Faculté des arts et des sciences
Département d'informatique
et de recherche opérationnelle

ÉVALUATION DU PROTOTYPE DE TRADUCTION DES AVERTISSEMENTS MÉTÉO

Fabrizio Gotti – gottif@iro.umontreal.ca
Guy Lapalme – lapalme@iro.umontreal.ca

17 JANVIER 2011

RALI

Recherche Appliquée en
Linguistique Informatique

Université de Montréal

Résumé

Ce rapport décrit l'évaluation effectuée en octobre 2010 du prototype de traduction automatique d'alertes météo (appelé WATT) produit par le RALI dans le cadre du projet de collaboration *Diffusion multiformat d'informations environnementales*. Cette évaluation en suit une autre, complétée par le Bureau de la traduction (BT) en juillet 2010, très critique à l'endroit de WATT.

Malheureusement, cette étude du BT repose sur une méthodologie discutable : les bulletins à traduire ne sont pas représentatifs de ceux habituellement émis, le repérage des erreurs ne s'est pas fait à l'aveugle, et ne tient pas compte des erreurs dans l'entrée (les phrases source à traduire). La présente étude cherche à réparer ces lacunes méthodologiques et à identifier les forces et faiblesses de WATT. À la lumière de ces informations, on propose un positionnement de WATT dans la chaîne de diffusion d'EC.

L'évaluation du RALI s'est faite à l'aveugle, par 5 annotateurs humains à qui l'on a demandé d'annoter les phrases source, les traductions humaines et machine de 32 bulletins d'alerte (101 phrases) tirés au hasard parmi ceux émis durant l'été 2010. L'analyse exhaustive de ces annotations révèle les points suivants :

- Les traductions humaines sont de meilleure qualité que les traductions machine non révisées. 66 % des bulletins traduits par WATT contiennent au moins une erreur de fidélité dans la traduction, comparé à 25 % pour l'humain. Notre étude est cependant moins critique à l'égard de WATT que celle du BT.
- 25 % des phrases source (l'intrant) contiennent au moins une erreur commise lors de la saisie du bulletin. Ces erreurs sont surtout de la ponctuation erronée et des noms de lieux mal écrits (casse, accents). On note aussi des problèmes dans la saisie des unités de mesure, de temps et quelques fautes d'orthographe.
- Lorsque les phrases source contiennent des erreurs, la qualité des traductions (humaines ou machine) en souffre. Normaliser les sources serait donc fructueux.
- Il est envisageable, en quelques mois, d'éliminer 51 % des erreurs de traduction commises par WATT, selon l'analyse des erreurs commises par ce dernier dans cette évaluation. Les erreurs majeures sont toutefois les plus difficiles à atténuer.

Suite à ces constatations, le RALI fait les recommandations suivantes :

1. Faire de WATT *non pas* un remplacement du traducteur humain, mais plutôt un **système de traduction quasi instantanée** (5 secondes en moyenne par bulletin) des informations dont l'urgence de diffusion est grande, quitte à ce que ces traductions soient révisées par la suite par un expert humain. Dans l'intervalle, le consommateur est avisé que ce qu'il lit est traduit automatiquement et qu'une révision est imminente. Un gain en productivité est ainsi possible.
2. Investir quelques mois pour **éliminer de WATT certaines erreurs** identifiées ici.
3. Assainir les entrées de la chaîne de diffusion en concevant un **système d'aide à la saisie des bulletins**, destiné au météorologue. Ce système filtrerait les mots mal orthographiés ou inconnus, uniformiserait la notation et le vocabulaire et permettrait donc des phrases source ainsi que des traductions de meilleure qualité.

Executive Summary

This report describes an evaluation performed in October 2010 of RALI's weather warning automated translation prototype (henceforth WATT), as part of the co-operation project *Multi-format Environmental Information Dissemination*. This evaluation follows a previous one, led by the Translation Bureau (TB) in July 2010, which was very critical of WATT.

Unfortunately, the TB's study was based on a questionable methodology: Translated weather warnings were not representative of those usually broadcast by EC, error identification was not blinded, and never takes into account the presence of errors in the source text (the inputs). Our study attempts to correct these flaws in methodology and to identify WATT's strengths and shortcomings. In light of this information, we then recommend an optimal positioning for WATT in EC's information dissemination pipeline.

RALI's blinded evaluation was performed by 5 human annotators (raters) who were asked to annotate source language sentences, as well as their respective human and machine translations, for 32 weather warnings (101 sentences) randomly selected from those issued during the summer of 2010. An exhaustive analysis of these annotations reveals that:

- Human translations are of a better quality than the automated, non-revised translations. 66 % of bulletins produced by WATT contain at least one fidelity flaw in their translation, compared to 25 % for human translations. However, our study is less critical of WATT than that produced by the Translation Bureau.
- 25 % of source sentences (inputs) contain at least one error, made when the bulletin was written. These errors pertain mostly to punctuation and misspelled place names (case, accents). We also note problems affecting units of measurement, of time, as well as a few orthographic mistakes and typos.
- When a source sentence contains errors, the quality of its translations (machine or human) will suffer. Cleaning up the sources would therefore help translation.
- It would be possible, within a few months, to suppress 51 % of all translation mistakes made by WATT, according to the error analysis performed in this study. However, major errors are the most difficult to eliminate.

Following these observations, RALI makes the following recommendations:

1. Implement WATT *not* as a replacement for human translators, but as an **almost instantaneous translation system** (5 seconds per bulletin) for information urgently needed, even if it means that the translation would be revised by a human expert afterwards. In the meantime, the consumer is advised that the bulletin he's reading has been translated automatically, and that a revision is pending.
2. Put a few month's worth of effort in **eliminating certain errors from WATT**.
3. Clean up the inputs of the information dissemination pipeline by designing a **computer-aided bulletin creation tool**, meant to help the meteorologist. Such a system would filter out misspelled or unknown words, standardize the notation, the vocabulary, and would allow for better quality source text and translations.

Remerciements

Nous tenons à remercier Marie-Claude L'Homme et Patrick Drouin, professeurs au département de linguistique et traduction de l'Université de Montréal, qui ont permis à leurs assistants de recherche de participer à cette évaluation.

Ce travail d'analyse n'aurait pas été possible sans la minutie des cinq annotateurs qui ont révisé les traductions : Gabriel Bernier-Colborne, Suzanne DesGroseilliers, Mélanie Gagnon, Annaïch Le Serrec et Amélie Paulus.

Table des matières

1 Introduction

Ce rapport décrit une évaluation effectuée en octobre 2010 du prototype de traduction d'alertes météo produit par le RALI dans le cadre du projet de collaboration *Diffusion multiformat d'informations environnementales*[1]. Ce projet de recherche, en cours, est financé par une contribution d'Environnement Canada (EC) depuis l'automne 2008 et est appuyé par MITACS depuis le printemps 2009.

Ce système de traduction automatique, que nous désignerons ici WATT (Warning-Alerte Translation-Traduction), est le résultat d'un travail de plusieurs années et de versions préliminaires ayant déjà fait l'objet de plusieurs publications scientifiques [1,2] depuis 2004. Le système s'appuie sur le moteur de traduction statistique Moses[2] et sur un corpus de plus de 40 millions de mots par langue extraits d'archives de bulletins météo et d'alertes qui ont été assemblées depuis près de 7 ans.

Les principes sous-jacents de WATT et une démonstration de celui-ci ont été présentés à des gestionnaires d'Environnement Canada le 7 janvier 2010. Ceux-ci s'étaient montrés très intéressés à intégrer à terme ce système dans leur environnement de production d'alertes météo pour en accélérer le processus de traduction.

WATT est disponible dans l'environnement Linux sous la forme d'un programme exécutable accessible *via* une interface « ligne de commande » ou *via* un script en langage Python. Une version de démonstration, basée sur ce programme, est aussi accessible sur le web à

http://rali.iro.umontreal.ca/EnvironmentalInfo/WarningTranslation.html

Cette version de démonstration a été utilisée épisodiquement par le personnel d'EC pour en évaluer le potentiel d'amélioration de la rapidité de diffusion des alertes météorologiques. Le processus *actuel* de traduction d'une alerte nécessitant environ 30 minutes, il retarde d'autant la diffusion des alertes dans les deux langues.

Suite aux résultats encourageants constatés lors d'essais informels, EC voulait explorer la possibilité d'une production rapide d'une première traduction machine des alertes qui serait immédiatement disponible (WATT nécessite environ 5 secondes pour traduire un bulletin), accompagnée d'une note indiquant qu'une révision humaine sera effectuée sous peu. La version révisée remplacerait cette version machine préliminaire et deviendrait la version officielle.

Au printemps 2010, M. Gaétan Deaudelin, Directeur régional – Opérations du SMC région du Québec, s'est adressé au Bureau de la traduction afin d'obtenir une évaluation de la qualité des sorties de WATT. Au début de juillet 2010, le BT a produit un rapport [3] qui a été transmis au RALI à la fin d'août 2010.

Ce rapport est le résultat d'un travail très minutieux d'annotation, de classification et de comptabilisation d'erreurs sur 135 bulletins rédigés expressément pour cette tâche. Ces bulletins ont été traduits, à la fin mai 2010, à l'aide de la version de démonstration

[1] http://rali.iro.umontreal.ca/EnvironmentalInfo/index.fr.html
[2] http://www.statmt.org/moses/

disponible sur Internet, probablement en utilisant une série d'opérations de copier-coller entre le document source et la traduction produite sur le site web.

Les conclusions du rapport sont très critiques:

> *65,19 % des échantillons traduits (88 des 135 échantillons jugés représentatifs par EC) présentent une ou plusieurs erreurs graves, dont 11,11 % (15 des 135 échantillons jugés représentatifs par EC) présentent au moins 5 erreurs graves.*

Malheureusement, le rapport repose sur une méthodologie d'analyse discutable :

- Les bulletins utilisés pour l'évaluation ont été créés de toutes pièces pour cette évaluation et aucune étude n'a été faite pour juger de leur similarité avec les bulletins habituellement émis par EC, même si on affirme qu'ils ont été jugés « représentatifs » par EC. Ces bulletins artificiels ne sont pas donnés dans le rapport quoiqu'on relève, en annexe 5, quelques « mauvais exemples » de traduction sans qualifier si elles se produisent souvent ou bien si ce sont des « perles ». Il serait probablement utile de construire un « palmarès » analogue avec certaines traductions humaines. À en juger par les bulletins de la période estivale, qui ont été par la suite fournis, ils comportent souvent des formulations différentes des bulletins habituels, ce qui évidemment biaise les résultats. Il aurait été préférable d'utiliser des bulletins réellement diffusés au cours des mois précédents. Le système utilisé pour les tests ayant été développé il y a plus d'un an, il n'y avait aucun risque de *surentraînement*. Il aurait alors été possible de comparer les sorties de la machine avec les traductions publiées.
- La comparaison n'a pas été faite *à l'aveugle* en comparant le taux d'erreurs similaires qui auraient pu être détectées dans des traductions humaines, l'évaluateur ne sachant pas s'il évalue une traduction humaine ou machine.
- L'évaluation n'a porté que sur la traduction de l'anglais vers le français alors que le système est bilingue et peut même déterminer automatiquement la langue de la source.
- L'évaluation ne tient pas compte des erreurs dans l'entrée. Le rapport se limite à une liste de tableaux d'erreurs sans donner les textes ni d'entrée ni de sortie. Il est donc difficile d'apprécier la qualité de l'annotation ou de tirer des leçons de cette évaluation pour améliorer le système. Pour la période d'été, certains bulletins accompagnent les statistiques, mais sans identifier l'endroit où l'annotateur a considéré qu'il avait détecté une erreur ni le type de l'erreur. Seuls certains bulletins jugés inacceptables ont été surlignés dans leur entièreté. Plusieurs erreurs d'entrée (fautes d'orthographe, de ponctuation ou problèmes de majuscules/minuscules) ont pu être constatées dans ces quelques bulletins donnés en entrée au système. Ces erreurs d'entrée sont inévitables étant donné le processus actuel de rédaction, mais il aurait été important d'en tenir compte dans l'évaluation de la sortie.
- L'évaluation des risques ne tient pas compte du temps de traduction qui peut retarder la publication d'une alerte. Il aurait donc fallu comparer le risque entre, d'une part, la diffusion d'un bulletin en langue source réputé sans erreur et de sa

traduction machine comportant une erreur mais restant tout même compréhensible avec, d'autre part, le risque qu'aucun des deux bulletins ne soit diffusé à temps pour avertir la population.

- Le rapport ne précise pas le processus de calcul des erreurs : où ont-elles été détectées ? Plusieurs erreurs ont-elles été attribuées à un même mot ? Les annotations ont-elles été effectuées par une seule personne ? Ont-elles été corroborées par une seconde personne ? L'annotateur savait-il qu'il traitait des traductions machine ?

Étant donné ces lacunes et le peu d'informations pouvant être tirées du rapport du BT, M. Gaétan Deaudelin a demandé, en septembre 2010, de proposer une nouvelle évaluation de WATT, cette fois en utilisant une méthodologie *état de l'art*.

Le RALI ne prétend pas disposer d'un système parfait, loin de là, mais le but de l'évaluation est de chercher à l'améliorer. Cette évaluation a été conçue pour répondre aussi objectivement que possible à la question de savoir si les sorties du système sont utiles pour la population concernée par les alertes météorologiques et pour nous permettre de cibler les principales lacunes pour y remédier.

La section suivante décrit le processus utilisé pour l'évaluation. La section 3 décrit les principaux résultats obtenus et offre des suggestions d'amélioration que le RALI pourrait apporter à WATT dans les prochains mois. La conclusion résume les résultats obtenus et offre notre opinion sur un positionnement éventuel de WATT dans la chaîne de diffusion d'EC.

2 Méthodologie

2.1 Objectifs et principes de l'évaluation

L'évaluation des sorties d'un système de traduction automatique est un problème complexe et demeure une question ouverte. Il existe des métriques automatiques, telles que BLEU [4], mais elles souffrent de divers problèmes de fiabilité [5]. Si l'on désire mesurer la qualité véritable des traductions en termes d'utilité pour un être humain, alors une évaluation humaine est indispensable [6].

Le RALI a donc mis en place une évaluation humaine du prototype en octobre 2010. Elle a consisté à examiner les sorties du prototype, lorsqu'on lui soumettait des phrases de bulletins d'avertissements météorologiques. **Aucune révision manuelle des sorties de WATT n'a été faite**. Aux fins de comparaison, des traductions humaines des mêmes phrases source ont également été évaluées.

Cette évaluation souhaitait répondre aux questions suivantes :

1. Comment se comparent les qualités des traductions machine et humaines ?
2. Quelles sont les erreurs commises par WATT, et par le traducteur humain ?
3. Le sens de la phrase du bulletin original est-il correctement transposé dans la langue cible par WATT et par le traducteur humain ?
4. Quelle est la qualité du texte source (l'intrant), et comment influe-t-elle sur la qualité des traductions ?

