou de documents formatés, et permet l'utilisation de dictionnaires personnalisés par les utilisateurs et de dictionnaires spécialisés fournis par SDL International, intégrant la terminologie spécifique à un secteur ou à une société.

SYSTRAN (voir www.systran.fr) est l'un des premiers concepteurs de logiciels de traduction automatique. En 2002, SYSTRAN s'est doté d'une nouvelle architecture qui a mis en valeur les nombreux perfectionnements ayant précédé et rendu l'ensemble du système plus efficace. Associé à la performance accrue des ordinateurs d'aujourd'hui, SYSTRAN assure désormais une couverture importante des traductions. Tous les systèmes SYSTRAN utilisent le même moteur de traduction et s'appuient sur les dernières technologies de traitement du langage naturel. Les systèmes intègrent une technologie à états finis pour accélérer l'accès à d'importantes ressources linguistiques. En outre, de nombreuses applications multilingues sont conçues par SYSTRAN, par exemple des applications web, messagerie électronique, Intranet, gestion de contenu. Le département R&D tire aujourd'hui parti des évolutions permanentes et des investissements dans la linguistique et le traitement du langage naturel. L'objectif est de développer de nouveaux systèmes de traduction automatique pour améliorer la qualité et la robustesse de la traduction via une architecture multi-agent et multi-tâche.

Le système développé par l'Université Polytechnique de Catalogne (UPC) est un système statistique qui apprend à partir d'un corpus parallèle plusieurs modèles : un modèle de traduction, un modèle de la langue de sortie, un modèle qui donne un

bonus pour chaque mot en sortie, et deux modèles lexicaux [MAR 06]. Le système de traduction utilise une combinaison log-linéaire de ces différents modèles pour donner une probabilité à chaque hypothèse de traduction. Le corpus d'entraînement est utilisé pour optimiser le poids respectif de chaque modèle. La particularité de ce système par rapport à d'autres systèmes de traduction automatique statistique réside dans le modèle de traduction, qui est un modèle de langage de trigrammes d'unités bilingues. Ces trigrammes sont extraits de l'alignement par la méthode de Viterbi et peuvent être décrits formellement comme le plus petit ensemble de syntagmes qui conduit à une segmentation monotone du corpus parallèle. Le modèle de la langue de sortie est aussi un modèle de trigrammes. Le modèle de bonus par mot compense la tendance du système à préférer les phrases courtes. Enfin, les deux modèles lexicaux utilisent, pour un trigramme bilingue donné, les probabilités de traduction du modèle IBM-1 entre ses côtés source et cible.

### 4.4. Métriques utilisées

La démarche adoptée dans les deux campagnes a été de prendre comme référence de la qualité les scores issus des jugements humains, et ensuite d'établir dans quelle mesure les métriques automatiques permettaient d'émuler ceux-ci, ou du moins le classement déduit. Les campagnes visaient ainsi une méthode de *classement* des systèmes et non pas, par exemple, l'identification des erreurs. De ce fait il a été légitime de retenir des méthodes automatiques dont les résultats se situent au niveau du texte ou de la collection de textes et non pas au niveau de la phrase. Le protocole d'évaluation était ainsi analogue à celui défini par le NIST (*National Institute of Standards and Technology*) aux États-Unis [NIS 03].

#### 4.4.1. Métriques automatiques

Les cinq méthodes ou métriques automatiques principales qui ont été déployées lors des deux campagnes se divisent en deux catégories. Celles de la première catégorie adoptent comme paramètre principal la proximité de la traduction automatique par rapport à une ou plusieurs traductions humaines, dites *traductions de référence*, fournies par des traducteurs professionnels. En revanche, les méthodes de la deuxième catégorie ne font pas appel à des traductions de référence des textes source, mais à des corpus autonomes représentatifs de la langue cible et, éventuellement, de la langue source.

Les méthodes fondées sur la proximité aux traductions de référence seront dénotées dans la suite par les abréviations BLEU, NIST et WNM. La métrique BLEU (*BiLingual Evaluation Understudy*) est une métrique automatique, développée par IBM [PAP 01]. Comme pour sa variante élaborée par le NIST [DOD 02], le principe en est de comparer les phrases sorties du système aux phrases de

référence correspondantes sur la base de séquences de n-grammes (à savoir une séquence ordonnée de *n* mots).

D'une manière simplifiée, on peut considérer que ces métriques comptent le nombre de mots d'une phrase à évaluer contenus dans une ou plusieurs phrases de référence. Une traduction est d'autant meilleure qu'elle partage un grand nombre de n-grammes avec une ou plusieurs de ces traductions de référence, tel un score de précision cumulée. Les unigrammes correspondraient alors à la fidélité et les valeurs plus élevées de *n* (bigrammes, trigrammes, etc.) rendraient compte de la fluidité. De plus, BLEU applique une pénalité de brièveté aux traductions ayant une différence de taille importante par rapport à la traduction de référence, ceci afin d'éviter qu'un mot apparaisse plus fréquemment dans le document traduit qu'il ne le peut dans le document de référence, et d'empêcher ainsi des scores trop importants. La méthode NIST applique différents poids aux n-grammes, utilisant le gain d'information et une pénalité sur la taille du document traduit. La méthode BLEU et sa variante NIST, même si elle sont contestées, restent actuellement les métriques les plus utilisées dans le domaine de la traduction automatique.

La méthode WNM (*Weighted N-gram Model*) [BAB 04 ; BAB 05] part du principe que, étant donné la variation légitime inhérente à toute traduction, tous les mots n'ont ni le même besoin ni le même droit d'être préservés par un processus de traduction. Tout en adoptant la formule des n-grammes, elle introduit une pondération pour les mots identifiés comme significatifs, cette identification se faisant sur la base de leur fréquence relative à l'intérieur d'un document par rapport à toute une collection de documents (il s'agit d'une variante du score *tf.idf* utilisé en recherche d'information). WNM permet alors de calculer trois types de scores : la précision, le rappel et la *f*-mesure qui pondère également la précision et le rappel. On peut considérer que le rappel correspond à l'adéquation, tandis que la *f*-mesure correspond à la fluidité.

Les méthodes faisant appel à des corpus autonomes étaient le X-score et le D-score [RAJ 01], [RAJ 02]. Celles-ci ont été utilisées à titre expérimental car leur fiabilité n'est pas entièrement démontrée, comme le suggèrent aussi les scores que nous citerons. Le X-score est basé sur la répartition de catégories morphosyntaxiques (dans le cas présent) ou de dépendances syntaxiques dans le document traduit. Il cherche à établir une mesure de la grammaticalité du texte qui est censé correspondre à sa fluidité. Il n'y a donc à aucun moment besoin d'un quelconque document source. Cependant, une phase préalable d'apprentissage est nécessaire pour établir une répartition type de l'information grammaticale choisie dans un document. Cette répartition se calcule à partir d'un corpus de documents en langue cible où chacun dispose d'un score de fluidité, comme, par exemple, le corpus du projet DARPA 94 sur l'anglais. Nous avons reproduit ce type de corpus pour le français au cours de CESTA.

