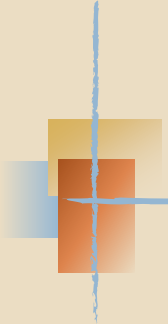


Vers une automatisation utilitaire de l'évaluation des traductions

Éric Poirier, Université du Québec à Trois-Rivières
Colloque « Vers une robotique du traduire ? » 2021
Strasbourg, 30 sept.-1^{er} oct. 2021



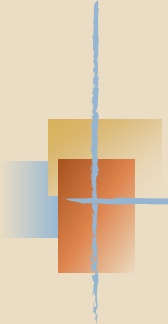
Objectifs

Possibilités offertes par la science des données traductologiques dans l'évaluation des traductions

Optimisation et objectivisation de l'évaluation des traductions

Avantages pour la formation et l'approfondissement des connaissances

Éléments automatisables de l'évaluation des traductions et du CQ?



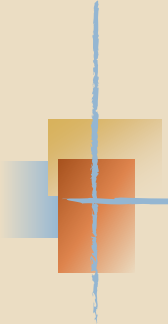
Contexte et environnement

Traduire c'est un art et une « technique »

- Double fidélité au TS et aux « procédés » récurrents
- Évaluation des traductions même réalité (CQ et révision)

Essor de la TA, données traductologiques massives

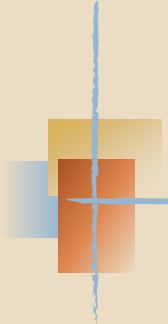
- panoplie d'outils producteurs de données traductologiques : mémoires, gestionnaires de corpus parallèles, concordanciers, banques de données terminologiques, etc.
- Éléments automatisables de l'évaluation des traductions et du CQ



Trois niveaux de qualité des traductions

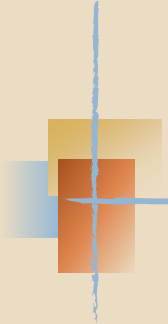
- a) Compréhension générale, des éléments essentiels du message (= gisting), niveau minimal de qualité de la traduction automatique (gratuite et très rapide)
- b) Communication interne (qualité suffisante et acceptable) = good enough quality – traduction partiellement automatisée – postédition
- c) Qualité optimale (publication, communication externe) = quality similar or equal to human translation; traduction manuelle intégrale

Source : TAUS et CNGL (2010).



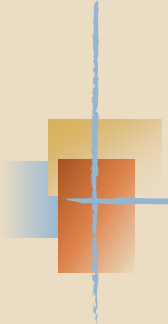
Trois modes de production

- a) Traduction automatique (TA)
- b) Optimisation des correspondances des mémoires de traduction, recours à la TA pour les paires de segments sous le seuil – postédition
- c) Traduction manuelle intégrale – TA non intégrée utilisée comme élément de comparaison



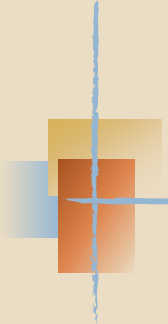
Trois formats de données

- a) Bi-textes simples (paires de segments)
- b) Triplets – bi-textes et segments cibles révisés
- c) Quadruplet et plus : segments cibles révisés et annotations : TALN (étiquetage POS, analyse syntaxique, analyse de dépendance)



Du CQ à l'évaluation – plan d'ensemble

- Contrôle de qualité (2 types)
 - démarche descriptive, identification et découverte de « métriques », repérage des paires de segments susceptibles de contenir des erreurs (échantillonnages)
- Révision professionnelle
 - démarche justificative, identification et description des “règles” ou techniques de traduction, repérage des erreurs de traduction
- Évaluation des traductions
 - démarche explicative, découverte de métriques et de règles ou techniques de traduction, définition opératoire des erreurs de traduction



Précision informative des traductions

Calcul et comparaison du volume informatif

Ratio PIT simple ($7 \div 7 = 1,0$)

Traduire, c'est produire avec des moyens différents des effets analogues (7) = To translate is to produce similar effects with different means (7)

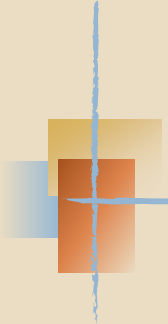
Inférieur à 1,0 = décalage négatif, supérieur à 1,0 décalage positif