Le RALI a privilégié trois principes directeurs lors de cette évaluation :

- L'évaluation s'est faite *à l'aveugle* (ou de façon anonyme), c'est-à-dire que les annotateurs ne connaissaient pas quels bulletins ni quel système de traduction (humain ou machine) ils examinaient. De cette façon, on évite qu'un éventuel parti pris pour la machine ou pour l'humain nuise à l'évaluation des traductions. Cet aspect d'évaluation à l'aveugle est une composante cruciale de toute évaluation de ce genre et favorise l'impartialité. Les évaluateurs sont des personnes qui n'ont jamais été impliquées dans le développement du système.
- L'évaluation est *représentative des conditions réelles* de fonctionnement de la chaîne de diffusion des informations météorologiques. Nous avons donc tâché de sélectionner des bulletins réels, aléatoirement, afin d'échantillonner de façon représentative les bulletins qui seraient éventuellement soumis au système de traduction.
- L'évaluation permet notamment la mesure de la *qualité extrinsèque* de la traduction des bulletins produits. La qualité extrinsèque est une métrique indiquant dans quelle mesure un système de traduction est utile à une tâche concrète donnée, en l'occurrence la diffusion d'informations météorologiques urgentes dans une langue de qualité, et dans un temps raisonnable. Son calcul s'appuie sur une interprétation de haut niveau des annotations et des impressions globales des annotateurs sur les bulletins produits.

Les sections qui suivent décrivent le protocole expérimental utilisé pour répondre aux questions soulevées, dans le respect des principes énoncés.

2.2 Contexte expérimental

Même si cette évaluation porte sur des bulletins d'alerte complets, nous avons jugé utile d'en annoter les phrases séparément. En effet, un examen préalable des traductions humaines et machine d'un bulletin complet montre qu'un utilisateur moyen y trouve une ou deux phrases qui lui mettent la puce à l'oreille quant à l'origine de la traduction. Par conséquent, ses annotations sur un bulletin complet ne se feraient plus véritablement à l'aveugle, car il serait renseigné sur le système qui l'a produit.

L'évaluation porte donc sur des paires de phrases. La source correspond à une phrase d'un bulletin d'alerte, la cible est une traduction. Celle-ci peut avoir été produite soit par WATT (« traduction machine »), soit par l'humain (« traduction humaine »).

Le fait de colliger des statistiques sur les phrases n'empêche pas l'évaluation de bulletins complets : nous montrons plus bas comment nous agrégeons ces statistiques pour évaluer les bulletins d'alerte du corpus de test.

2.2.1 Corpus de test et échantillonnage aléatoire des bulletins

Afin que cette évaluation soit représentative des conditions réelles de traduction des bulletins, les paires de phrases évaluées ont été extraites de réels bulletins d'alerte météorologique. Ces bulletins nous ont été transmis par Miguel Tremblay, coordonnateur national des services commerciaux de données pour Environnement Canada. Ils concernent les mois de juin, juillet et août 2010, et correspondent à la forme finale du bulletin tel que diffusé sur le site Web d'Environnement Canada[3]. Le texte est donc accentué et la casse a été rétablie (format de fichier `ACC-MP`), contrairement au format intermédiaire en toutes majuscules utilisé ailleurs dans la chaîne de diffusion et d'archivage d'Environnement Canada. Le RALI tient donc à souligner l'efficace collaboration de Miguel Tremblay qui nous a fourni ce corpus. Cet effort est d'autant plus apprécié qu'il n'existe aucunes archives officielles de ces textes diffusés sur le Web.

Nous avons utilisé un corpus d'été, car c'est la saison où les avertissements urgents sont les plus nombreux, et pour laquelle EC juge que le besoin de diffusion rapide est très grand. Par ailleurs, il nous aurait été difficile d'obtenir des bulletins pour les autres saisons, étant donné le manque de données disponibles.

Le choix de ce corpus permet de considérer la justesse de la restauration des accents et de la casse des bulletins lors de l'examen des bulletins. De plus, les bulletins évalués ici sont identiques à ceux que les internautes ont pu lire sur le site d'Environnement Canada, ce qui permet une évaluation du « produit fini » de la chaîne de diffusion des informations météorologiques.

Le corpus complet comptait 20 230 bulletins (soit environ 10 000 dans chaque langue officielle), et couvrait les 10 provinces et les 3 territoires du pays. Le RALI a dépouillé les bulletins, en a apparié les versions françaises et anglaises, puis en a finalement extrait le texte des « discussions », c'est-à-dire les sections étiquetées `EVENT_DISCUSSION` et

[3] http://www.meteo.gc.ca/warnings/warnings_f.html

`OMNI_DISCUSSION`. Cette section des bulletins est typiquement rédigée dans un texte libre, donc présentant un défi de traduction plus grand que d'autres sections au texte stéréotypé (p.ex. `Environnement Canada continue à surveiller la situation.`), facilement traduit par une mémoire de traduction.

Le texte de chacune de ces discussions a été segmenté en phrases, puis le RALI a apparié chaque phrase anglaise à sa traduction française. Ce travail s'est révélé fort complexe, du fait du grand nombre d'étapes nécessaires, et à cause du format quelque peu obscur de ces bulletins. Un exemple de paires de bulletins appariés est présenté à l'annexe A, de même que certains des procédés qui ont permis l'appariement.

Nous avons sélectionné aléatoirement 32 paires de bulletins en tâchant de représenter équitablement chaque mois échantillonné (juin, juillet et août) et chacun des 13 territoires et provinces, lorsque le nombre de bulletins disponibles le permettait. Nous avons considéré que la langue source est le français pour les avertissements issus du Québec, et l'anglais pour les autres. Toutes les paires de phrases de ces bulletins tirés au hasard ont ensuite servi à l'évaluation. Le tableau 2.1 montre la répartition des paires de bulletins et paires de phrases, par province ou territoire. On y constate notamment qu'un bulletin compte en moyenne 3 phrases de discussion.

Tableau 2.1
Nombre de paires de bulletins et de phrases sélectionnés aléatoirement, par province et territoire.

Province ou territoire	Nb. bulletins archivés	Nb. bulletins évalués	Nb. phrases évaluées
Alberta	1234	3	10
Colombie-Britannique	101	3	9
Île-du-Prince-Édouard	13	1	3
Manitoba	1071	3	11
Nouveau-Brunswick	103	3	6
Nouvelle-Écosse	41	2	4
Nunavut	9	3	10
Ontario	1188	3	9
Québec	1388	3	13
Saskatchewan	1902	3	12
Terre-Neuve	87	1	3
Territoires du Nord-Ouest	19	2	7
Yukon	35	2	4
Total	**7191**	**32**	**101**

Ces 101 phrases sources ont été appariées d'abord avec leur traduction humaine (repérée dans les bulletins) et ensuite à leur traduction machine, produite par WATT. Ceci donne donc un total de 101 × 2 = 202 paires de phrases à évaluer.

La traduction automatique des 101 phrases source (1887 mots) a pris 147 s sur un PC typique, soit une seconde et demie par phrase (13 mots à la seconde), en moyenne. Ceci correspond à 5 s en moyenne pour la traduction d'un bulletin complet.

2.2.2 Mesure et comparaison de représentativité des bulletins

Étant donné que la représentativité des bulletins utilisés dans cette évaluation fait partie des principes directeurs de celle-ci, nous présentons ici les résultats de mesure de similarité entre, d'une part, les bulletins météo habituellement émis par EC et, d'autre part, ceux utilisés dans notre évaluation.

La métrique statistique habituellement utilisée en informatique linguistique pour juger de la similarité entre deux corpus est appelée *perplexité*. Cette perplexité repose sur la construction d'un modèle de langue statistique à partir de bulletins et avertissements météo émis par EC entre 2004 et 2009 (c'est le corpus d'entraînement). Pour obtenir la perplexité sur un nouveau corpus, le modèle de langue est ensuite interrogé pour déterminer dans quelle mesure ce nouveau corpus ressemble au corpus d'entraînement.

Plus la perplexité est basse, plus le nouveau corpus ressemble au corpus d'entraînement. En d'autres termes, plus la perplexité est basse, plus le nouveau corpus est représentatif du corpus d'entraînement et, dans ce cas, du langage habituellement utilisé dans les bulletins et avertissements d'Environnement Canada.

Une autre métrique simple de représentativité est le pourcentage de mots inconnus par le modèle de langue. Cette métrique mesure le pourcentage de mots jamais rencontrés auparavant parmi les mots du nouveau corpus. Plus le pourcentage de mots inconnus est faible, plus le nouveau corpus est représentatif du corpus d'entraînement.

On a calculé ces deux métriques (perplexité et pourcentage de mots inconnus) entre le corpus d'entraînement et notre corpus d'évaluation. De plus, nous avons calculé ces deux métriques entre notre corpus d'entraînement et le corpus d'évaluation du BT, tel que mentionné dans leur rapport d'évaluation [3], qui contenait 135 bulletins élaborés afin que le BT puisse mener cette évaluation. Nous avons récupéré le texte intégral de ces bulletins soumis dans le journal (*log*) informatisé de WATT, pour la journée du 26 mai 2010, date à laquelle les bulletins ont été soumis à notre moteur *via* le Web. Nous avons travaillé avec la suite logicielle SRILM [10] pour gérer le modèle de langue. Les résultats sont présentés au tableau 2.2.

Tableau 2.2

Mesures de représentativité de deux corpus d'évaluation, par rapport aux bulletins météo habituellement émis par EC. Plus le pourcentage de mots inconnus est bas, plus le corpus d'évaluation est représentatif. Plus la perplexité est basse, plus le corpus d'évaluation est représentatif. La perplexité affichée ici tient compte du nombre de mots différents entre corpus d'évaluation.

Corpus	Nb. de mots	% de mots inconnus	Perplexité
Corpus d'évaluation du RALI	1571	3,4 %	27,6
Corpus d'évaluation du BT	12 018	5,1 %	38,1

On constate que :

- La proportion de mots inconnus est plus élevée dans le corpus d'évaluation du Bureau de la traduction.
- Les deux corpus ont une perplexité relativement faible. En effet, par expérience sur des corpus de nature différente, ce chiffre peut atteindre un ordre de grandeur supérieur. La faible perplexité relative s'explique sans doute par la taille restreinte du vocabulaire dans le domaine des bulletins météo.
- La perplexité est plus élevée pour le corpus du BT, par rapport à celle sur le corpus du RALI.

La représentativité du corpus d'évaluation du BT est donc quelque peu inférieure à celle de celui du corpus du RALI, qui est un tirage au hasard parmi les bulletins émis récemment par Environnement Canada. En d'autres termes, l'élaboration manuelle d'une prose « représentative » des bulletins d'EC est un exercice plus périlleux qu'il n'y paraît. Les bulletins créés de toutes pièces par le Bureau sont ainsi plus difficiles à traduire par WATT que ne le seraient des bulletins réels émis par EC.

2.2.3 Déroulement de l'évaluation

Cinq annotateurs bilingues détenant une formation universitaire en linguistique ont été invités à lire attentivement chaque paire de phrases, et à indiquer les erreurs relevées dans le texte de la source et de la traduction. Chacune des 202 paires de phrases a été ainsi annotée. Conformément à l'état de l'art, nous avons fait évaluer chaque paire de phrases deux fois, par deux annotateurs différents. Ceci est essentiel pour limiter l'effet du biais personnel d'un annotateur donné. Puisque chacune des 202 paires a été annotée deux fois, on a en fin de compte $202 \times 2 = 404$ paires annotées.

Afin d'obtenir une évaluation aussi impartiale que possible, les annotateurs ignoraient si la phrase traduite qu'ils évaluaient était une traduction machine ou humaine. Pour compléter le processus d'anonymisation de l'évaluation, les paires de phrases ont été mises en désordre avant d'être soumises aux annotateurs.

2.2.4 Typologie d'erreurs

La typologie d'erreurs désigne l'ensemble des erreurs que nous demandions aux annotateurs d'utiliser afin d'évaluer la qualité des phrases source (l'intrant) et des traductions.

Cette typologie s'inspire fortement de celle utilisée par le Bureau de la traduction du Canada, décrite dans leur rapport [3]. Cette liste permet de mesurer la qualité linguistique et se base sur les principes fondamentaux du Système canadien d'appréciation de la qualité linguistique (SICAL).

Nous avons regroupé certains types d'erreur afin de rassembler sous la même rubrique des erreurs très semblables ou difficiles à différencier à partir de leur définition. Le travail d'annotation s'en trouve simplifié, sans perdre en rigueur. Le tableau 2.3 montre les 7 types d'erreur que les annotateurs étaient invités à repérer, ainsi que leur définition.

Ces repères pour l'annotation ont été expliqués aux annotateurs lors d'une présentation orale, et une page Web de ressources a été créée pour leur rappeler les consignes. Cette page Web est reproduite à l'annexe B.

Tableau 2.3

Typologie d'erreurs pour l'évaluation, telle que présentée aux annotateurs. Il y a 7 types d'erreur en tout.

Erreurs sur le sens		
	Majeures	**Mineures**
Charabia	**CHA1** Passage présentant un langage, style incompréhensible ou grossièrement incorrect.	S.O.
Omission	**OMIS1** Omission d'une donnée ou d'un élément de sens important pour la compréhension du texte.	**OMIS2** Omission sans gravité ne nuisant pas à la compréhension de façon significative.
Faux-sens	**FAUX1** Ajout d'une information incorrecte. Sens contraire à l'original.	**FAUX2** Insertion d'une ambiguïté gênant un peu la compréhension. Sens différent mais voisin de l'original.
Erreurs sur la forme (linguistiques)		
LANG	Non-respect des règles d'orthographe, de grammaire et de syntaxe.	
TYPO	Non-respect des règles de casse, ponctuation et diacritiques (accents, cédilles, etc.) et césure des mots.	

En plus du repérage des erreurs expliquées au tableau 2.3, on a demandé à l'annotateur de se prononcer sur la qualité globale du transfert du sens lors de la traduction. Pour ce faire, on lui demandait de répondre par oui ou par non à la question suivante :

Cette traduction respecte-t-elle l'exactitude du sens et des données de l'original ?

Cette question mesure de façon directe et simple un critère central de la qualité de l'exécution de la traduction, soit la compréhension du message du bulletin. Cette formulation est issue du rapport du Bureau de la traduction [3]. Nous cherchons ainsi à évaluer, comme le souligne justement le Bureau de la traduction, le risque qui

> […] *réside d'une part dans une mauvaise compréhension ou une interprétation erronée du message et d'autre part dans l'obligation pour le lecteur de se référer à la version d'origine pour comprendre le message.*

Cette question concourt au calcul de la qualité extrinsèque des systèmes de traduction.

2.2.5 Interface d'annotation

L'évaluation s'est faite dans Microsoft Word, en utilisant un système intuitif de macros permettant aux annotateurs de marquer les erreurs relevées. L'utilisateur n'a qu'à sélectionner le segment de phrase où l'erreur est repérée et à cliquer sur le bouton correspondant pour annoter. Une capture d'écran du système est présentée à la figure 2.1.

2.2.6 Annotations de la phrase source

Comme expliqué plus haut (et illustré à la figure 2.1), les annotateurs ont également examiné la qualité linguistique des phrases source. La qualité de l'intrant influence de façon directe la qualité des traductions automatiques. À la figure 2.1, la phrase source contient `rivière rounge` au lieu de `rivière Rouge`. Comme le prototype ne fait pas la correction avant de traduire, il transpose cette erreur dans la traduction. Ce phénomène doit donc être quantifié, afin d'identifier équitablement les erreurs commises par le prototype, et celles attribuables à des problèmes d'intrant.