Le D-score est une mesure de la conservation du contenu sémantique après la traduction d'un document. L'idée principale est d'utiliser un modèle vectoriel sémantique. Il s'agit de construire un espace vectoriel pour les langues source et cible, sur la base d'un corpus de textes dans la langue source et le corpus correspondant de traductions dans la langue cible. A partir de cet espace vectoriel le D-score calcule la position d'un document source, et compare cette position à celle du document cible sorti du système de traduction. En d'autres termes, la mesure est basée sur la similarité des espaces vectoriels d'un document source et sa traduction, par rapport à des corpus de référence dans les deux langues, source et cible. L'hypothèse est que plus la structure de l'espace vectoriel du document source est similaire à celle du document cible, plus la qualité de la traduction est élevée.

### 4.4.2. Jugements humains

Les deux campagnes d'évaluation ont inclus une évaluation humaine afin de permettre la méta-évaluation des métriques automatiques, en comparant les scores humains aux scores automatiques. Deux critères d'évaluation ont été retenus, la *fluidité* et l'*adéquation*, d'après les campagnes d'évaluation DARPA [WHI 94].

Pour que les juges puissent évaluer les traductions automatiques, une interface Web a été développée. À l'aide de cette interface, chaque segment traduit – correspondant à une phrase la plupart du temps – a été évalué par deux juges différents. Les segments de toutes les traductions disponibles ont été mélangés, pour que les juges ne puissent pas mémoriser des informations sur deux segments adjacents, ce qui aurait biaisé l'évaluation.

Pour chacun des deux critères les évaluateurs ont répondu à une seule question. Pour la fluidité, la question posée était : « Ce texte est-il écrit en bon français ? ». Les évaluateurs avaient alors à choisir un score sur une échelle de cinq valeurs, allant de « français impeccable » (note égale à 5) à « français incompréhensible » (note égale à 1). Pour l'adéquation, la question posée était : « À quel point le sens exprimé dans la traduction à noter est le même que celui du texte de référence ? ». Une échelle de cinq valeurs était également proposée aux évaluateurs, allant de « tout le sens » (valeur notée avec 5) à « aucun sens » (valeur notée avec 1).

Les échelles de valeur intermédiaires étaient également nommées explicitement lors de la première campagne. Mais il a finalement été décidé de laisser le libre choix aux évaluateurs des valeurs intermédiaires en ne spécifiant par leurs noms, mais seulement les valeurs des scores (2, 3, 4) pour la seconde campagne. Les évaluations de la fluidité et de l'adéquation ont été réalisées en même temps pour la première campagne, mais de manière séparées pour la seconde : les évaluateurs évaluaient à la suite tous leurs segments selon la fluidité, puis selon l'adéquation. La

dernière modification entre les deux campagnes d'évaluation concerne l'évaluation de la traduction officielle en plus des traductions automatiques lors de la seconde campagne. En effet, seules les traductions automatiques ont été évaluées pour la première campagne.

### 4.4.3. Méthodes de méta-évaluation des métriques

Afin de prouver la validité des métriques utilisées, automatiques et humaines, il est nécessaire de mesurer leur robustesse intrinsèque, à savoir leur capacité à produire des scores constants pour des données ayant une qualité constante (la similarité de la qualité étant jugée par les juges humains). Le but est notamment d'observer la stabilité des métriques par rapport à la variation des données, à qualité supposée constante. La meilleure façon de mesurer cette stabilité, lorsque les données sont limitées, est d'échantillonner les données afin d'obtenir une grande quantité d'ensembles de traductions différents. L'hypothèse est que des échantillons différents produits par un même système auront une « qualité stable », qui devra être mise en évidence par une métrique « stable ». Nous avons choisi de créer 100 échantillons des textes à évaluer, chaque échantillon étant produit aléatoirement. Les échantillons étaient de la même taille que l'ensemble de départ, et de nombreuses données (paires de segments source-cible) étaient par conséquent dupliquées dans chaque échantillon.

Pour tester la robustesse de l'évaluation *humaine*, nous avons échantillonné d'une part sur les segments ainsi que d'autre part sur les juges. Quatre types de calculs ont été produits pour les deux échantillonnages : le score moyen des échantillons ; le rang moyen des échantillons ; le classement des scores moyens des échantillons ; le classement des rangs moyens des échantillons. Ainsi avec les deux critères de l'évaluation humaine (la fluidité et l'adéquation), nous avons obtenu seize mesures différentes pour tester la robustesse de l'évaluation humaine. L'ensemble de ces résultats figure dans le rapport scientifique détaillé de la campagne CESTA.

Pour tester la robustesse de chacune des évaluations automatiques, nous avons repris les échantillons obtenus précédemment sur les segments, et effectué les mêmes calculs, sur les scores moyens aussi bien que sur les rangs moyens. Ainsi nous obtenons (seulement) huit mesures différentes par métrique, l'échantillonnage sur les juges ne pouvant être considéré.

Lors de l'évaluation humaine, chaque segment est évalué par deux juges différents, ce qui nous permet de justifier la validité des évaluations. Nous avons utilisé trois méthodes pour définir l'accord entre les juges. La première méthode calcule le ratio entre le nombre de scores identiques sur deux jugements d'un même segment et le nombre total de segments. La deuxième calcule le ratio entre le

nombre de scores différents d'au plus *n* unités (*n* = 1, 2,…) sur deux jugements d'un même segment et le nombre total de segments (la première méthode correspond donc à *n* = 0). La troisième méthode prédit la probabilité de similarité entre les paires de jugements (ses résultats figurent dans le rapport final CESTA seulement).

Il est également important de valider les métriques automatiques par rapport à l'évaluation humaine, c'est-à-dire de réaliser une méta-évaluation des métriques. Deux types de résultats étant disponibles pour chacune des métriques – les scores et les rangs de chacun des systèmes – il est possible de calculer une corrélation avec chacun d'eux. En ce qui concerne les scores, nous avons calculé une corrélation de Pearson entre les métriques humaines et les métriques automatiques notamment. Pour les rangs, un calcul de proximité est nécessaire, nous avons choisi pour cela de calculer la distance de Hamming [HAM 50].

### 4.5. Première campagne

#### 4.5.1. Organisation et déroulement

L'objectif de la première campagne était de fournir une première mesure de la qualité des systèmes avec l'ensemble des métriques prévues pour les campagnes CESTA, en utilisant des textes du domaine général. Les deux sens de traduction, de l'anglais et respectivement de l'arabe vers le français ont été testés [SUR 05].

Cinq des six systèmes traduisant de l'anglais vers le français et deux des trois systèmes traduisant de l'arabe vers le français ont participé dès la première campagne. Afin de préserver l'anonymat de ces systèmes, selon les conventions de CESTA, nous utiliserons le système de notation suivant. Les différentes instances des systèmes seront numérotées de S1 à S13, avec duplication entre la première et la seconde campagne. Cette duplication rend impossible l'identification des systèmes d'une campagne à une autre, ce qui correspond d'ailleurs à un changement de version et de données, qui fait qu'il n'est pas pertinent de comparer un système de la première campagne à la seconde. En outre, nous préciserons entre parenthèses la langue source (« en » ou « ar ») et numéro de la campagne (1, 2a ou 2b). La première campagne concerne ainsi les systèmes de S1(en,1) à S5(en,1) pour l'anglais, et les systèmes S6(ar,1) et S7(ar,1) pour l'arabe, alors que la seconde concerne les systèmes S8(en,2) à S12(en,2) pour l'anglais et S13(ar,2) pour l'arabe.