Analyse manuelle – décalages

Prix Montréal Centre-ville 2021 Fr -> Ang

Description en français	Description en anglais
Prix ¹ Hommage ² : reconnaît ³ la contribution ⁴ hors pair ⁵ d'une personnalité ⁶ ayant ⁰ contribué ⁷ de façon ⁸ exemplaire ⁹ à la relance ¹⁰ du centre-ville ¹¹ .	The Tribute ¹ Award ² recognizes ³ an individual ⁴ who has ⁰ outstandingly ⁵ contributed ⁶ to the relaunch ⁷ of the downtown ⁸ core ⁹ .
5 = peut être aligné à outstandingly ⁵ 8,9 = peut être aligné à outstandingly ⁵ – en redondance avec 5 11 = composé aligné à un terme complexe ⁹	



La distance euclidienne

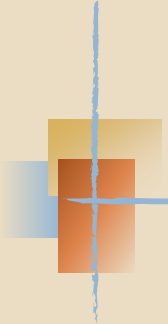
Calcul de la « distance » nominale qui « sépare » le segment cible du segment source (2 ou 3 dimensions)

p = nombre de caractères, q = nombre de mots lexicaux
(volume informatif) Poirier 2021

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Concept intéressant pour les groupes de données qui possèdent des valeurs corrélées, ce qui est le cas des traductions

Pas de norme numérique : DE de 1 à 300



Analyse outillée

Algorithme PIT (précision de l'information traduite),
version 2.0 (source : Poirier, 2019, 2021)

3 modules

- saisie du texte (format csv), unité par défaut
paragraphe, étiquetage POS module spaCy
- Calcul du volume informatif et du ratio PIT, DE
- Criblage des décalages informatifs positifs et négatifs –
liste des paires de segments avec décalages

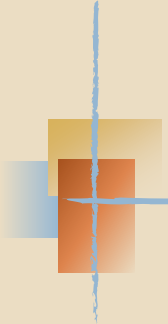


Données empiriques sur la qualité des traductions

Analyse outillée des décalages informatifs dans l'évaluation de la qualité de la traduction de l'ensemble des lois codifiées du Québec (Poirier et Roy, à paraître)

Étude sur un corpus de 515 fichiers :

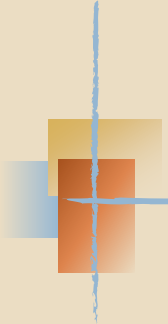
- 196 544 unités d'analyse ou paires de segments
- 12 916 602 mots, soit 6 974 679 en français (51,72 %) et 6 510 232 en anglais (48,28 %)
- 6 719 110 mots lexicaux, soit 3 409 147 en français (50,7 %) et 3 309 963 en anglais (49,3 %) ce qui donne un ratio PIT (précision de l'information traduite) moyen global de 0,9709
- 78 552 455 caractères soit 40 149 565 en français (51,11 %) et 38 402 890 en anglais (48,89 %).



Deux modes de traduction

Traductions isomorphes (81 % des traductions dans le corpus LégisQuébec)

Traductions hétéromorphes (19 % des traductions dans le corpus LégisQuébec)

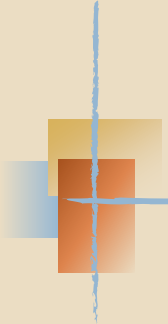


Outils de CQ formel 1

Outils axés sur le TALN (pas encore sur la compréhension des langues naturelles – NLU)

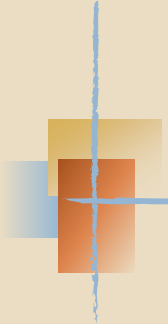
CQ formel : la qualité des alignements

- Outil : Ratio de précision de l'information traduite (PIT);
- Volume informatif = 0 (données vides ou sans information) – omissions de segments



Outils de CQ formel 2

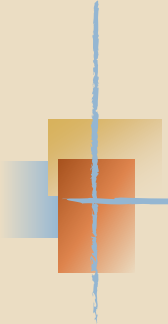
- Ratio PIT à X écarts-types (absolu) : très fortement hétéromorphes
- Positifs : problèmes de mise à jour des textes, corpus (LégisQuébec : énumération de villes qui ont été vraisemblablement retirées)
- Négatifs : omissions partielles attribuables à des mésalignements de phrases asymétriques 2:1 (paragraphe 5:4, 5:3, 3:2, etc.)