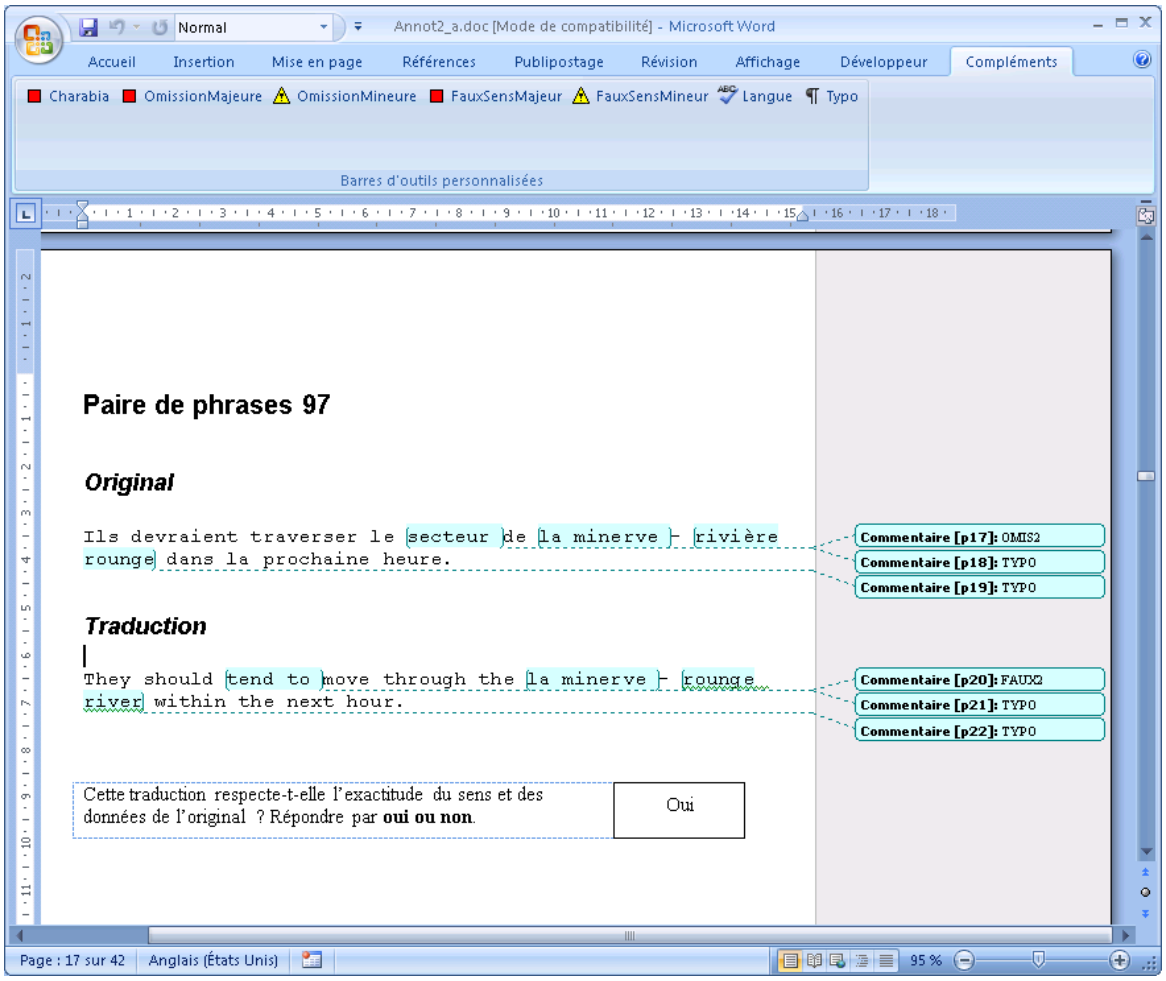


Figure 2.1

Interface d'annotation d'une des 404 paires de phrases dans Microsoft Word, avec les marques faites par un annotateur pour la paire de phrases anonyme n° 97. La phrase source est en fait extraite d'un bulletin québécois, et sa traduction est automatique, à l'insu de l'annotateur. La barre d'outils en haut de l'image permet d'ajouter les marques d'annotation. Dans cet exemple, l'annotateur a relevé des erreurs de typographie (la casse des noms de lieux) dans la phrase source. De plus, il signale que le passage `le secteur` de la phrase source n'apparaît pas dans la phrase cible (code `OMIS2`). Enfin, il relève un faux-sens mineur dans la traduction `tend to`, qui ne correspond pas exactement à la source `devraient`, selon lui. Cependant, globalement, il estime que l'exactitude du sens est respectée par la traduction (annotation finale `oui`).

3 Résultats

Nous présentons dans cette section les résultats de l'évaluation manuelle ainsi que leur analyse. On y utilise les codes d'erreurs montrés au tableau 2.3.

3.1 Quelques définitions

À plusieurs endroits dans cette section, on a jugé bon de regrouper certains résultats pour en faciliter l'interprétation. Ainsi, on définit les mesures agrégées suivantes.

- Les erreurs **mineures** sont la somme des erreurs OMIS2 et FAUX2.
- Les erreurs **majeures** sont la somme des erreurs CHA1, OMIS1 et FAUX1.
- Les erreurs **linguistiques** sont la somme des erreurs LANG et TYPO.
- Les **omissions** sont la somme des erreurs OMIS1 et OMIS2.

Il est bon de noter que les omissions apparaissent donc dans deux mesures agrégées.

Dans ce genre d'étude, il est utile de calculer l'accord inter-annotateurs (*inter-annotator agreement*). Typiquement, celui-ci est le coefficient « kappa » défini par Cohen [7]. Cette mesure est controversée (voir p.ex. [8]) et s'applique difficilement à notre cas. En effet, ces mesures s'appliquent aux cas où 2 annotateurs classent *n* objets dans *m* catégories discrètes et mutuellement exclusives. Les valeurs *n* et *m* doivent être fixes et connues d'avance. Or, dans notre cas, le nombre *n* correspond au nombre de sites d'annotations, qui varie selon l'annotateur. De plus, les *m* catégories ne sont pas mutuellement exclusives, puisqu'un site d'annotation peut contenir plus d'un type d'erreur. Nous proposons donc de remettre le calcul de l'accord inter-annotateurs à une étude ultérieure, si le besoin s'en faisait sentir.

3.2 Qualité de la traduction de bulletins complets

Afin d'obtenir des statistiques sur la qualité des traductions des bulletins *complets*, comme utilisé dans le rapport du BT [3], nous avons regroupé, pour chaque bulletin, les annotations des phrases qui le composaient et les avons agrégées.

Dans un bulletin donné, pour calculer la statistique correspondante à chacun des types d'erreurs du tableau 2.3, on a fait la moyenne des annotations de chaque annotateur sur chaque phrase, puis nous avons fait la somme de ces moyennes pour chacune des phrases du bulletin.

> **EXEMPLE.** Dans un bulletin fictif de 2 phrases, la sortie machine de la phrase 1 a été annotée avec 1 erreur CHA1 par l'annotateur A et 0 erreur CHA1 par l'annotateur B. On compte ainsi une moyenne de 0,5 erreur CHA1 pour la phrase 1. Si la phrase 2, elle, a reçu une moyenne de 1 erreur CHA1, alors le bulletin complet reçoit un nombre d'erreurs CHA1 de 0,5 + 1 = 1,5 erreur CHA1.

Cette méthode de calcul pénalise donc les bulletins plus longs. Nous avons jugé qu'il n'y avait pas lieu de normaliser par le nombre de phrases du bulletin, car notre calcul traduit la réalité qui veut que la traduction d'un bulletin plus long offre plus de chances pour le système de traduction (humain ou machine) de commettre une erreur. Par ailleurs, les comparaisons subséquentes entre humain et WATT ne sont pas touchées par ce détail.

Quant à elle, la mesure d'exactitude de transfert du sens (voir la question à la section 2.2.4) est une mesure de type oui/non. « Oui » signifie que le sens est maintenu, « non » signifie le contraire. Pour regrouper les annotations de toutes les phrases pour savoir si un bulletin est exact ou non, on a examiné les annotations de toutes les phrases et de tous les annotateurs pour un système donné (machine ou humain). Si une phrase ou plus était annotée « non », alors tout le bulletin recevait une mesure « non » pour l'exactitude. C'est donc la plus sévère des interprétations des résultats que nous avons utilisée.

> **EXEMPLE.** Dans un bulletin fictif de 6 phrases, la sortie humaine de 5 phrases, pour tous les annotateurs est « oui » à la question d'exactitude. Ceci fait donc 10 annotations (2 annotateurs par phrases) donnant « oui ». Cependant, la dernière phrase est annotée « oui » par l'annotateur C et « non » par l'annotateur D. Alors, puisqu'une seule annotation négative suffit pour rendre un bulletin inexact au sens de cette métrique, le bulletin reçoit un « non » pour la mesure.

Enfin, pour obtenir le nombre d'erreurs moyen par bulletin, nous avons fait la moyenne des métriques expliquées plus haut, pour chaque type d'erreurs, sur tous les bulletins annotés, au nombre de 32.

Puisque l'influence de la qualité des sources nous intéresse également, nous avons fait le calcul expliqué plus haut dans deux configurations différentes :

- **Sans filtre**, ce qui correspond au calcul expliqué précédemment et
- Filtré par les phrases **source sans faute**, ce qui consiste à retirer du calcul d'erreur toute paire de phrases pour laquelle au moins un annotateur a signalé au moins une erreur dans la phrase source.

  > **EXEMPLE.** Dans le bulletin fictif de l'exemple précédent comptant 6 phrases, l'annotateur A a trouvé une faute de frappe dans la source de la phrase 1 et une autre dans la source de la phrase 6. Par conséquent, toutes les erreurs qui auraient pu survenir dans les traductions des phrases 1 et 6 sont éliminées des statistiques extraites du bulletin, sans distinction.

Ces 2 configurations permettent de voir dans quelle mesure les sources engendrent, en moyenne, une augmentation des erreurs de traduction, chez la machine et l'humain. Il est bon de noter à ce stade que 25 des 101 phrases source contiennent au moins une erreur, selon au moins un annotateur. Nous revenons sur la qualité de l'intrant à la section 3.3.

La figure 3.1 montre le nombre moyen d'erreurs de chaque type, par bulletin. On a ces statistiques pour les sorties humaines et pour les sorties machine, en utilisant les configurations « sans filtre » et « source sans faute » expliquées plus haut, ce qui correspond à 4 cas de figure.

Dans la même veine, on a calculé le pourcentage des bulletins traduits contenant au moins une erreur d'exactitude (selon la question de la section 2.2.4), pour ces 4 cas de figure. On présente ces résultats à la figure 3.2.

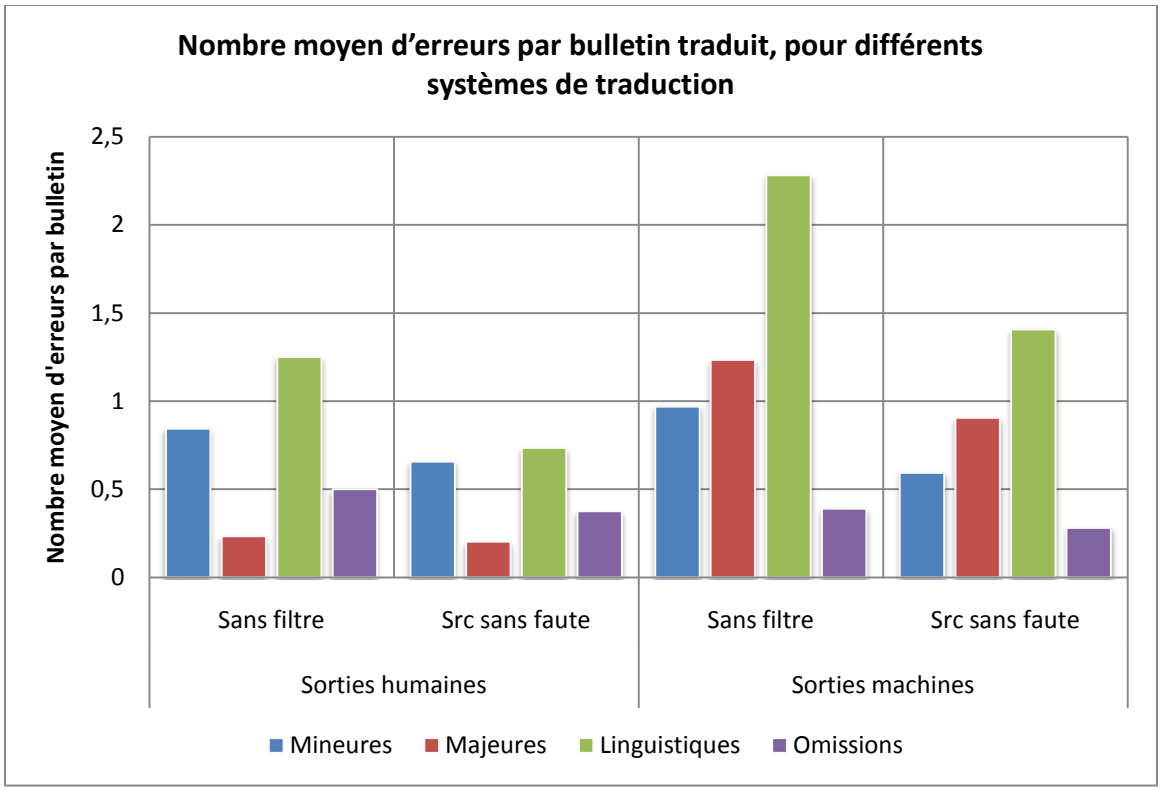


Figure 3.1

Nombre moyen d'erreurs de différents types, par bulletin. Pour chaque système de traduction, il y a deux ensembles de statistiques : « sans filtre » signifie que toutes les traductions sont considérées, « Src sans faute » ne considère que les traductions de phrases source sans faute. 25 des 101 phrases source contiennent au moins une erreur. La moyenne des améliorations entre traductions humaines et machine est de 0,4 erreur en moyenne, et l'amélioration entre les traductions de phrases source sans faute et celles non filtrées est de 0,3 erreur.

La figure 3.1 permet de faire les observations et précisions suivantes :

- Les traductions humaines sont de meilleure qualité : le nombre moyen d'erreurs par bulletin est inférieur pour presque tous les types d'erreurs. Il y a par exemple 1 erreur majeure en plus par bulletin lorsque l'on passe de la traduction humaine à la traduction machine. Cet écart est moins prononcé pour les autres types d'erreurs, et vaut en moyenne 0,3 erreur.
- Les omissions sont plus rares dans les traductions machine que dans les traductions humaines. Cela s'explique par la « rigidité » des règles de traduction embarquées par le prototype : il ne s'autorise aucune reformulation majeure ni aucun oubli. Les quantités, les noms de lieux sont invariablement transmis dans la

traduction avec une très haute fidélité, ce qui est notoirement plus difficile (et plus pénible) pour l'être humain[4].