Pour la seconde campagne, déclinée en deux phases, les systèmes seront plus spécifiquement notés avec 2a (avant adaptation) ou 2b (après adaptation), par exemple S8(en,2a) ou S13(ar,2b). Il n'est donc pas possible de comparer les performances entre les deux campagnes pour les systèmes ayant participé aux deux, mais il sera possible de comparer les performances d'un même système avant et

après adaptation du domaine, dans la seconde campagne, par exemple les scores de S9(en,2a) et S9(en,2b), ou S13(ar,2a) et S13(ar,2b). S13 est le seul système français/arabe ayant participé à la seconde campagne, alors que deux systèmes français/arabe ont participé à la première campagne.

### 4.5.2. Données et métriques utilisées

Les textes utilisés pour la première campagne proviennent du *Journal officiel des Communautés Européennes* (JOC, 15 documents) pour le sens anglais/français, et des *Actes de la 32e Conférence générale de l'UNESCO* (16 documents) pour le sens arabe/français. Les deux corpus, segmentés en phrases de longueur assez variable, contiennent environ 20 000 mots chacun en langue source. Ces textes ont été introduits dans un corpus beaucoup plus vaste d'environ 200 000 mots pour chaque langue, thématiquement semblable, ce qui permet le masquage des données de test et évite que les participants ne prennent directement connaissance des données de test et ne modifient (consciemment ou non) leur système pour améliorer les performances sur ces données. Toutes les données, encodées au format UTF-8, suivent le format d'annotation défini par le NIST [NIS 03].

Quatre traductions de référence ont été produites pour chaque document testé, afin que les évaluateurs humains et surtout les métriques d'évaluation automatique puissent les utiliser. Outre la traduction originale de chaque document (le JOC et les actes de l'UNESCO sont traduits, voire rédigés en français), trois agences de traduction ont été sollicitées par les organisateurs de CESTA pour produire chacune une traduction française de bonne qualité. Les quatre traductions ont été utilisées par les métriques automatiques, mais une seule a été donnée comme référence aux juges humains, en sélectionnant à l'aide d'un expert la meilleure des quatre traductions.

### 4.5.3. Résultats

Nous présentons d'abord dans le Tableau 4.1 ci-dessous les résultats de l'évaluation humaine des traductions produites par les systèmes. Ces traductions ont été évaluées phrase par phrase en termes de fluidité et d'adéquation (métriques définies ci-dessus) sur une échelle de 1 à 5, où 5 est le score d'une traduction idéale, et le tableau présente la moyenne de ces scores sur tous les segments ou phrases.

Plusieurs mesures permettent de juger de la fiabilité de ces scores. Ainsi, chaque segment ayant été évalué par deux juges, on observe qu'environ 40% des segments (toutes langues et métriques confondues) reçoivent exactement la même note. Lorsque l'on considère, outre l'identité de la note, la distance entre les notes des deux juges pour un même segment, on observe que le nombre de jugements qui sont

identiques ou qui diffèrent d'au plus une unité (p.ex. 4 *versus* 5) est d'environ 84% pour la fluidité et d'environ 78% pour l'adéquation, les chiffres exacts étant semblables pour l'anglais et pour l'arabe.

Les résultats des mesures de fiabilité obtenus par échantillonnage sur les segments évalués (comme expliqué dans la section 4.4.3) conduisent à des écarts-type compris entre 0.03 et 0.09 pour la fluidité et l'adéquation. Il est également possible de calculer des intervalles de confiance directement à partir des scores des segments, pour chaque système, sans passer par l'échantillonnage : ces intervalles sont compris entre ±0.05 et ±0.08 pour les différents scores et les différents systèmes, avec une majorité d'intervalles compris entre ±0.06 et ±0.07. La même technique d'échantillonnage permet d'ailleurs de calculer la probabilité du classement observé pour chaque système, qui est indiquée dans le même Tableau 4.1 en la calculant sur les scores moyens (l'utilisation des rangs moyens pour chaque échantillon réduit quelque peu cette probabilité).

| Système | Fluidité | | Adéquation | |
|---|---|---|---|---|
| | Score (1-5) | Classement | Score (1-5) | Classement |
| S1(en,1) | 2.41±.05 | 5 (p=1) | 2.96±.06 | 5 (p=1) |
| S2(en,1) | 3.04±.06 | 1 (p=.99) | 3.54±.06 | 1 (p=1) |
| S3(en,1) | 3.01±.06 | 2 (p=.99) | 3.43±.06 | 2 (p=1) |
| S4(en,1) | 2.67±.06 | 4 (p=1) | 3.18±.07 | 4 (p=.89) |
| S5(en,1) | 2.84±.07 | 3 (p=1) | 3.24±.06 | 3 (p=.89) |
| S6(ar,1) | 1.79±.08 | 1 (p=1) | 2.24±.08 | 1 (p=1) |
| S7(ar,1) | 1.33±.06 | 2 (p=2) | 1.66±.07 | 2 (p=2) |

**Tableau 4.1.** *Résultats des jugements humains pour la première campagne : scores sur une échelle de 1 à 5 (5 étant la meilleure note et 1 la moins bonne) avec leurs intervalles de confiance, et classements avec leurs probabilités*.

L'évaluation humaine indique donc avec une fiabilité suffisante le même classement des systèmes selon la fluidité et l'adéquation de leurs traductions. Toutefois, on constate que les deux meilleurs systèmes, S2 et S3, paraissent impossibles à départager avec certitude, car les intervalles de confiance à 95% ont une intersection non nulle. De même, S4 et S5 présentent une différence non significative des scores d'adéquation, ainsi que des scores de fluidité proches. Dans le sens arabe/français, S6 apparaît clairement meilleur que S7. Notons que, bien que les scores pour le sens arabe/français soient nettement en dessous de ceux obtenus pour le sens anglais/français, il est difficile de comparer globalement ces scores puisque les données de test sont différentes, et donc potentiellement aussi de

difficultés différentes. Si ces scores avaient été obtenus sur un même corpus (trilingue) nous aurions pu conclure que les systèmes pour le sens arabe/français sont moins développés que ceux pour le sens anglais/français. Enfin, notons que l'adéquation est globalement plus élevée que la fluidité pour la plupart des systèmes.