Outils de CQ substantiel 1

Ratio PIT – échantillonnage, RAS – Paires de segments isomorphes (80 %) méthodes directes

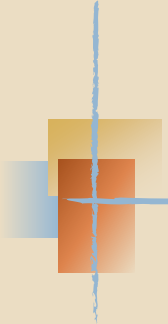
- Formellement des traductions correctes : comparaison automatique avec la TA; examen manuel des différences
- Analyses automatiques internes : arbres de dépendances ; problème de granularité des analyses pour les décalages « internes »



Outils de CQ substantiel 2

Ratio PIT – échantillonnage, RAS – Paires de segments hétéromorphes (19 %)

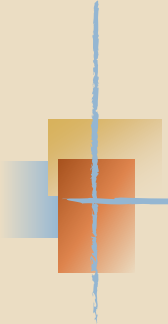
- Paires positives : ajouts possibles (dépouillement manuel) – comparaison automatique avec la TA; décalages informatifs
- Paires négatives : omissions possibles (dépouillement manuel) – comparaison automatique avec la TA; décalages informatifs



Outils de révision professionnelle 1

Paires de segments isomorphes

- Validations des traductions par listes positives ou négatives (terminologies courantes, à éviter)
- FS, NS -> fluidité en LC et comparaison avec TA
- CS -> exactitude par comparaison de la représentation du sens en LS et en LC : outils du natural language understanding, et comparaison avec TA



Outils de révision professionnelle 2

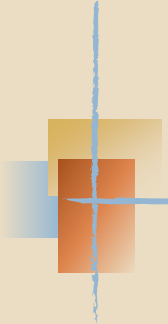
Paires de segments hétéromorphes

Positifs

- redondances fautives en LC (surtraductions ou pléonasmes) : fluidité de la LC et comparaison manuelle avec la TA

Négatifs

- ellipses fautives en LC (sous-traductions) : fluidité de la LC et comparaison manuelle avec la TA



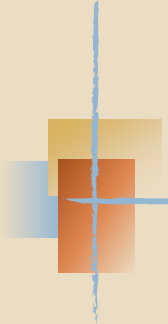
Outils de révision professionnelle 3

Comme pour les paires isomorphes

Validations des traductions par listes positives ou négatives (terminologies courantes, à éviter)

FS, NS -> fluidité en LC et comparaison avec TA

CS -> exactitude par comparaison de la représentation du sens en LS et en LC : outils du natural language understanding, et comparaison avec TA

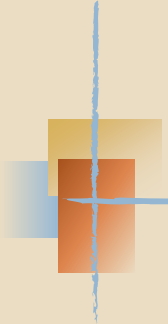


Outils d'évaluation des traductions 1

Paires isomorphes

Différences entre traduction mot à mot et traduction littérales

Anisomorphisme des langues et isomorphisme de traduction



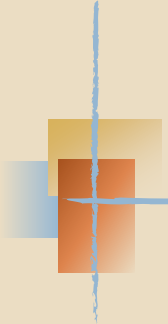
Outils d'évaluation des traductions 2

Paires hétéromorphes

Techniques de traduction : définition d'équivalences discursives (anisonomorphisme des langues), traduction des entités nommées, dépronominalisation, stylistique comparée (langues de spécialité), traduction de la phraséologie, etc.

Comparaison des redondances et ellipses entre les langues

Définitions opératoires des erreurs de traduction



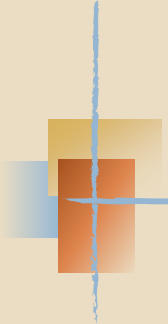
Conclusion

Possibilités d'automatiser certains éléments du CQ, de la révision et de l'évaluation des traductions grâce à des fonctions utilitaires qui procèdent par échantillonnage/cribleage.

Nouvelles définitions des erreurs de transfert linguistique cardinales : ajout, omission, faux sens, non-sens, contresens et objectivisation de ces erreurs

Comparaison ratio PIT et DE

Seuils de DE



Merci de votre attention!

Questions?

Eric.Poirier@uqtr.ca