- La qualité des phrases source influence de façon significative la qualité des traductions, et ce, pour l'humain et pour la machine. On observe en effet, toutes configurations confondues, une diminution moyenne de 0,2 erreur par bulletin chez l'humain lorsqu'on lui présente des sources irréprochables, et une diminution du double (0,4 erreur) pour la machine. Ceci confirme l'intuition qui veut qu'un système de traduction automatique traduira une entrée corrompue en une sortie elle aussi corrompue.

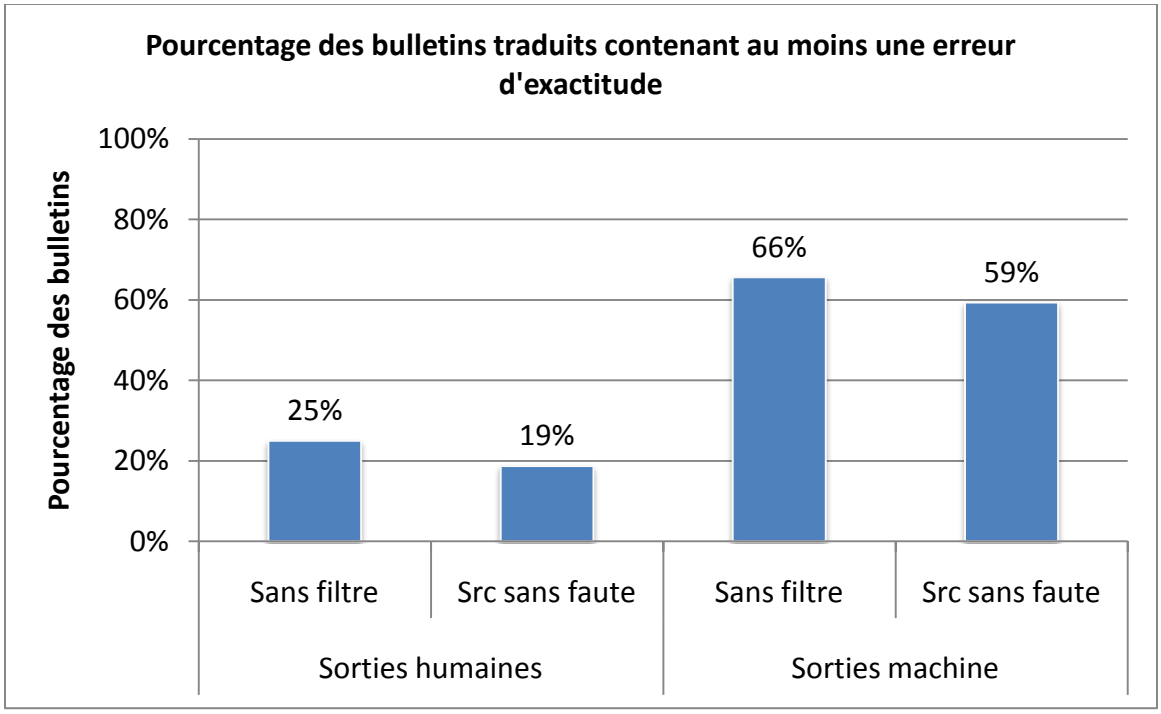


Figure 3.2
Pourcentage des 32 bulletins traduits contenant au moins une erreur d'exactitude, selon la question de la section 2.2.4).

La figure 3.2, qui traite des erreurs au niveau des bulletins, permet de constater que :

- Elle corrobore les constatations faites à partir de la figure précédente.
- En moyenne, 66 % des bulletins traduits par le prototype contiennent au moins une traduction infidèle, selon au moins un annotateur. Comme indiqué à la page 7, le BT avait indiqué un taux d'erreurs très similaire sous la forme du taux de bulletins contenant une erreur. Notre évaluation permet d'expliquer les

[4] Des outils de contrôle de la qualité automatisés s'appuient d'ailleurs sur cette fidélité des systèmes informatiques pour valider les traductions. Voir notamment TransCheck, http://rali.iro.umontreal.ca/Traduction/TransCheck.fr.html.

hypothèses fortes sous-jacentes à cette mesure et de la mettre en contraste avec la performance humaine.

En effet, 25 % des bulletins traduits par des humains contiennent une erreur d'exactitude, un nombre relativement élevé. Il faut toutefois rappeler que la méthode de mesure (expliquée à la section précédente) est sévère, puisqu'il suffit qu'un seul annotateur annote une phrase comme ayant un problème de transfert de sens pour que le bulletin soit considéré comme inexact. La métrique ignore ainsi un désaccord entre annotateurs sur ce point.

Les annotateurs ont été très exigeants dans le transfert du sens, puisque cette mesure extrinsèque de la qualité des bulletins est critique. On montre ainsi, à la figure 3.3, deux exemples de traductions acceptables de prime abord, mais jugées inexactes par les annotateurs. On y observe leur minutie.

Comme dans toute évaluation humaine, certains cas sont discutables, mais, puisque cette évaluation se fait à l'aveugle et sur un bon nombre de traductions, on peut considérer que les deux méthodes de traduction, humaine et machine, sont jugées de façon comparable et impartiale, et que, au fil des annotations, certaines annotations plus dures sont compensées par d'autres plus clémentes.

Traduction humaine
Source
Diminishing northerly winds and clear skies will give ideal conditions tonight for temperatures to fall to the low single digits over Eastern Nova Scotia.
Cible
Des vents du nord diminuant d'intensité et le ciel dégagé fourniront les conditions idéales cette nuit pour faire tomber les températures à 1 à 3 degrés sur l'est de la Nouvelle-Écosse.
Traduction machine
Source
The moisture from the recent rains combined with daytime heating are creating conditions favourable for the development of thunderstorms.
Cible
L'humidité par les récentes pluies, combiné au réchauffement diurne qui tombe crée des conditions propices à la formation d'orages.

Figure 3.3

Exemples de deux traductions qui nous semblent acceptables de prime abord, mais que les annotateurs ont jugées inexactes dans leur transfert du sens de l'original, pour l'humain et pour la machine. On a souligné les mots que l'annotateur considère comme le site de la perte de fidélité.

Nous avons également reporté les décomptes d'erreurs par bulletin, faits par le Bureau de la traduction dans son rapport [3] (page 11) afin de faciliter la comparaison de nos résultats aux leurs. Ceci est présenté à la figure 3.4 pour les erreurs majeures et à la figure 3.5 pour les erreurs mineures. On constate que :

- Pour les erreurs majeures, les résultats de cette étude sont très semblables à ceux du BT, ce qui tend à prouver qu'elle est cohérente avec la leur, même si nos métriques sont sensiblement différentes. Les évaluateurs ont été moins critiques envers les traductions machine, où l'on ne relève jamais plus de 4 erreurs majeures par bulletin, contrairement à 13 pour l'étude du Bureau.
- Les erreurs majeures sont beaucoup plus rares chez l'humain, ce qui est cohérent avec nos observations précédentes.
- Les erreurs mineures sont plus rares chez la machine, ce qui s'explique par le total d'omissions, qui sont plus rares pour WATT, comme on l'a vu à la figure 3.1. Cela s'explique aussi par des définitions différentes des métriques utilisées, ce qui est particulièrement visible pour les erreurs mineures, qui vont jusqu'à 38 pour l'étude du BT, alors que le maximum est de 5 pour notre étude.

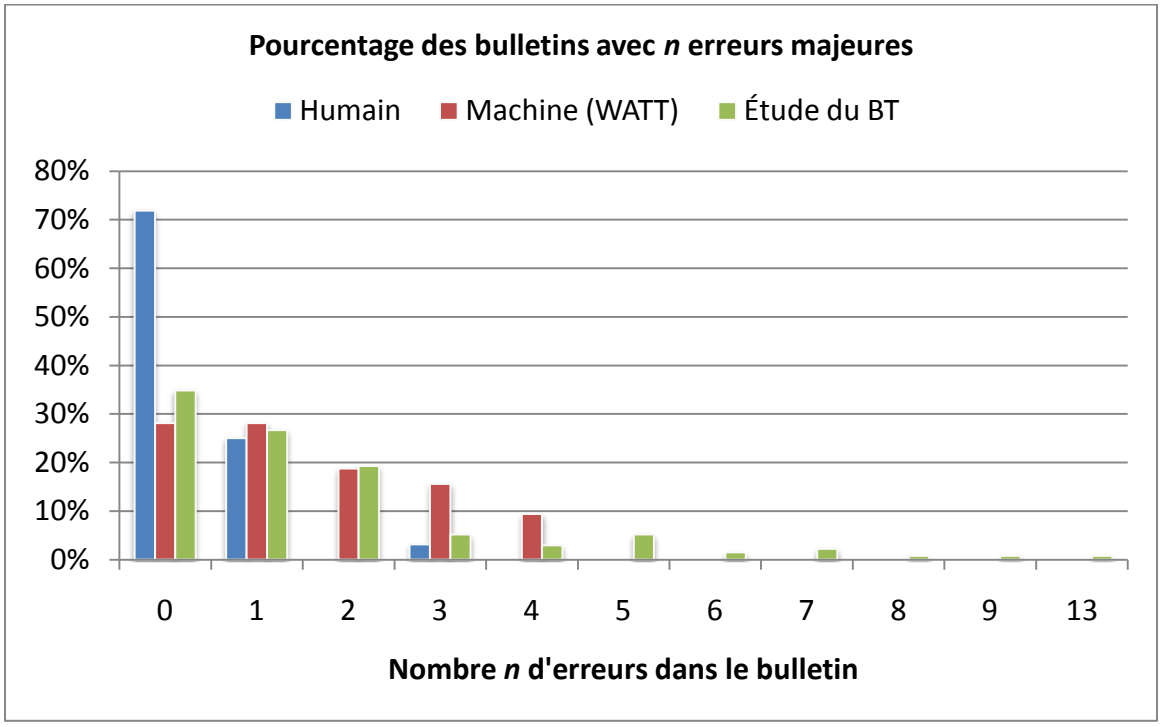


Figure 3.4
Ventilation par nombre d'erreurs majeures, montrant le pourcentage de bulletins contenant 0, 1, 2, … erreurs majeures.

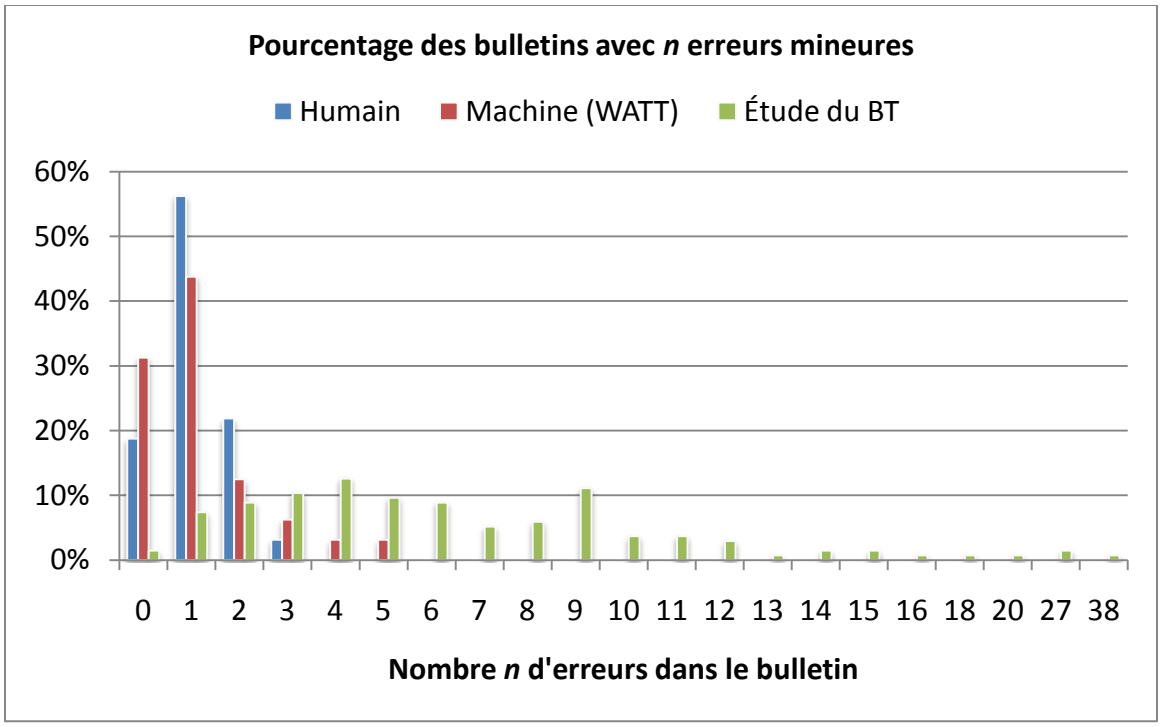


Figure 3.5
Ventilation par nombre d'erreurs mineures, montrant le pourcentage de bulletins contenant 0, 1, 2, … erreurs mineures.

3.3 Qualité de la source

Nous nous sommes intéressés à la qualité de la source, car elle représente la qualité du texte véhiculé dans la chaîne de diffusion des informations météorologiques. De plus, elle influence directement la qualité des traductions, comme on l'a vu à la section précédente. On rappelle que l'on considère que la langue source d'un bulletin est le français au Québec, et l'anglais ailleurs.

3.3.1 Méthodologie

Les métriques que nous présentons ici concernent les phrases source, donc écrites par des humains et non traduites. Les annotations portent sur 404 phrases source, puisque l'on a autant de paires de phrases à évaluer (voir section 2.2.3). Ces 404 phrases source annotées correspondent au travail de 4 annotateurs sur 101 phrases source chacun.

Les erreurs repérées dans les sources par les annotateurs sont exclusivement des erreurs d'ordre linguistique, soit celles de type LANG et TYPO (voir tableau 2.3, page 14); les autres types ne s'appliquent pas.

3.3.2 Fréquence des erreurs source

Des 404 annotations de phrases source,

- 3,7 % contenaient au moins une erreur de type LANG
- 10,6 % contenaient au moins une erreur de type TYPO

- 13,3 % contenaient au moins une erreur de type LANG ou TYPO (on n'obtient pas la somme des deux précédents nombres, car une phrase source annotée peut contenir à la fois une erreur LANG et une TYPO).

On obtient une métrique agrégée si l'on ramène ces 404 annotations aux 101 phrases source et que l'on considère le nombre de phrases pour lesquelles au moins un annotateur a trouvé une erreur.

> **EXEMPLE.** Une phrase source fictive *p* a été annotée 4 fois, comme toutes les phrases source. L'annotateur B a trouvé une erreur de typographie (TYPO) dans cette source et l'annotateur C, une erreur dans un nom de lieu (LANG). Les annotateurs ne sont pas d'accord, mais on compte néanmoins cette phrase source comme contenant une erreur. On compte ensuite le nombre de phrases source dans le même cas que la phrase *p*, divisé par 101 pour obtenir le pourcentage.

En utilisant cette dernière métrique, on constate que 25 phrases source contiennent une erreur, soit 25 % des 101 phrases.

3.3.3 Nature des erreurs source

Pour obtenir un portrait aussi précis que possible des erreurs dans la source des traductions, nous avons réexaminé toutes les annotations faites dans les sources. Nous avons considéré uniquement les erreurs uniques dans les sources, c'est-à-dire que si plus d'un annotateur repérait une erreur, alors on la comptait une seule fois.