Les scores obtenus grâce aux métriques automatiques sont quant à eux indiqués dans le Tableau 4.2 ci-dessous, en utilisant les n-grammes jusqu'à 4 pour la métrique BLEU, et jusqu'à 5 pour la métrique NIST, dans les deux cas en tenant compte également de la casse des mots. La mesure WNM, dont on fournit ici les valeurs de f-mesure (notées WNMf) est en réalité la moyenne des scores WNMf obtenus en considérant tour à tour chaque traduction humaine comme référence. Le tableau fait figurer également les écarts-types des scores, obtenus par échantillonnage des segments comme décrit ci-dessus.

| **Système** | **BLEU** | | **NIST** | | **WNMf** | | **X-score** | | **D-score** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **%** | **cl.** | **v. a.** | **cl.** | **%** | **cl.** | **v. a.** | **cl.** | **v. a.** | **cl.** |
| S1(en,1) | 37.60±2.6 | 5 | 9.03±.30 | 5 | 49.48±.33 | 5 | *40.87* | *1* | *69.50* | *1* |
| S2(en,1) | 44.76±1.7 | 3 | 9.78±.22 | 3 | 55.57±.39 | 3 | *40.02* | *3* | *59.82* | *3* |
| S3(en,1) | 57.33±2.0 | 1 | 11.03±.25 | 1 | 57.29±.39 | 1 | *40.53* | *2* | *58.22* | *5* |
| S4(en,1) | 46.64±1.9 | 2 | 9.98±.25 | 2 | 56.98±.35 | 2 | *37.97* | *5* | *60.01* | *3* |
| S5(en,1) | 43.87±2.9 | 4 | 9.65±.53 | 4 | 53.22±.42 | 4 | *38.91* | *4* | *69.29* | *2* |
| S6(ar,1) | 19.13±1.0 | 1 | 7.18±0.15 | 1 | 40.93±.28 | 1 | *38.38* | *2* | - | |
| S7(ar,1) | 8.21±0.5 | 2 | 4.69±0.11 | 2 | 33.14±.28 | 2 | *40.99* | *1* | - | |

**Tableau 4.2**. *Scores des métriques automatiques pour les traductions de la première campagne : valeurs, en pourcent (%) ou en valeur absolue (v. a.), écarts-types (±) et classements (cl.). Les deux métriques expérimentales figurent en italique.*

Les scores des métriques qui utilisent les n-grammes apparaissent particulièrement homogènes, non tant par leurs valeurs que par les classements *identiques* qui en résultent. Pour ce qui est du sens anglais/français, les écarts-types, affichés ici à la place des intervalles de confiance, indiquent une proximité assez grande de S4, S2 et S5, alors que S3 semble se démarquer comme ayant les scores les plus élevés, et S5 les plus bas. Pour ce qui est de l'arabe, les métriques BLEU, NIST et WNMf mettent en évidence une différence nette entre S6 et S7, le premier étant clairement meilleur.

Les métriques expérimentales X-score et D-score présentent des résultats moins concordants avec les autres métriques et avec les jugements humains au niveau des

classements. Le D-score ne peut pas d'ailleurs pas être calculé pour l'arabe par manque de ressources, et le X-score est en contradiction avec les autres classements.

### 4.5.4. Discussion

La comparaison des classements établis par les juges humains avec ceux résultant des métriques automatiques permet aussi de mesurer la confiance que l'on peut accorder à ces dernières. On peut également calculer une corrélation de Pearson entre les scores, affichée dans le Tableau 4.3 ci-dessous pour la fluidité et l'adéquation respectivement. Les corrélations qui sont acceptables, mais plus faibles que celles observées dans la littérature pour ces mêmes métriques appliquée à l'anglais cible.

Le classement établi par les juges humains correspond à (S2>S3)>(S5>S4)>S1, les parenthèses indiquant des scores comparables (cf. Tableau 4.1). Or, le classement des trois métriques automatiques est S3>S4>S2>S5>S1, ce qui diffère en plusieurs points du précédent. La fiabilité des métriques automatiques n'est donc pas parfaite. Comme annoncé par ses créateurs, la métrique WNMf [BAB 04] est relativement plus proche des juges humains que BLEU ou NIST.

Le X-score ne présente pas de corrélation significative avec les juges humains, et son utilisation n'est donc pas concluante. Le D-score présente une corrélation négative, ce qui ne compromet pas son utilité si on inverse ses résultats, en le considérant désormais dans ce qui suit comme une mesure de distance, et non de proximité [HAM 06b].

| | **BLEU** | **NIST** | **WNMf** | **X-score** | **D-score** |
|---|---|---|---|---|---|
| **Fluidité** | 0.69 | 0.63 | 0.72 | *0.02* | *-0.63* |
| **Adéquation** | 0.69 | 0.64 | 0.72 | *0.08* | *-0.71* |

**Tableau 4.3**. *Corrélation de Pearson (échelle -1 à 1) entre les métriques automatiques et les juges humains, première campagne, sens anglais/français.*

Enfin, pour expliquer la différence de classement observée entre les juges humains et les métriques automatiques – (S2 > S3) > (S5 > S4) > S1 contre S3 > S4 > S2 > S5 > S1 – nous faisons l'hypothèse que les métriques automatiques à base de n-grammes favorisent le système S4, qui utilise en réalité un modèle de langage pour sélectionner les phrases traduites, et que les performances de S2 et S3 sont objectivement trop proches pour être distinguées. Dans tous les cas, les différences majeures entre les systèmes semblent bien capturées par les métriques automatiques, bien que leur fiabilité paraisse moindre lorsque le français est la langue cible que

lorsqu'il s'agit de l'anglais, par comparaison avec les résultats disponibles dans la littérature.

### 4.6. Seconde campagne : adaptation à un domaine

#### 4.6.1. Organisation et déroulement

Au delà de l'évaluation comparative des performances, l'un des objectifs de la seconde campagne était d'améliorer la robustesse et la fiabilité du protocole d'évaluation [HAM 06a]. Mais l'objectif spécifiquement innovant de cette seconde campagne était surtout de tenter de mettre en évidence l'impact d'une adaptation des systèmes au domaine des textes à traduire. C'est pourquoi, deux séries de scores ont été établies avec le protocole d'évaluation amélioré : l'une utilise des versions « génériques » des systèmes, et l'autre utilise des versions adaptées au domaine choisi par les organisateurs de CESTA, à savoir celui de la santé. Les systèmes participant à la seconde campagne ont ainsi reçu un corpus bilingue d'adaptation jugé représentatif du domaine de la santé, leur permettant d'adapter leurs systèmes pendant deux semaines avant l'évaluation proprement dite.

Cinq systèmes ont participé à la seconde campagne pour le sens anglais/français, dont quatre avaient déjà participé à la première campagne, et un seul pour le sens arabe/français. Afin de préserver l'anonymat de ces systèmes, comme indiqué déjà, nous utiliserons des numéros différents de la première campagne, de S8 à S13. Nous indiquerons également, comme ci-dessus, la langue source et la version du système, à savoir avant ou après l'adaptation au domaine (« 2a » ou « 2b »). Ainsi, par exemple, S10(en,2a) désigne le système S10 pour le sens anglais/français avant adaptation au domaine.

Selon la nature du système, l'adaptation au domaine, appelée parfois aussi enrichissement terminologique, peut par exemple impliquer l'utilisation d'un dictionnaire préexistant du domaine de la santé, ou l'extraction de termes du corpus représentatif fourni par les organisateurs de CESTA et leur insertion dans un dictionnaire utilisateur, ou encore l'entraînement du système sur ce corpus bilingue.

Aucune contrainte n'était donc imposée quant à la procédure d'adaptation, le risque étant par conséquent que les participants cherchent à obtenir davantage de données similaires au corpus fourni (extrait de sites Internet, comme indiqué ci-après) et parviennent à se procurer les données de test elles-mêmes. Lorsque cela s'est produit, dans un seul cas, un accord avec les développeurs du système a permis de revenir à une version du système non adaptée aux données de test, et ce sont ces résultats qui figurent ci-après. En effet, l'évaluation de la version du système qui a tenu compte des données de test découvertes par hasard ne rend pas compte des

performances véritables du système sur des textes non vus auparavant, car elle conduit à des scores artificiellement élevés.

### 4.6.2. Données et métriques utilisées

Le corpus de test anglais/français spécifique au domaine de la santé est composé de seize documents provenant du site Internet bilingue canadien *Santé Canada* (http://www.hc-sc.gc.ca). Le corpus arabe/français, dans le même domaine, est composé de trente documents provenant des sites Internet multilingues de deux organismes internationaux et d'une ONG : l'UNICEF (http://www.unicef.org), l'Organisation Mondiale de la Santé (http://www.who.int) et *Family Health International* (http://www.fhi.org). Chacun de ces corpus dispose d'une traduction officielle. Les textes source contiennent environ 20 000 mots, correspondant à 917 segments pour l'anglais et à 824 segments pour l'arabe.

Outre ces données de test, un corpus d'entraînement représentatif du même domaine a été fourni aux participants afin de leur permettre d'adapter leurs systèmes avant l'évaluation (phase 2b). Ce corpus provenait des mêmes sources que le corpus de test (avec des documents différents) et comptait également environ 20 000 mots dans chaque langue. Par ailleurs, les données de test ont une fois de plus été dispersées dans un corpus de masquage environ dix fois plus grand.

Comme dans la première campagne, les documents ont été segmentés au niveau de la phrase, et trois traductions de référence, en plus de la traduction officielle, ont été produites par des agences de traduction. La qualité de la traduction provenant de la même organisation que le texte source ayant été jugée inférieure aux traductions commandées (par un expert indépendant), c'est l'une des trois traductions commandées qui a été employée comme référence pour tous les évaluateurs humains. En revanche, toutes les quatre traductions ont été utilisées comme références pour les métriques automatiques.

En ce qui concerne l'évaluation humaine, seulement une partie des phrases traduites par les systèmes participants a été évaluée – environ un tiers des segments produits par chaque système – à cause de la difficulté à trouver des juges qualifiés. Parmi les 7150 segments produits par les systèmes et par l'un des traducteurs de référence (i.e. 6 fois 917 plus deux fois 824), un total de 2304 segments ont été évalués par 48 juges au total. Chaque segment a été évalué deux fois par deux juges différents, ce qui correspond ainsi à 96 segments par système et par juge. Le protocole humain d'évaluation a été identique à la première campagne, mis à part que l'adéquation et la fluidité ont été estimées en deux étapes séparées, pour éviter des corrélations potentielles entre les scores dues à l'évaluation simultanée par le même juge humain. Les juges humains étaient par ailleurs familiers avec le domaine

de la santé, pour garantir un jugement objectif de la qualité terminologique des traductions.

### 4.6.3. Résultats

Nous présenterons d'abord les performances enregistrées par les métriques automatiques pour les systèmes non adaptés (campagne 2a). Toutefois, ce sont seulement les résultats des systèmes adaptés (campagne 2b) qui ont été soumis également aux juges humains et ont fait l'objet d'une analyse de fiabilité approfondie. La comparaison avec les systèmes adaptés au domaine pourra donc se faire suivant les scores des métriques automatiques, dont la robustesse est acceptable selon les chiffres observés pour la campagne 2b.

| Système | BLEU | | NIST | | WNMf | | X-score | | D-score | |
|---|---|---|---|---|---|---|---|---|---|---|
| | % | cl. | v. a. | cl. | % | cl. | v. a. | cl. | v. a. | cl. |
| S8(en,2a) | 32.83 | 4 | 7.76 | 4 | 48.09 | 4 | *34.91* | *4* | *42.56* | *2* |
| S9(en,2a) | 37.96 | 1 | 9.14 | 1 | 51.37 | 1 | *36.68* | *3* | *44.02* | *3* |
| S10(en,2a) | 33.80 | 3 | 8.58 | 3 | 50.02 | 2 | *38.57* | *1* | *44.13* | *4* |
| S11(en,2a) | 35.19 | 2 | 8.71 | 2 | 49.79 | 3 | *37.86* | *2* | *46.61* | *5* |
| S12(en,2a) | 25.61 | 5 | 7.38 | 5 | 48.06 | 5 | *34.60* | *5* | *40.67* | *1* |
| S13(ar,2a) | 36.71 | 1 | 8.72 | 1 | 54.29 | 1 | *40.66* | *1* | - | - |

**Tableau 4.4**. *Scores des métriques automatiques pour les résultats de la seconde campagne avant adaptation au domaine (2a).*

Les scores avant adaptation au domaine (voir Tableau 4.4) obtenus grâce aux métriques fondées sur les n-grammes conduisent encore une fois à un classement concordant, ce qui permet d'accorder à ces résultats une confiance acceptable. L'un des systèmes, S9(en,2a) se détache comme particulièrement meilleur, suivi du groupe S10-S11, puis du groupe S8-S12. Au sein de ces deux paires, il semble difficile de départager de manière fiable les systèmes. Pour la paire S8-S12 une différence très notable est toutefois observée pour la métrique BLEU, mais celle-ci n'est pas confirmée par le score WNMf. Enfin, le X-score présente cette fois-ci des résultats proches des trois métriques précédentes, alors que le D-score semble avoir une corrélation inverse.

Les traductions automatiques produites après adaptation au domaine (campagne 2b) ont quant à elles fait l'objet d'une évaluation plus détaillée, en commençant par des juges humains utilisant les interfaces améliorées après la première campagne. Les moyennes des scores obtenus, avec leurs intervalles de confiance calculés

directement sur l'ensemble des segments traduits par chaque système, figurent dans le Tableau 4.5, accompagnées des classements correspondants et de leurs probabilités (ou fréquences) calculées également par échantillonnage. L'accord entre les juges a été globalement amélioré par rapport à la première campagne, puisque celui-ci est de 43% et respectivement 46% pour l'identité des notes de fluidité et d'adéquation des segments, contre 41% et 38% pour la première campagne. Lorsqu'une une différence d'une unité est autorisée entre les scores, la similarité est d'environ 80% pour les deux mesures, ce qui améliore l'accord sur l'adéquation par rapport à la première campagne, mais pas sur la fluidité.