Ceci nous a permis d'identifier 30 erreurs uniques dans les sources, que nous avons réparties en 5 classes, correspondant à 5 problèmes distincts dans la chaîne de diffusion, selon nous.

Ces 5 classes sont décrites et illustrées par des exemples tirés des phrases sources annotées au tableau 3.1. Leur nombre et répartition sont montrés à la figure 3.6.

Tableau 3.1
Exemples d'erreurs dans les sources, telles que repérées par les annotateurs. On a souligné le site des erreurs signalées afin de clarifier le travail de l'annotateur. Lorsqu'un problème de casse survenait dans un nom de lieu, on a assigné ce problème à la classe « Nom de lieu mal écrit » plutôt que « Casse », qui est plus générique.

Catégorie	Phrase source annotée
Ponctuation	As a result..Patchy frost will form over low lying inland areas where ground temperatures will be near zero.
Nom de lieu mal écrit	Ils devraient traverser le secteur de la minerve - rivière rounge dans la prochaine heure.
Incohérence de notation	About 20 millimetres of rain has already fallen. Another 15 to 25 mm of rainfall is expected in Old Crow before the disturbance begins to weaken this evening.
Orthographe	Weather conditons are no longer favourable to produce funnel clouds.
Casse	attention...

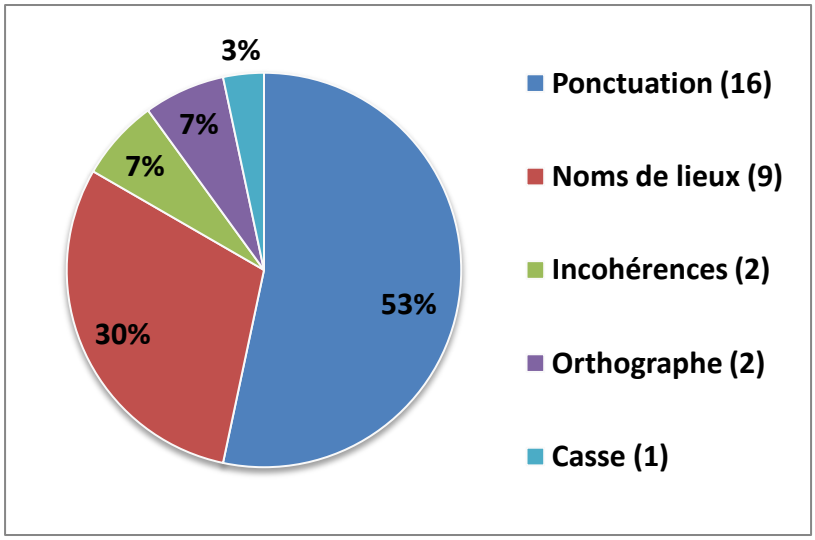


Figure 3.6
Ventilation des 30 erreurs uniques repérées par les annotateurs dans les phrases sources du prototype. Entre parenthèses, on retrouve le compte (sur 30 erreurs uniques).

Comme on peut s'y attendre dans ce genre d'évaluation, certaines erreurs repérées sont plus graves que d'autres, et certaines sont discutables. Le but de ce travail est de signaler que les entrées du prototype sont perfectibles et d'en montrer les aspects qui bénéficieraient d'une révision.

Il est utile de noter les points suivants :

- 30 erreurs uniques ont été repérées par les annotateurs, touchant 25 phrases source (voir section 3.3.2). Certaines phrases source contiennent donc plus d'une erreur.
- Les erreurs les plus communes sont des erreurs de ponctuation. Il semble que la chaîne de traitement actuelle ait de la difficulté à éliminer des bulletins « grand

public » certaines marques spécifiques aux formats internes de la chaîne de diffusion, nommément les séparateurs de phrases « .. » et « - » en lieu et place d'une marque de fin de phrase. Nous avons vérifié cela de façon anecdotique en lisant certains avertissements au moment de la rédaction de ce rapport et avons trouvé au moins un exemple de ce problème, que nous illustrons à l'annexe C.

- Les erreurs de noms de lieux sont relativement fréquentes. La casse est souvent incorrecte (*« Great Slave lake » pour « Great Slave Lake »), les mots sont mal accentués (*« st-zenon » pour « St-Zénon ») ou simplement mal orthographiés (*« rivière rounge » pour « rivière Rouge »).
- Des incohérences dans la notation des quantités ont fait tiquer les annotateurs. Ainsi, dans le même bulletin, on peut utiliser l'abréviation « mm » pour millimètre, puis écrire au long l'unité « centimètres ». À notre avis, ceci n'est pas là du purisme : uniformiser facilite la lecture. Cela est d'autant plus vrai que l'on retrouve parfois des périphrases du type « de la grêle de la taille d'une balle de golf » ou « de la taille d'une balle de baseball »[5] dans les bulletins, des informations assez vagues pour le non-sportif. Des erreurs de notation se retrouvent également dans l'expression des heures, où tous les annotateurs ont découvert la coquille *« 8;20 PM EDT » pour « 8:20 PM EDT ».
- On observe moins fréquemment des erreurs d'orthographe ailleurs que dans les cas cités plus haut. Néanmoins, les annotateurs ont repéré un *« Weather conditons » pour « Weather conditions », qui est sans doute une faute de frappe.

À la lumière de ces constatations, il est envisageable d'améliorer la qualité des bulletins produits en intervenant dès la saisie du texte dans la langue source, par le météorologue. Les rectifications suivantes profiteraient au système de saisie :

- Vérification des marqueurs de fin de phrase (« .. », par exemple) et leur remplacement automatique par leur équivalent correct.
- Validation des noms de lieux en comparant ceux saisis avec une banque préalablement constituée des noms possibles dans les bulletins météorologiques. Nous avons d'ailleurs utilisé une liste de noms de lieux pour améliorer notre prototype de traduction. Cette liste nous a été transmise par EC.
- Remplacement automatique de toutes les variantes d'unités de mesure et périphrases par des unités du Système international d'unités (SI), et utilisation de leurs abréviations respectives partout, de façon cohérente. Ceci devrait également s'appliquer à l'expression des heures et des dates.

[5] Ces exemples sont tirés d'autres corpus de bulletins d'alertes météorologiques d'EC, pas de celui utilisé dans cette étude. La phrase complète est « DE LA GRELE DONT LA GROSSEUR VARIAIT ENTRE UNE BALLE DE GOLF ET UNE BALLE DE BASEBALL PROVENANT D UN ORAGE A L OUEST DE RADVILLE A ETE SIGNALEE ET UNE TORNADE PROVENANT DU MEME ORAGE A BRIEVEMENT TOUCHE LE SOL . ».

- Ajout d'un correcteur orthographique et/ou grammatical pour éviter les coquilles ou d'autres erreurs malheureuses qui polluent le reste de la chaîne de traitement et de diffusion[6].

Il faut noter également qu'on pourrait pousser plus loin l'évaluation de la qualité de l'intrant en annotant non pas 101 phrases source, mais un nombre plus grand, dans le cadre d'une évaluation humaine dédiée à ceci. Une étude de ce type pourrait révéler des problèmes supplémentaires dans la saisie des bulletins, et pourrait donner des statistiques plus précises de ventilation des erreurs. Par ailleurs, elle aussi pourrait être effectuée à l'aveugle de façon à ce que l'annotateur ne sache pas s'il annote une phrase humaine ou produite par une machine.

L'annexe D présente les annotations des phrases source.

L'effet des erreurs dans la source sur la qualité de la traduction a été expliqué à la section 3.2.

3.4 Erreurs de traduction machine

Cette section explore la nature et la fréquence des erreurs commises par le prototype. Nous cherchons également ici à déterminer la nature des efforts nécessaires pour l'améliorer et à quantifier ces améliorations potentielles. Nous n'incluons donc pas dans ces résultats les statistiques sur les traductions humaines.

3.4.1 Nature et fréquence des erreurs

Chacune des 101 phrases source a été traduite par le prototype, et les sorties ont été annotées deux fois, par deux annotateurs différents. Ici, on considère la somme des erreurs de chaque type (voir tableau 2.3).

La figure 3.7 montre le nombre moyen d'erreurs par traduction machine, pour chacun de ces types d'erreurs. Les annotateurs ont repéré 271 erreurs de traduction en tout, ce qui correspond à une moyenne de 271/202 = 1,34 erreur en moyenne par phrase traduite par le prototype. Par comparaison, les humains ont commis 130 erreurs de traduction, soit 0,64 erreur en moyenne par phrase traduite.

La figure 3.8 montre la fréquence relative des erreurs de traduction par le prototype.

[6] Il est utile de noter que le simple fait de coller les phrases source dans des documents Microsoft Word pour l'annotation et pour la rédaction de ce rapport a permis de révéler plusieurs des erreurs signalées par les annotateurs, *via* la correction grammaticale de Word. Il est donc clair que l'utilisation d'un outil de ce type bien commun améliorerait la qualité des intrants du système.

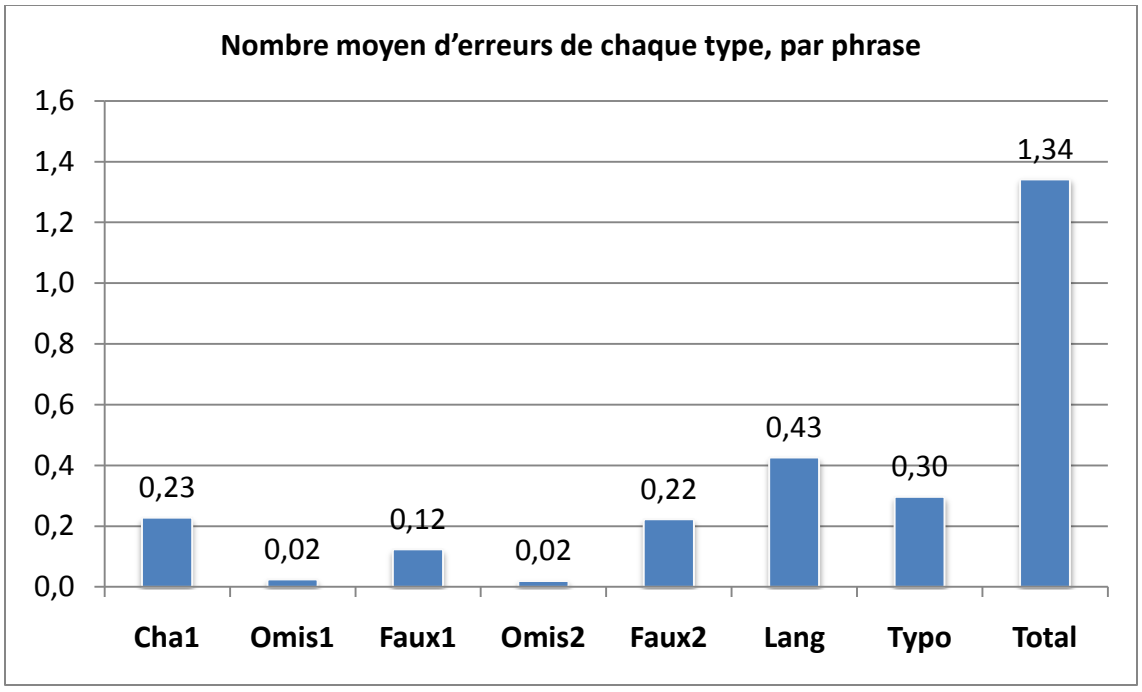


Figure 3.7
Nombre moyen d'erreurs par phrase traduite par le prototype, pour chacun des types d'erreurs expliqués au tableau 2.3. Le nombre de traductions machine annotées est de 202.

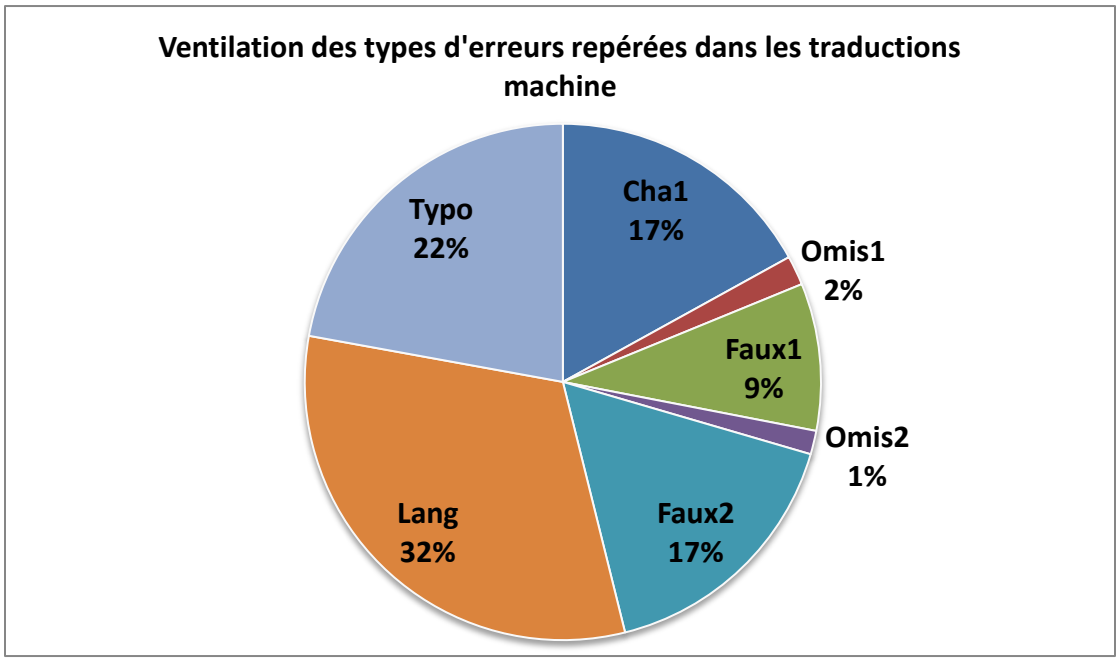


Figure 3.8
Ventilation des 271 erreurs de traduction repérées dans les 202 traductions machine annotées.

Des deux précédentes figures, on retient les points suivants :

- Les erreurs majeures (CHA1, OMIS1, FAUX1) comptent pour 28 % des erreurs repérées.
- Les erreurs mineures (OMIS2, FAUX2) comptent pour 18 % des erreurs repérées.
- Les erreurs linguistiques (TYPO, LANG) comptent pour 54 % des erreurs repérées.

Les types d'erreurs relevés sont fort différents, à deux égards :

1. Certains affectent de façon plus déterminante que d'autres la qualité des traductions (les faux-sens et le charabia, nommément).
2. Certains pourraient être corrigés relativement aisément, avec les ressources adéquates, alors que d'autres représentent des limites inhérentes à la traduction statistique, même lorsque cette dernière met en œuvre les derniers développements dans le domaine, comme c'est le cas de WATT.

Il y a donc lieu d'examiner davantage les erreurs du prototype, ce que nous faisons à la section qui suit.

3.4.2 Atténuation des erreurs par rectification du prototype

Le but de cette section est de déterminer dans quelle mesure, et avec quel effort, le prototype peut être amélioré pour engendrer de meilleures performances de traduction.