| **Système** | **Fluidité** | | **Adéquation** | |
|---|---|---|---|---|
| | **Score (1-5)** | **Classement** | **Score (1-5)** | **Classement** |
| S8(en,2b) | 2.28±.10 | 5 (p=1) | 2.84±.11 | 5 (p=1) |
| S9(en,2b) | 3.19±.11 | 3* (p=.51) | 3.15±.10 | 4 (p=1) |
| S10(en,2b) | 3.30±.10 | 2 (p=.95) | 3.44±.11 | 2 (p=.88) |
| S11(en,2b) | 3.19±.10 | 3* (p=.51) | 3.38±.11 | 3 (p=.88) |
| S12(en,2b) | 3.57±.09 | 1 (p=1) | 3.78±.09 | 1 (p=1) |
| S13(ar,2b) | 3.08±.11 | 1 (p=1) | 2.70±.12 | 1 (p=1) |

**Tableau 4.5.** *Résultats des jugements humains pour la seconde campagne, après adaptation des systèmes au domaine : scores sur une échelle de 1 à 5 avec leurs intervalles de confiance, et classements avec leurs probabilités (* dénote deux systèmes* ex aequo*).*

On constate sur le Tableau 4.5 que les scores les plus élevés sont obtenus par le système S12, alors que les performances des systèmes S9, S10 et S11 sont très proches, parfois indiscernables, et que le système S8 est loin derrière. Dans la plupart des cas, l'adéquation des traductions est supérieure à leur fluidité.

En outre, les humains ont aussi évalué (sans explicitement le savoir) la traduction officielle de chaque segment (i.e. celle disponible sur le site d'origine), à des fins de comparaison avec les résultats des systèmes. Pour le sens anglais/français, la traduction officielle a obtenu un score de fluidité de 4.55 et un score d'adéquation de 4.20, alors que pour le sens arabe/français ces valeurs étaient respectivement de 4.70 et 3.51. Le score maximal étant de 5, on constate que les valeurs de fluidité sont proches du maximum, mais que l'adéquation se situe bien en dessous, avec des valeurs étonnamment faibles pour une traduction humaine. Plusieurs explications peuvent être envisagées, sans que l'on puisse les départager : sévérité des juges, compréhension trop littérale par les juges de la notion d'adéquation, faible qualité intrinsèque des traductions officielles, ou différences linguistiques entre les juges français et les traductions québécoises (pour le sens anglais/français).

Les scores de fluidité/adéquation des traductions officielles sont toujours supérieurs à ceux des meilleurs systèmes du Tableau 4.5, bien que la distance entre eux ne soit pas toujours très grande. À l'inverse des systèmes de traduction automatique, les traductions humaines reçoivent de bien meilleures notes pour leur fluidité que pour leur adéquation, ce qui renforce peut-être la deuxième hypothèse ci-dessus, celle d'une compréhension trop restrictive de la notion d'adéquation. Cette différence peut être également due à la fréquence plus élevée des reformulations dans les traductions humaines, alors que les systèmes traduisent plus souvent en suivant l'ordre du texte.

| Système | BLEU | | NIST | | WNMf | | X-score | | D-score | |
|---|---|---|---|---|---|---|---|---|---|---|
| | % | cl. | v. a. | cl. | % | cl. | v. a. | cl. | v. a. | cl. |
| S8(en,2b) | 33.04±3.00 | 2 | 8.35±0.40 | 5 | 50.05±0.66 | 4 | *35.58* | *5* | *41.52* | *1* |
| S9(en,2b) | 38.07±2.70 | 4 | 9.13±0.34 | 2 | 51.50±0.71 | 3 | *36.71* | *4* | *44.06* | *3* |
| S10(en,2b) | 36.60±2.40 | 5 | 8.97±0.31 | 3 | 52.47±0.68 | 2 | *38.50* | *1* | *44.06* | *3* |
| S11(en,2b) | 35.74±4.60 | 3 | 8.77±0.49 | 4 | 50.59±0.66 | 5 | *38.15* | *2* | *46.16* | *5* |
| S12(en,2b) | 40.43±1.00 | 1 | 9.27±0.17 | 1 | 56.25±0.77 | 1 | *37.65* | *3* | *42.20* | *2* |
| S13(ar,2b) | 40.82 | 1 | 8.95 | 1 | 54.15 | 1 | *42.04* | *1* | - | - |

**Tableau 4.6**. *Scores des métriques automatiques pour les résultats de la seconde campagne, après adaptation (pourcentages ou valeurs absolues, écarts-types et classements). Les deux métriques expérimentales figurent en italique. Les écarts-types ne sont pas significatifs pour le sens arabe/français car aucune comparaison n'est possible avec un seul système.*

Les scores calculés par les métriques automatiques à base de n-grammes reproduisent à quelques exceptions près le classement obtenu par les juges humains (de manière bien plus coûteuse) pour le sens anglais/français. L'avance du système S12, classé premier, est encore plus marquée, alors que les scores du groupe S9-S10-S11 sont plus hétérogènes, et plus proches de ceux de S8. Ainsi S8 apparaît comme favorisé par les métriques automatiques fondées sur les n-grammes, par rapport aux jugements humains. L'hypothèse d'une optimisation de ce système pour la métrique BLEU ne peut pas être exclue, mais une évaluation humaine complète de la sortie du système devrait d'abord confirmer la différence observée.

Les intervalles de confiance calculés par échantillonnage ne permettent en général pas de départager les quatre systèmes S8 à S11 selon les métriques automatiques à base de n-grammes. Quant aux métriques expérimentales X-score et D-score, leurs scores et classements semblent assez éloignés de ceux des autres métriques (cf. aussi Tableau 4.7) et n'apportent pas d'informations supplémentaires.

Quant au sens arabe/français, les scores du seul système participant paraissent être dans la même gamme que ceux des systèmes anglais/français, mais cette comparaison n'est pas précise car les données de référence sont différentes. Lorsque l'on compare les scores produits par les juges humains pour le système S13 avec ceux de la traduction de référence – fluidité : 3.08 *versus* 4.70 et adéquation : 2.70 *versus* 3.51 – on constate que l'adéquation du système, à savoir sa capacité à reproduire le sens des phrases source, est plus proche des performances humaines que sa fluidité. Cela peut aussi refléter un biais des juges humains, moins tolérants envers les fautes de français.