Pour ce faire, nous avons réexaminé manuellement[7] toutes les erreurs de traduction découvertes par les annotateurs (au nombre de 271 sur 202 traductions) et les avons regroupées dans trois classes dénotant chacune la durée approximative du travail nécessaire pour les éliminer d'une mouture future du prototype. Ces classes sont les suivantes :

- **Classe facile (quelques semaines de travail).** Cette classe regroupe toutes les erreurs pour lesquelles peu ou pas d'effort est nécessaire pour les éliminer d'un nouveau prototype. Elle regroupe notamment les erreurs de langue que les annotateurs ont découvertes, mais qui sont en fait des tournures propres à la langue des bulletins météo. On y retrouve par exemple l'absence d'article devant certains noms, comme dans le début de phrase « Quantité de pluie totale de … », ou encore l'omission volontaire de certains mots, tels « est prévue », pour désigner un événement météorologique imminent. Ces erreurs pourraient être facilement filtrées ou simplement laissées telles qu'elles, puisqu'elles font partie du jargon du domaine, compris par les consommateurs des bulletins. Dans tous les cas, elles requièrent peu d'efforts pour être éliminées, et pourraient nécessiter un travail de quelques semaines.
- **Classe moyenne (quelques mois de travail).** Cette classe regroupe les erreurs d'accentuation, de restauration de la casse et de transfert de certaines quantités numériques.

[7] À nouveau, nous avons effectué ceci de façon semi-automatisée en utilisant le langage de programmation embarqué dans Microsoft Word.

Il faut en effet rappeler que dans la chaîne actuelle de diffusion des informations météorologiques, deux formats coexistent, soit le format accentué et à la casse habituelle, et un autre, plus technique, en toutes majuscules et sans accents. Le processus de restauration de la casse et des accents est par conséquent important, et pourrait être raffiné au sein du prototype. La nature de ces rectifications est bien connue et celles-ci pourraient être mises en œuvre à moyen terme. Ces corrections toucheraient notamment la casse des noms de lieux, dont on a souligné l'importance plus tôt.

Enfin, l'analyse des erreurs de traduction nous a clairement montré qu'il nous faut corriger certaines règles régissant le transfert des quantités numériques de la phrase source à la traduction. Il suffit en théorie de rajouter des règles dans le prototype, et ceci peut être effectué à moyen terme, en quelques mois.

Cette dernière étape de correction pourrait profiter de l'avis d'Environnement Canada, afin d'uniformiser la façon d'exprimer les unités, les dates et les heures contenues dans les bulletins pour se conformer à leurs directives de qualité. On a vu en effet que ceci peut poser problème à la section 3.3.3.

- **Classe difficile (à plus long terme).** Cette classe recense toutes les erreurs dont la correction est plus problématique, puisqu'elles touchent aux limites de l'approche de traduction automatique. Leur résolution demandera des solutions innovatrices développées à moyen ou long terme.

  Ces erreurs appartiennent généralement aux types suivants :

  — Fautes de grammaire, notamment d'accord en genre et nombre.

  — Mots inconnus, ou très rares, qui déconcertent le prototype. Parmi ces erreurs, on retrouve par exemple une traduction incorrecte de « These winds can break branches and knock down diseased trees. » par « Ces vents peut pause branches d'arbres et jeter par terre diseased. », un charabia qui illustre le fait que le mot « diseased », inconnu, perturbe la traduction automatique.

    Une uniformisation de la langue source lors de la saisie contribuerait à diminuer les erreurs de mots inconnus.

  — Formulations incompréhensibles dans la traduction, dues à des règles de traduction incorrectes.

  — Anaphore non résolue, par exemple, une traduction automatique utilisera un pronom « elles » au lieu de « précipitations », ce qui diminue forcément la compréhensibilité de la traduction.

Cette classification autorise en fin de compte une étude de la facilité avec laquelle on peut améliorer le prototype à court, moyen et long terme. La figure 3.9 montre la répartition des types d'erreurs dans chacune des classes précitées, ainsi qu'une répartition du total de toutes les erreurs de traduction repérées (271 en tout). On y a regroupé les classes « facile » et « moyenne » pour faciliter la lecture des résultats.

On constate que :

- Les erreurs mineures et linguistiques sont les plus faciles à atténuer, avec 60 % d'erreurs évitables pour des efforts à moyen terme.
- Les erreurs majeures sont les plus difficiles à éliminer, ce qui est, intuitivement, ce à quoi l'on s'attendait.
- En tout, une erreur sur deux pourrait être éliminée de façon « facile » ou « moyenne », donc dans un temps raisonnable de recherche et développement.

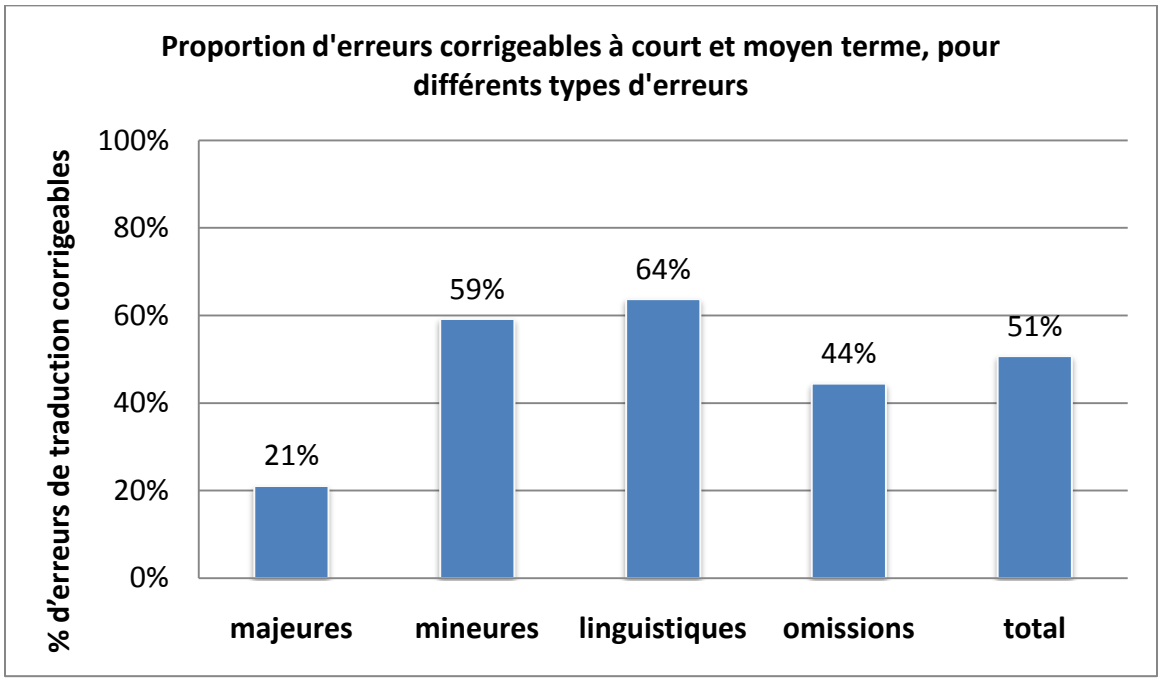


Figure 3.9
Proportion des différents types d'erreurs de traduction dans la catégorie « facile-moyenne » dénotant le pourcentage des erreurs de chaque type pouvant être éliminées à court et moyen terme. La dernière colonne « total » montre le total pour les 271 erreurs repérées, toutes catégories confondues.

En outre, il est intéressant de noter que la résolution de problèmes engendrant des erreurs mineures de traduction pourrait fort bien contribuer également à empêcher l'apparition de certaines erreurs majeures. À ce stade, cependant, cette « coopération » entre erreurs est difficilement quantifiable.

3.4.3 Atténuation des erreurs par assainissement des phrases source

Un autre aspect de l'amélioration de la chaîne de traduction des bulletins est l'assainissement des phrases source. Comme on l'a montré aux sections 3.2 et 3.3, la qualité de la source influe de façon décisive sur la qualité des deux systèmes de

traduction. Il est donc utile de tenter de quantifier l'atténuation des erreurs du prototype en épurant les sources qu'il traduit.

Cependant, cette métrique est difficile à obtenir, dans la mesure où il faudrait, pour chaque erreur dans la source, déterminer comment elle a pu induire le prototype en erreur, si c'est le cas. Telle faute d'orthographe a-t-elle conduit à telle erreur; telle faute dans la ponctuation a-t-elle été sans conséquence ? Une telle analyse, pour les quelque 271 erreurs de traduction du prototype, n'a pu être envisagée dans les temps que nous avons prévu accorder à cette évaluation.

On s'est intéressés donc à une approximation de l'effet des erreurs de la source sur les traductions. Cette métrique ressemble à celle utilisée à la figure 3.7, et compte le nombre d'erreurs moyen par phrase, mais cette fois-ci en éliminant les erreurs de traduction qui sont associées à une phrase source contenant au moins une erreur.

Cette statistique repose donc sur une hypothèse forte, rendue nécessaire par les circonstances pratiques de notre contexte d'évaluation. Cette hypothèse veut que *toutes* les erreurs de traduction d'une phrase source erronée sont dues à l'erreur (ou aux erreurs) dans la source. En d'autres termes, elle signifie que si la source avait été irréprochable, les traductions l'auraient été aussi. Ce n'est pas forcément le cas, il faut donc lire cette métrique en gardant à l'esprit qu'elle constitue une *limite supérieure* à l'influence des erreurs source sur les erreurs cible.

Ces mises en garde faites, le contraste est maintenant possible entre la moyenne d'erreurs par phrase telle que présentée à la figure 3.7 (« sans filtre »), et cette limite supérieure (« Sources sans faute »). Nous montrons ce résultat à la figure 3.10.

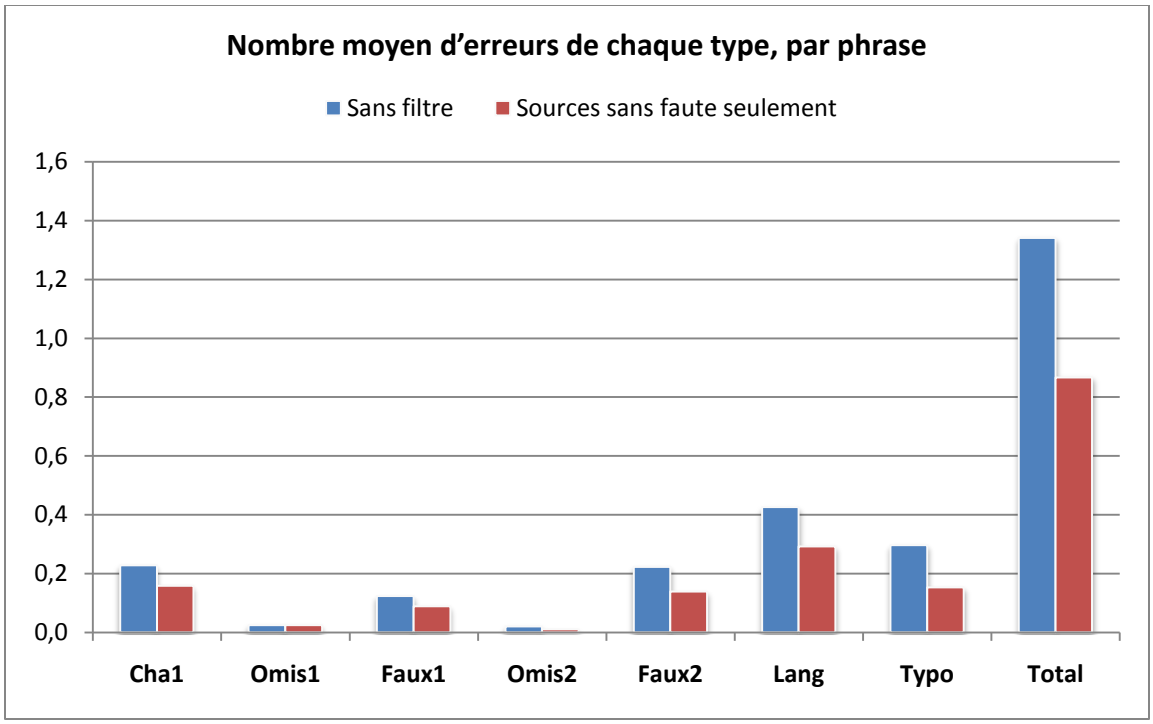


Figure 3.10
Contraste entre le nombre moyen d'erreurs par phrase tel que présenté à figure 3.7 et ce nombre réévalué en éliminant les erreurs de traduction associées à des phrases source contenant au moins une erreur. Le dénominateur dans chaque cas est 202, soit le nombre de traductions machine annotées.

On observe qu'un gain est possible en corrigeant les sources, mais il est difficile de le quantifier. Dans le meilleur des cas, le nombre moyen d'erreurs total (la dernière colonne de la figure 3.10) pourrait passer de 1,3 à 0,9, ce qui serait désirable, naturellement. Au-delà de ces chiffres, c'est avant tout une intuition qui se confirme, soit qu'une source épurée viendrait en aide au prototype.

Cela est d'autant plus vrai que le prototype *apprend* des couples source-traduction humains, dans son processus de modélisation statistique. Lui présenter des sources erronées est donc doublement handicapant pour lui.

3.4.4 Portrait complet de l'atténuation des erreurs de traduction machine

À ce stade, on a identifié deux stratégies possibles pour atténuer les erreurs de traduction du prototype soit, d'une part, des réfections en temps raisonnable du prototype (classes « moyenne » et « facile » à la section 3.4.2), et, d'autre part, l'assainissement des sources qui lui sont présentées.

Nous avons jugé opportun de résumer dans un seul graphique les influences de ces deux stratégies, d'abord prises isolément, puis lorsque combinées. Nous présentons ceci à la figure 3.11.

La figure 3.11 montre que chaque amélioration apporte sa part d'atténuation des erreurs, et que leur effet combiné est encore plus grand. Lorsque l'on considère les erreurs totales, le nombre moyen d'erreur passe de 1,3 sans améliorations, à 0,49 lorsqu'on emploie les deux stratégies en même temps. En tout, cela pourrait faire baisser le nombre d'erreurs de plus de la moitié.

Il n'en reste pas moins, en dernière analyse, que les types d'erreurs les plus dommageables au transfert fidèle du contenu, soit CHA1 et FAUX1, restent relativement stables. Leur somme passe de 0,35 à 0,21 après amélioration. Ce dernier chiffre signifie qu'une erreur de ce type est commise en moyenne toutes les 5 phrases, ou qu'un bulletin de 5 phrases a $(100\ \% - 21\ \%)^5 = 31\ \%$ de chance de ne *pas* contenir une erreur de ce type, donc 69 % de chance d'en contenir une.

Les erreurs LANG et TYPO, elles, se prêtent à merveille à ces améliorations. Leur somme passe de 0,73 à 0,2, une division par près de 3. Les erreurs de typographie, remarquablement, pourraient être complètement mises en sourdine par ces stratégies.