### 4.6.4. Méta-évaluation et discussion

Les corrélations entre les métriques automatiques et les juges humains, pour le sens anglais/français, augmentent considérablement pour la seconde campagne (2b), comme on peut le constater en comparant le Tableau 4.3 ci-dessus avec le Tableau 4.7 ci-dessous. La meilleure corrélation est obtenue à égalité par la métrique NIST et par la f-mesure WNMf – 0.86/0.87 avec la fluidité et 0.95 avec l'adéquation – confirmant ainsi les expériences des auteurs de cette métrique [BAB 04]. Ces chiffres sont légèrement inférieurs à ceux obtenus pour l'anglais langue cible, mais confirment toutefois que les métriques automatiques peuvent se substituer, lorsque cela est nécessaire, aux métriques humaines, beaucoup plus coûteuses. L'augmentation de la corrélation s'explique probablement par un meilleur protocole de récolte des jugements humains, ainsi que par l'utilisation de traductions de référence de meilleure qualité, en particulier de qualité plus homogène. La nature des textes utilisés, eux-mêmes plus homogènes quant à leur contenu, pourrait également expliquer la stabilité accrue des jugements de qualité humains.

| | **BLEU** | **NIST** | **WNMf** | **X-score** | **D-score** |
|---|---|---|---|---|---|
| Fluidité | 0.85 | 0.87 | 0.86 | *0.52* | *0.05* |
| Adéquation | 0.94 | 0.95 | 0.95 | *0.39* | *0.25* |

**Tableau 4.7**. *Corrélation de Pearson (échelle -1 à 1) entre les métriques automatiques et les juges humains, seconde campagne (2b), sens anglais/français.*

Les métriques expérimentales X-score et D-score demeurent assez mal corrélées avec l'adéquation et la fluidité, bien que les corrélations augmentent par rapport à la première campagne. Une explication possible pour le X-score réside dans la nature du corpus d'entraînement lui permettant d'estimer le score de fluidité, qui n'est pas un corpus du domaine de la santé. Les fréquences des relations grammaticales peuvent être différentes d'un domaine à un autre, ce qui fait que les résultats du X-score sont moins fiables sur un domaine inconnu. Il est possible aussi que ces deux

métriques mesurent des paramètres de qualité autres que ceux mesurés par les juges humains.

La seconde campagne se proposait également, de manière entièrement originale par rapport à des campagnes précédentes, de mesurer la capacité des systèmes à s'adapter rapidement à un domaine spécifique, ici celui de la santé. Le Tableau 4.8 reproduit à des fins de comparaison les scores des métriques automatiques pour les traductions produites par les systèmes avant (2a) et respectivement après (2b) l'adaptation au domaine. Cette comparaison illustre les difficultés à s'adapter correctement à un domaine de spécialité. Si la plupart des scores augmentent bien après l'adaptation, pour les systèmes S9, S10 et S11, l'amélioration reste toutefois faible, ce qui peut être dû à un niveau de qualité initial déjà élevé, ou bien par la difficulté d'adapter le système, eu égard à la faible taille des données d'adaptation ainsi qu'aux délais relativement brefs accordés pour l'adaptation.

En revanche, le système S11, et dans une moindre mesure S8, parviennent à s'adapter au domaine avec une grande efficacité. Le cas de S11, passant de la dernière à la première place, est en ce sens particulièrement édifiant, et permet de penser que ce système peut atteindre d'excellentes performances si le domaine d'application a été bien appréhendé initialement.

| **Sys.** | **BLEU (%)** | | **NIST** | | **WNMf (%)** | | **X-score** | | **D-score** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | avant | après | avant | après | avant | après | avant | après | avant | après |
| S8 | 32.83 | 33.04 | 7.76 | 8.35 | 48.09 | 50.05 | *34.91* | *35.58* | *42.56* | *41.52* |
| S9 | 37.96 | 38.07 | 9.14 | 9.13 | 51.37 | 51.50 | *36.68* | *36.71* | *44.02* | *44.06* |
| S10 | 33.80 | 36.60 | 8.58 | 8.97 | 50.02 | 52.47 | *38.57* | *38.50* | *44.13* | *44.06* |
| S11 | 35.19 | 35.74 | 8.71 | 8.77 | 49.79 | 50.59 | *37.86* | *38.15* | *46.61* | *46.16* |
| S12 | 25.61 | 40.43 | 7.38 | 9.27 | 48.06 | 56.25 | *34.60* | *37.65* | *40.67* | *42.20* |

**Tableau 4.8**. *Comparaison des scores des métriques automatiques pour les résultats de la seconde campagne avant et après adaptation au domaine, pour le sens anglais/français.*

La différence entre les données de référence pour la première et la seconde campagne ne permet pas d'en comparer directement les résultats, d'autant que les systèmes participants n'étaient pas exactement les mêmes. Les enseignements globaux ne peuvent donc être tirés qu'au niveau de la méta-évaluation des métriques, à savoir des jugements de fiabilité qui permettent de mieux comprendre leurs qualités et leurs limites pour une utilisation future.

Quoique souvent utilisées dans la communauté, les mesures BLEU et NIST n'ont pas entièrement répondu aux attentes des organisateurs. Les corrélations et les

mesures de distance avec les jugements humains paraissent dans l'ensemble acceptables, mais moins bonnes que prévu. Les résultats varient quelque peu avec le nombre de n-grammes considérées, comme nous le montrons dans le rapport technique associé à la campagne. Ainsi, la métrique NIST obtient de meilleures corrélations que BLEU pour les unigrammes ou les bigrammes, tandis que la tendance est inversée pour les trigrammes et les quadrigrammes.

Enfin, la métrique WNMf obtient de meilleures corrélations que BLEU et NIST et justifie la recherche de nouvelles métriques pour améliorer BLEU. Toutefois, WNMf réagit différemment selon la traduction de référence utilisée, passant par exemple d'une corrélation avec les humains de près de 40% à près de 80% selon que la traduction de référence est, respectivement, la traduction officielle ou l'une des traductions effectuées par des agences (cette observation est tirée de la première campagne). Une utilisation de plusieurs références en faisant la moyenne des scores paraît une solution de compromis acceptable lorsque celles-ci sont disponibles, et c'est ce qui a été fait ci-dessus. En outre, les mesures mWER et mPER ont également été testées, et leurs corrélations avec les jugements humains sont similaires avec BLEU/NIST – les résultats figurent dans le rapport technique CESTA.

Les mesures expérimentales X-score et le D-score ne semblent actuellement pas en mesure de remplacer les métriques utilisant les n-grammes, car les corrélations sur l'ensemble des deux campagnes sont insuffisantes, et présentent de surcroît une grande variation. Bien qu'elles ne répondent pas aux attentes initiales du projet CESTA, ces mesures ont toutefois permis de mettre en lumière les difficultés à mettre en place des mesures d'évaluation fondées sur la syntaxe et la sémantique, et ont également permis de considérer des directions d'étude prometteuses.

### 4.7. Conclusions et perspectives

Les deux campagnes CESTA ont été riches en enseignements, comme l'attestent aussi les nombreux chiffres contenus dans le rapport final. Un protocole d'évaluation a été mis en place, pour réaliser à la fois des évaluations automatiques et tenter de confirmer leurs indications par des évaluations humaines, qui demeurent la référence de la capacité de traduction. Les évaluations humaines se sont révélées assez coûteuses, financièrement et temporellement, confirmant l'utilité d'une évaluation automatique, moins coûteuse des deux points de vue.

Il est toutefois nécessaire de relativiser les résultats automatiques et de prendre en compte leurs limites avant de prétendre remplacer entièrement l'évaluation humaine. CESTA a montré ainsi que la subjectivité relative des jugements humains pouvait être atténuée par la multiplication des évaluations sur une même phrase. On observe ainsi sur les deux campagnes que les juges humains produisent des

évaluations tout à fait cohérente, l'accord strict sur la note étant juste en dessous de 50% (pour des notes de 1 à 5), mais la similarité entre les jugements étant bien meilleure.