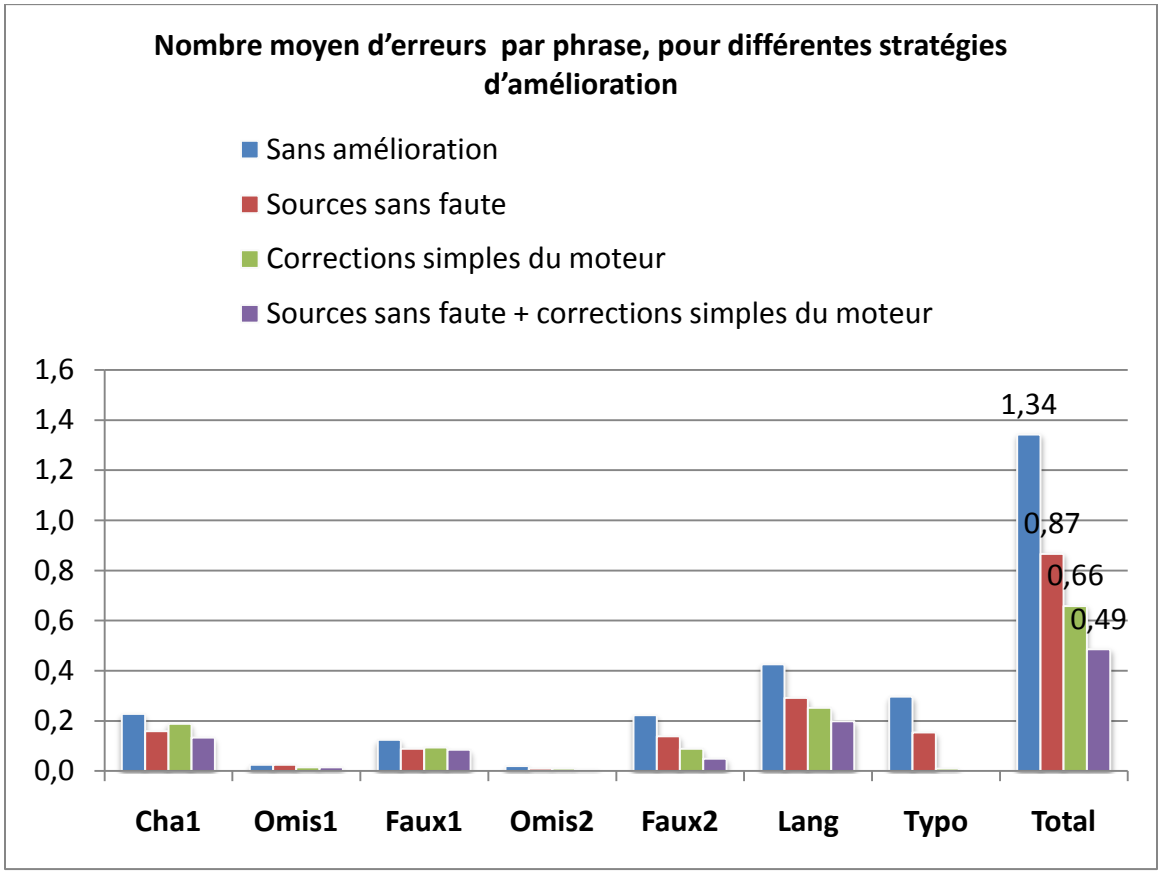


Figure 3.11

Influence respective de deux types d'améliorations possibles de la chaîne de traduction automatique, sur le nombre moyen d'erreurs par phrase traduite automatiquement. « Source sans faute » désigne l'amélioration *théorique maximale* apportée par des phrases source assainies (voir section 3.4.3), « Corrections simples du moteur » fait référence aux corrections « faciles » et « moyennes » expliquées à la section 3.4.2.

4 Conclusion

4.1 Réponses aux questions de la section 2.1

Ce rapport a présenté les résultats de l'évaluation humaine de WATT. La méthodologie d'évaluation a privilégié une évaluation à l'aveugle, dans des conditions réelles de fonctionnement du prototype, lorsque celui-ci était confronté à des exemples tirés de corpus vraiment diffusés sur le Web, et accessibles à tous les Canadiens. Notre analyse des résultats a mis l'accent sur une transparence des métriques utilisées, exemples à l'appui. Nous avons choisi le plus souvent les métriques les plus « sévères » pour analyser les résultats. Nous avons également souligné les limites de chacune, le cas échéant. Ceci est prévisible lorsqu'on traite les données bruitées d'une évaluation humaine, soumises aux aléas et aux dissensions de la réflexion des annotateurs humains.

Nous avons pu répondre à l'essentiel des questions posées à la section 2.1, que nous reproduisons ici, assorties de leurs réponses respectives. Nous tirons enfin des conclusions supplémentaires après celles-ci.

1. **Comment se comparent les qualités des traductions machine et humaines ?**
2. **Quelles sont les erreurs commises par WATT, et par le traducteur humain ?**

Les traductions humaines sont de meilleure qualité que les traductions machine non révisées. En effet, le nombre moyen d'erreurs par bulletin d'alerte est inférieur de 0,4 erreur à celui des traductions machine. Cette différence de qualité touche surtout les erreurs mineures (glissements mineurs de sens), majeures (par exemple les faux-sens ou le charabia). De plus, la qualité linguistique des bulletins traduits par les humains est supérieure. WATT commet 1,34 erreur en moyenne par phrase. Ceci corrobore certaines des conclusions du Bureau de la traduction tirées dans le cadre d'une évaluation du prototype [3], ce qui est bon signe. Ni l'humain ni la machine ne sont irréprochables lorsqu'ils traduisent, cependant.

Il faut souligner par exemple que les omissions d'éléments de sens sont plus rares chez la machine, ce qui s'explique par l'inflexibilité des règles de traduction embarquées dans le moteur de traduction, qui transposent presque sans faille les quantités, tournures, et lieux dans la traduction.

3. **Le sens de la phrase du bulletin original est-il correctement transposé dans la langue cible par WATT et par le traducteur humain ?**

Une mesure extrinsèque de fidélité de la traduction montre que 66 % des bulletins produits par le prototype contiennent au moins une traduction infidèle, comparativement à 25 % pour les traductions humaines.

4. **Quelle est la qualité du texte source (l'intrant), et comment influe-t-elle sur la qualité des traductions?**

La qualité de la source influence de façon importante la qualité des traductions automatiques et humaines. Ces erreurs touchent 25 % des phrases source, et sont en majorité des fautes de ponctuation, ainsi que des erreurs touchant souvent des noms de lieux. On relève aussi quelques fautes d'orthographe et des incohérences dans l'expression des unités de mesure et de temps. Ces constatations nous mènent à penser

qu'un système de contrôle de la qualité lors de la saisie des bulletins serait utile. De même, on pourrait aussi imaginer un système d'aide à la saisie des bulletins, dont profiterait le météorologue.

Systématiquement, lorsqu'on filtre les résultats d'évaluation en supprimant les phrases pour lesquelles les sources sont erronées, on observe une atténuation des erreurs de traduction. Un gain est donc envisageable en normalisant les entrées, ce qui est attendu, puisque les erreurs dans la source perturbent le moteur de traduction automatique. La section suivante montre comment cette normalisation pourrait se faire.

4.2 La place de *WATT* dans une chaîne de diffusion bilingue future

À la lumière des résultats obtenus ici, il nous semble que la chaîne de diffusion d'avertissements météo pourrait profiter grandement de l'intégration de WATT. Le moteur de traduction ne peut manifestement remplacer un être humain, étant donné qu'il commet plus d'erreurs, en moyenne, que le traducteur humain. Cela n'a d'ailleurs pas été notre prétention en entamant cette étude. La vocation de WATT est plutôt, à notre avis, de permettre une traduction quasi instantanée (quelque 5 s par bulletin) pour des informations dont l'urgence de diffusion est évidente, quitte à ce que ces traductions puissent bénéficier de la révision d'un expert humain par la suite.

Dans l'intervalle, l'avertissement serait affiché sur le Web avec une notice avisant le lecteur que le texte lu a été produit par un outil informatique et qu'il sera révisé sous peu[8]. Un exemple de notice possible est illustré à la figure 4.1. Cette notice discrète pourra inclure un hyperlien (icône dans notre figure) vers une page vulgarisant le mécanisme de diffusion d'EC. Cette dernière pourra notamment expliquer que le bulletin est en cours de révision et qu'il sera amendé sous peu.

Le processus de révision de traductions est fort différent de celui de la traduction en tant que telle. Essentiellement, il vise à décharger le traducteur de tâches répétitives sur du texte stéréotypé pour lui permettre de focaliser son attention sur des passages plus problématiques où le moteur de traduction a péché. Les études ne manquent pas dans le domaine, et portent sur le gain en productivité et en qualité du processus, sur l'interface de travail idéale, sur la nature des révisions, etc.

L'étude du gain potentiel en productivité d'un processus de révision des traductions en aval de WATT dépasse le cadre de la présente étude, mais pourrait s'avérer très utile pour chiffrer l'utilité du prototype dans la chaîne de diffusion des avertissements météo. Dans tous les cas, il est bon de souligner qu'un outil tel que WATT ne remplace pas l'être humain, mais l'épaule dans son travail.

[8] C'est d'ailleurs ce que fait Microsoft pour bon nombre de ses articles techniques. Voir par exemple http://support.microsoft.com/kb/315939/fr.

Figure 4.1
Exemple de modification du site Web d'EC incluant une notice de traduction automatique (en haut à droite, pointée par la flèche). Le texte a été extrait de la page http://www.weatheroffice.gc.ca/warnings/report_e.html?qc18, qui est une traduction humaine (cette figure ne cherche qu'à illustrer l'artifice graphique de notice).

En outre, si l'on désire que le prototype fonctionne le mieux possible, il est essentiel d'explorer deux stratégies d'amélioration, qui sont évoquées dans ce rapport.

Rectifications du prototype La première consiste à investir des efforts supplémentaires dans le raffinement du prototype. Maintenant que nous avons en main une typologie des erreurs commises par le moteur durant la traduction (voir la section 3.4), il est naturel de tenter de les éliminer. Conséquemment, à la section 3.4.2, on montre que des réfections à moyen terme pourraient atténuer la moitié des erreurs commises par WATT. Si l'on suit notre recommandation de faire de WATT un outil de traduction qui aide le traducteur, alors les atténuations des erreurs possibles ici seraient autant de simplifications du travail de l'expert en révision.

Assainissement des sources La section 3.4.3 suggère que l'élimination des coquilles et l'uniformisation de certains éléments de la source dès la saisie contribueraient à améliorer

le prototype. Cet effort, combiné aux rectifications proposées au paragraphe précédent, pourrait améliorer les qualités des sorties de façon décisive. De plus, on se trouve à épurer les sources qui seront diffusées dans la langue originale des bulletins.

En fin de compte, il nous apparaît que la chaîne de diffusion des informations météorologiques profiterait grandement d'une intégration solide entre, d'une part, un système d'aide à la saisie des avertissements météo et, d'autre part, le système WATT amélioré, en aval. Le système d'aide à la saisie faciliterait l'entrée des avertissements, normaliserait le vocabulaire, la grammaire, les unités de mesure et de temps, afin de présenter à WATT une entrée irréprochable dans un vocabulaire connu, et donc plus facilement traduisible. Si toutefois WATT a de la difficulté à traduire une phrase, il serait en mesure d'en aviser tout de suite le météorologue dans son interface de saisie, qui pourrait alors apporter des corrections à son texte source (ou bien outrepasser WATT complètement et demander une traduction humaine, pour un bulletin plus difficile).

Dans la foulée, les intrants du système seraient épurés, et le consommateur canadien y trouverait son compte lorsqu'il lirait les bulletins dans leur langue originale.

Un système solidement intégré ainsi facilite donc la saisie, la traduction automatique et humaine, la révision éventuelle des bulletins et favorise une langue de qualité à la fois dans les intrants et dans les sorties du système, peu importe le média de diffusion.

5 Bibliographie

1. Thomas Leplus, Philippe Langlais and Guy Lapalme. *Weather Report Translation using a Translation Memory*. Machine Translation: from Real Users to Research: 6th Conference of AMTA, series. Lecture Notes in AI #3265, p. 154-163, Washington, septembre 2004
2. Philippe Langlais, Simona Gandrabur, Thomas Leplus and Guy Lapalme. *The Long-Term Forecast for Weather Bulletin Translation. Machine Translation,* vol. 19, number. 1, p. 83-112, mars 2005.
3. *Rapport d'évaluation linguistique des sorties machine produites par l'outil de traduction automatique du RALI pour Environnement Canada*, Bureau de la traduction, Juillet 2010.
4. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). *"BLEU: a method for automatic evaluation of machine translation"* in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318
5. Callison-Burch, C., Osborne, M. and Koehn, P. (2006) *"Re-evaluating the Role of BLEU in Machine Translation Research"* in 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006 p. 249–256
6. D. Vilar, J. Xu, L. F. D'Haro and H. Ney: *Error Analysis of Statistical Machine Translation Output*. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 697-702, Genova, Italy, May 2006.
7. J. Cohen, *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, vol. 20, numéro 1, p. 37–46, 1960.
8. B. Di Eugenio et M. Glass, *The kappa statistic:a second look.* Computational Linguistics, vol. 30, numéro 1, p. 95–101, 2004.
9. *Summer warnings document analyser.doc*, Bureau de la traduction, juillet 2010.
10. A. Stolcke, *Srilm – An extensible language modeling toolkit*. Actes de ICSLP, vol. 2 (2002), p. 901–904.

Annexe A
Exemple de paire de bulletins d'alerte météorologique

Pour aligner les versions anglaises et françaises des bulletins, des programmes Python ont été écrits. Puisque ni les noms de fichiers ni les en-têtes de fichiers (1res lignes) ne semblent permettre l'alignement, nous nous sommes basés sur le champ `MSG_SEQID`, qui permettait l'alignement du plus grand nombre de bulletins.

Seulement les bulletins dont le champ `BULLETIN_TEXT_STATUS` indiquait `complete` ont été considérés, et seulement lorsque du texte était présent dans la discussion (`OMNI_DISCUSSION`, ou, si absent, `EVENT_DISCUSSION`).

Nous avons fait ce choix après avoir examiné plusieurs champs texte des bulletins, et éliminé ceux qui étaient soit inutilisés, soit absents dans un membre de la paire de bulletins, soit contenant du texte stéréotypés sans intérêt pour l'évaluation. Les champs non considérés sont `OMNI_CLOSING_TEXT`, `EVENT_NAME`, `OMNI_INEFFECT_TEXT`, `OMNI_ENDED_TEXT`, `EVENT_CLOSING_TEXT` et `EVENT_STANDARD_TEXT`. Le grand nombre de ces champs aurait complexifié d'autant le travail d'extraction du texte, à plus forte raison parce qu'il nous est difficile de comprendre l'utilité et les particularités de chacun.

Pour déterminer la province ou le territoire, nous avons compilé la liste des 23 `AREA_NAME` différents et les avons liés à leur province ou territoire respectif. Par exemple, `qikiqtaaluk area` est associé au Nunavut.

Un exemple de paire de bulletins alignés est présenté aux pages suivantes.