En ce qui concerne les résultats des systèmes de traduction automatique, deux systèmes ressortent de la première campagne alors qu'un système est clairement moins performant. Les deux systèmes restants ont des résultats assez similaires, tout en restant proches des deux premiers. En observant également les scores humains, la seconde campagne met également en place une certaine hiérarchie, un système étant clairement au-dessus des autres systèmes, suivi par un groupe de trois systèmes ayant des résultats très proches, et enfin par un dernier système ayant des performances moins importantes. Trois systèmes (S8, S10 et S12) tirent le meilleur parti des données d'adaptation, et enregistrent une progression significative entre les deux tests de la seconde campagne.

Les protocoles d'évaluation mis en place par CESTA – données et métriques notamment – seront désormais disponibles pour la communauté des développeurs et des utilisateurs de systèmes de traduction automatique. Ces ressources nous paraissent constituer ainsi une première pour les directions anglais/français et arabe/français. Les ressources sont distribuées par l'ELRA sur support CD-ROM, accompagnées de documentations et des rapports préservant l'anonymat. Les deux métriques expérimentales, X-score et D-score, demandent encore à être étudiées en vue de leur validation. Les résultats de CESTA ont été discutés lors de l'atelier de clôture et présentés à l'ensemble des participants de la plateforme EVALDA du programme Technolangue.

## 4.8. Remerciements

Nous voudrions remercier chaleureusement tous les systèmes participants aux campagnes CESTA, qui sont décrits dans la section 4.3, et notamment leurs représentants qui ont bien voulu participer aux réunions CESTA et les développeurs qui ont adapté les systèmes aux spécifications de la campagne CESTA. Nous remercions également tous les évaluateurs humains pour leur travail.

## 4.9. Bibliographie

[BAB 04] BABYCH B., HARTLEY A., « Extending the BLEU MT Evaluation Method with Frequency Weightings », *42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelone, p. 622-629, 2004.

[BAB 05] BABYCH B., HARTLEY A., ELLIOTT D., "Estimating the predictive power of n-gram evaluation metrics across languages and text types", *MT Summit X*, Phuket, Thaïlande, p. 412-418, 2005.

[DAB 04] DABBADIE M., MUSTAFA EL HADI W., TIMIMI I., "CESTA: The European MT Evaluation Campaign", *Multilingual Computing and Technology*, vol. 15, n° 5, p. 10-12, 2004.

[DOD 02] DODDINGTON G., "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", *Human Language Technology Conference (HLT 2002)*, San Diego, 2002.

[HAM 50] HAMMING R., "Error-detecting and error-correcting codes", *Bell System Technical Journal*, vol. 29, n° 2, p. 147-160, 1950.

[HAM 06a] HAMON O., POPESCU-BELIS A., CHOUKRI K., DABADIE M., HARTLEY A., MUSTAFA EL HADI W., RAJMAN M., "CESTA: First Conclusions of the Technolangue MT Evaluation Campaign", *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Gênes, Italie, p. 179-184, 2006.

[HAM 06b] HAMON O., RAJMAN M., "X-Score: Automatic Evaluation of Machine Translation Grammaticality", *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Gênes, Italie, 2006.

[HAS 06] Hasan S., El Isbihani A., Ney H., "Creating a Large-Scale Arabic to French Statistical Machine Translation System", *LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Gênes, Italie, 2006.

[HOV 02] HOVY E.H., KING M., POPESCU-BELIS A., "Principles of Context-Based Machine Translation Evaluation", *Machine Translation*, vol. 17, n° 1, p. 1-33, 2002.

[ISO/IEC 01] ISO/IEC, *ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model*, Geneva, International Organization for Standardization / International Electrotechnical Commission, 2001.

[LAN 06], LANGLAIS P., GOTTI F., PATRY A., "De la Chambre des communes à la chambre d'isolement : adaptabilité d'un système de traduction basé sur les segments de phrases", *TALN 2006*, Leuven, Belgique, 2006.

[MAR 06] MARINO J.B., BANCHS R., CREGO J.M., DE GISPERT A., LAMBERT P., FONOLLOSA J.A.R., COSTA-JUSSA M.R., "N-gram Based Machine Translation", *Computational Linguistics*, vol. 32, n° 4, 2006.

[MUS 02] MUSTAFA EL HADI W., TIMIMI I., DABBADIE M., "Terminological Enrichment for non-Interactive MT Evaluation", *LREC 2002 (Third International Conference on Language Resources and Evaluation)*, Las Palmas, Espagne, p. 1878-1884, 2002.

[MUS 04] MUSTAFA EL HADI W., DABBADIE M., TIMIMI I., RAJMAN M., LANGLAIS P., HARTLEY A., POPESCU-BELIS A., "CESTA: Machine Translation Evaluation Campaign", *Coling 2004 Workshop on Language Resources for Translation Work, Research and Training*, Genève, p. 8-17, 2004.

[NIS 03] *The 2004 NIST Machine Translation Evaluation Plan* (MT-04), 24th Dec. 2003, http://www.nist.gov/speech/tests/mt/doc/mt04_evalplan.v2.1.pdf.

[PAP 01] PAPINENI K., ROUKOS S., WARD T., ZHU W.-J., BLEU: a Method for Automatic Evaluation of Machine Translation, Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022), 2001.

[PAT 05] PATRY A., LANGLAIS, P. "Corpus-Based Terminology Extraction", *7th International Conference on Terminology and Knowledge Engineering*, Copenhague, p. 313-321, 2005.

[RAJ 01] RAJMAN M., HARTLEY A., "Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores", *Workshop on MT Evaluation at MT Summit VIII*, Compostelle, Spain, p. 29-34, 2001.

[RAJ 02] RAJMAN M., HARTLEY T., "Automatic ranking of MT systems", *LREC 2002*, Las Palmas, Espagne, p. 1247-1253, 2002.

[SUR 05] SURCIN S., HAMON O., HARTLEY A., RAJMAN M., POPESCU-BELIS A., MUSTAFA EL HADI W., TIMIMI I., DABBADIE M., CHOUKRI K., "Evaluation of Machine Translation with Predictive Metrics beyond BLEU/NIST: CESTA Evaluation Campaign #1", *Machine Translation Summit X*, Phuket, Thaïland, p. 117-124, 2005.

[THO 94] THOMPSON H., BREW C., "Automatic Evaluation of Computer Generated Text", *ARPA/ISTO Workshop on Human Language Technology*, p. 104-109, 1994.

[WHI 94] WHITE J.S., O'CONNELL T.A., "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches", *AMTA Conference, 5-8 October 1994*, Columbia, MD, USA, 1994.

[WHI 01] WHITE J.S., "Predicting Intelligibility from Fidelity in MT Evaluation", *Workshop on MT Evaluation at Mt Summit VIII*, Compostelle, Espagne, 2001.

[ZHA 04] ZHANG Y., VOGEL S., WAIBEL A., "Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System?" *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, Lisbonne, p. 2051-2054, 2004.