ACC-MP71WG3004230601	ACC-MP11WG3004230657
MPCN71 CWWG 042345 HEADER=ww AREA=71 COVERAGE=71 AREA_NAME=le sud du Manitoba COVERAGE_NAME=le sud du Manitoba OFFICE=cwwg MSG_SEQID=cwwg000883 MSG_LANGUAGE=français SOURCE_APPLICATION=bullprep SOURCE_SERVER=opsr1 OMNI_ISSUETIME=201006042339 OMNI_LCL_FULL_ISSUETIME=18h39 HAC le vendredi 4 juin 2010 OMNI_TIMEZONE=HAC OMNI_COR_LEVEL=- OMNI_FILENAME=ww71_omni_wg OMNI_INITIALS= OMNI_DISCUSSION=" L'humidité issue des dernières pluies et le réchauffement diurne créent des conditions propices à la formation d'orages. Certains d'entre eux pourraient devenir violents et s'accompagner de grosse grêle, de vents forts et de pluie torrentielle de près de 30 mm par endroits. On a signale de nombreux nuages en entonnoir sur le sud du Manitoba. Ce type de nuage en entonnoir se forme à partir de gros cumulus ou de très faibles orages et n'à normalement pas suffisamment de force pour atteindre le sol. Environnement Canada continue de surveiller la situation de près. Restez à l'écoute de vos médias locaux ou de radio météo pour les mises à jour à venir. " OMNI_CLOSING_TEXT=" Veuillez consulter les dernières prévisions publiques pour plus de précisions. " BULLETIN_TEXT_STATUS=complète OMNI_STANDARD_TEXT=" " BULLETIN_TEXT_CASE=lower EVENT_ID=877cwwg EVENT_NAME=veille d'orages violents EVENT_KIND=orage violent EVENT_TYPE=veille EVENT_RGN_CONFIG=fp71wg.Rgn EVENT_DURATION=600 EVENT_NAME_MODIFIER= EVENT_HEADER=ww EVENT_AREA=71	MPCN11 CWWG 042339 HEADER=ww AREA=11 COVERAGE=11 AREA_NAME=Southern Manitoba COVERAGE_NAME=Southern Manitoba OFFICE=cwwg MSG_SEQID=cwwg000883 MSG_LANGUAGE=english SOURCE_APPLICATION=bullprep SOURCE_SERVER=opsr1 OMNI_ISSUETIME=201006042339 OMNI_LCL_FULL_ISSUETIME=6:39 PM CDT Friday 4 June 2010 OMNI_TIMEZONE=CDT OMNI_COR_LEVEL=- OMNI_FILENAME=ww11_omni_wg OMNI_INITIALS= OMNI_DISCUSSION=" The moisture from the recent rains combined with daytime heating are creating conditions favourable for the development of thunderstorms. Some of these storms have the potential to become severe..With large hail..Strong winds..And possible localized downpours of around 30 mm. There have been numerous reports of funnel clouds over Southern Manitoba. These types of funnel clouds form out of large cumulus clouds or very weak thunderstorms and normally do not have the energy to reach the ground. Environment Canada continues to monitor the situation closely. Please continue to monitor your local media or weatheradio for further updates. " OMNI_CLOSING_TEXT=" Please refer to the latest public forecasts for further details. " BULLETIN_TEXT_STATUS=complete BULLETIN_TEXT_CASE=lower EVENT_ID=877cwwg EVENT_NAME=severe thunderstorm watch EVENT_KIND=severe thunderstorm EVENT_TYPE=watch EVENT_RGN_CONFIG=fp11wg.Rgn EVENT_DURATION=600 EVENT_NAME_MODIFIER= EVENT_HEADER=ww EVENT_AREA=11 EVENT_COVERAGE=11 EVENT_AREA_NAME=Southern Manitoba EVENT_COVERAGE_NAME=Southern Manitoba EVENT_OFFICE=cwwg EVENT_ISSUETIME=201006042339

EVENT_COVERAGE=71
EVENT_AREA_NAME=le sud du Manitoba
EVENT_COVERAGE_NAME=le sud du Manitoba
EVENT_OFFICE=cwwg
EVENT_ISSUETIME=201006042339
EVENT_LCL_FULL_ISSUETIME=18h39 HAC le vendredi 4 juin 2010
EVENT_TIMEZONE=HAC
EVENT_COR_LEVEL=-
EVENT_FILENAME=ww71_omni_wg
EVENT_STACKED_FILENAME=ww_omni
EVENT_STATUS=updated
EVENT_PRINT_REGION=false
OMNI_INEFFECT_TEXT="
Risque d'orages violents ce soir.

Ceci est une mise en garde quant à la formation possible d'orages
violents accompagnés de gros grêlons et de vents destructeurs.

Veuillez surveiller les bulletins météorologiques et leurs mises à
jour. Si le temps devient menaçant, prenez sans délai les précautions
qui s'imposent.
"
EVENT_PRIORITY=low
EVENT_SELECTABLE=true
EVENT_SAME_CODE=sva
OMNI_ENDED_TEXT="
"
EVENT_CLOSING_TEXT="
"
EVENT_STANDARD_TEXT="
"
EVENT_DISCUSSION="
"
EVENT_EC_CODE=stv
REGION=Selkirk - Gimli - Stonewall - Woodlands - Eriksdale
ZONE=Selkirk - Gimli - Stonewall –

[...]

SUBREGION=
SUBZONE=
STATUS=ended
End/mpcn71

EVENT_LCL_FULL_ISSUETIME=6:39 PM CDT Friday 4 June 2010
EVENT_TIMEZONE=CDT
EVENT_COR_LEVEL=-
EVENT_FILENAME=ww11_omni_wg
EVENT_STACKED_FILENAME=ww_omni
EVENT_STATUS=updated
EVENT_PRINT_REGION=false
OMNI_INEFFECT_TEXT="
Severe thunderstorms are possible this evening.

This is an alert to the potential development of severe thunderstorms with large hail and damaging winds.

Monitor weather conditions..Listen for updated statements. If threatening weather approaches take immediate safety precautions.
"
EVENT_PRIORITY=low
EVENT_SELECTABLE=true
EVENT_SAME_CODE=sva
EVENT_EC_CODE=stv
REGION=Selkirk - Gimli - Stonewall - Woodlands - Eriksdale
ZONE=Selkirk - Gimli - Stonewall - Woodlands - Eriksdale

[...]

SUBREGION=
SUBZONE=
STATUS=ended
END/MPCN11

Annexe B
Page web d'instructions aux annotateurs

Évaluation des traductions d'alertes météorologiques

Merci de nous aider dans le projet de diffusion d'informations environnementales!

Voici quelques instructions afin d'annoter les traductions d'alertes météorologiques.

Utilisation de Word pour annoter

Nous utiliserons Microsoft Word (toutes versions) pour ajouter des annotations expliquant la qualité perçue des traductions. Les annotations prendront la forme de simples commentaires Word dont le contenu est préétabli.

Les documents à annoter se trouvent ici : documents d'annotation. Il y en a deux par annotateur.

Un document de test est disponible ici.

Activez les macros dans Word pour que l'annotation fonctionne. Si vous rencontrez des problèmes, voici comment activer les macros sous Word 2007 pour Windows, Word 2003 pour Windows (choisissez le niveau de sécurité moyen en 2003), et Word 2004 pour Macintosh.

Remplissage de la première page

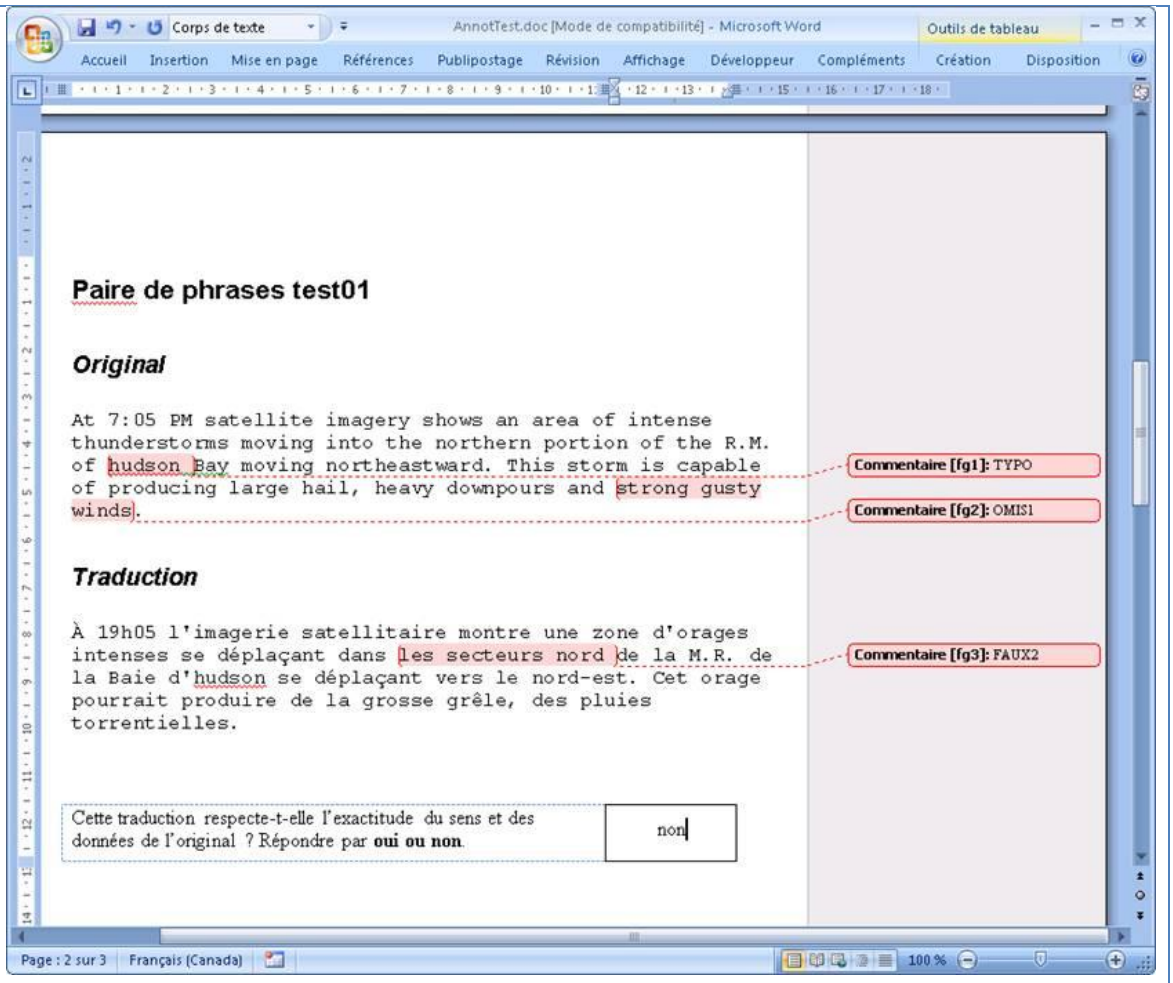
La première page contient quelques liens pour simplifier votre travail.

Ajoutez-y vos initiales.

Annotations de la qualité des traductions

Pour chaque paire de phrases, les boutons activés par les macros de Word (en haut de l'image) permettent l'ajout de commentaires dénotant des erreurs dans la paire de phrases à évaluer.

La typologie d'erreur est disponible ici.



Vous n'avez pas à taper quoi que ce soit, il vous suffit d'utiliser les boutons d'annotations.

a. Commencez par lire la phrase source pour y repérer d'éventuelles erreurs. Si la phrase source contient des fautes de langue ou de typographie, on ne pénalise pas une traduction erronée *de ce passage*.

b. Lisez ensuite la traduction et pour y marquer les fautes de langue ou de typographie.

c. Comparez ensuite la traduction à l'original pour évaluer la qualité du transfert du contenu. Pour marquer une omission dans la traduction, sélectionnez le passage dans l'original qui a été omis et marquez-le avec une omission (mineure ou majeure).

d. Enfin, relisez complètement l'original et la traduction et répondez à la question finale sur l'exactitude du sens. Pensez-vous que le Canadien moyen qui lit la traduction de l'alerte météo est aussi bien informé que s'il avait pu lire l'original ?

Vous pouvez annuler vos annotations comme vous annuleriez toute autre opération dans Word, ou en supprimant simplement le commentaire que vous voulez amender.

Annexe C
Capture d'écran du site d'EC

Un problème de ponctuation typique est illustré dans cette capture d'écran du site d'EC, récupéré le 17 novembre 2010, à l'url

http://www.weatheroffice.gc.ca/warnings/report_e.html?nu24

Noter le « .. » séparant les deux phrases.

Annexe D
Erreurs identifiées dans les sources

Voici une copie des phrases dans la langue source pour lesquelles au moins une erreur a été repérée par au moins un annotateur. On ne montre pas ici toutes les annotations, mais toutes les phrases sources avec une erreur sont identifiées. Les annotations sont encadrées et sur fond bleu.

A few of these thunderstorms could become severe giving nickel sized hail, gusty winds and locally heavy rain.
A low pressure system off the coast of Labrador is forecast to track north through the Davis strait and pass south of Qikiqtarjuaq on Friday.
A slow moving low pressure system west of Great Slave lake will continue to bring rain to the Fort Liard region today.
A warm, humid airmass over much of southern and Eastern Saskatchewan this afternoon.
About 20 millimetres of rain has already fallen. Another 15 to 25 mm of rainfall is expected in Old Crow before the disturbance begins to weaken this evening.
As a result..Patchy frost will form over low lying inland areas where ground temperatures will be near zero.
At 2 PM radar indicates isolated severe thunderstorms developing over Eastern Ontario.
attention...
Certains de ces orages sont exceptionnellement intenses et l'un d'eux pourrait produire une tornade.
Ces orages produiront des rafales de 90 km/h ou plus - de la grêle de 2 centimètres ou plus - de fortes pluies - et de nombreux éclairs.
Des lignes orageuses affectent l'ouest de la région de Québec et d'autres de st-zenon à prevost dans les Laurentides Ces orages produiront des rafales de 90 km/h ou plus - de la grêle de 2 cm ou plus - de fortes pluies - et de nombreux éclairs.
Des orages affectent Montréal et se dirigent vers St-Hyacinthe.
Des orages de forte intensité sur le secteur du mont Laurier se dirigent vers le sud-est à près de 40 km/h.
Further rainfall amounts of 15 to 20 mm can be expected with higher amounts possible in upslope flow over higher terrain.
Ils devraient traverser le secteur de la minerve - rivière rounge dans la prochaine heure.

Individual storms are moving to the northeast and are capable of torrential downpours..Wind gusts to 90 km/h and intense lightning.
Intense lightning..Torrential downpours giving 25 to 50 mm in under an hour..Wind gusts to 90 km/h and 2 cm hail are possible in these thunderstorms.
Local amounts of over 50 millimetres of rain have already been reported with this system.
Nearly 3 hours of torrential rainfall associated with thunderstorms has been affecting an area between st. Leonard and Edmundston.
Please continue to monitor your local media or weatheradio for further updates.
Some of these storms have the potential to become severe..With large hail..Strong winds..And possible localized downpours of around 30 mm.
Torrential downpours in excess of 25 millimetres per hour are possible as the line moves through Southwestern Nova Scotia.
Weather conditons are no longer favourable to produce funnel clouds.
Weather radar at 8;20 PM EDT shows a line of strong to severe thunderstorms extending from 30 km southeast of Dryden towards Emo.