Traduire automatiquement des articles dans les sciences dites dures

Nicolas BACAËR

Institut de recherche pour le développement

nicolas.bacaer@ird.fr

Résumé

On présente quatre applications de la traduction automatique dans les sciences dites dures : la messagerie électronique, la lecture d'articles, la traduction à partir du français et la traduction vers le français. Les questions d'ordre typographique jouent un rôle important à cause de la présence de nombreuses formules mathématiques.

Soyez résolus de ne servir plus, et vous voilà libres.
La Boétie (La Servitude volontaire)

1. La messagerie électronique

Avant d'en venir au point plus spécifique de la traduction automatique de textes contenant des formules mathématiques, il n'est peut-être pas inutile de rappeler qu'il est désormais relativement facile d'envoyer systématiquement des courriels aux scientifiques étrangers dans la langue du destinataire. Pour cela, la double traduction est particulièrement utile. Par exemple, pour traduire du français vers l'italien ou vers le japonais (langues supposées inconnues), on demande la traduction à un traducteur automatique (par exemple DeepL ou Google Traduction) puis on retraduit soit vers le français, soit vers une

troisième langue, par exemple l'allemand (supposé connu). Si le résultat à l'arrivée correspond au message initial, il est raisonnable de supposer que la traduction intermédiaire est correcte. Sinon, il faut modifier légèrement le message initial jusqu'à ce que le résultat de la double traduction convienne. Certes les traducteurs automatiques ne couvrent qu'une centaine de langues parmi les milliers qui existent. Certes les messages doivent rester assez simples. Certes la qualité de la traduction est variable selon la langue cible. Mais c'est déjà un pas important vers une plus grande diversité linguistique.

2. Lire des articles scientifiques

C'est bien connu : il serait obligatoire de connaître l'anglais, ne serait-ce que pour s'informer de la littérature scientifique dans son domaine. Si cela a pu être vrai à une époque récente, ce n'est plus le cas en 2020. En effet, demandons-nous ce qu'il est possible de lire en français avec le navigateur **Chrome**, qui intègre le traducteur de Google. Considérons les deux plus gros éditeurs de revues scientifiques, Elsevier et Springer, avec chacun plus de deux mille revues, principalement en anglais. Depuis quelques années, les articles ne sont plus seulement publiés en PDF (la traduction automatique fonctionne très mal avec ce format) mais aussi en HTML. Avec les articles de **Elsevier**, il suffit donc d'aller sur la version HTML, de cliquer sur le bouton droit de la souris, de demander la traduction en français et de faire défiler l'article jusqu'au bout car le traducteur automatique ne traduit que ce qui est visible à l'écran. On peut éventuellement sauvegarder la traduction en imprimant la page vers un fichier PDF. On bénéficie au passage du fait que la traduction automatique entre l'anglais et le français marche assez bien. Il y aura probablement certains mots techniques qui seront mal traduits. Mais si l'article est dans votre spécialité, vous aurez vite fait de deviner de quoi il s'agit.

Avec les articles HTML de l'éditeur **Springer**, une étape supplémentaire est nécessaire si l'article contient des formules mathématiques. Il faut cliquer avec le bouton droit de la souris sur une

formule quelconque, choisir « Math settings » puis « Math renderer » puis « SVG » (abréviation de « Scalable Vector Graphics » ou « graphique vectoriel adaptable »). Toutes les formules du document passent alors en mode SVG et il suffit de poursuivre sur le navigateur Chrome comme avec les articles de Elsevier. On remarque au passage que Elsevier et Springer utilisent tous les deux la bibliothèque logicielle MathJax pour l'affichage des formules mathématiques : elle permet notamment l'affichage sur une page HTML de formules écrites dans le langage LaTeX.

Après ces deux éditeurs « internationaux », bien connus pour profiter de leur oligopole pour facturer leurs abonnements à des prix exorbitants aux bibliothèques universitaires, considérons les principaux éditeurs basés en France dans les sciences dites dures (hors médecine):

- EDP Sciences avec environ 70 revues, récemment passé sous pavillon chinois, et qui a reçu d'importantes subventions publiques pour mettre en accès libre une partie de ses articles ;
- le Centre Mersenne à Grenoble avec une vingtaine de revues, toutes en accès libre sans frais de publication ;
- les éditions du Muséum national d'histoire naturelle avec une dizaine de revues, également en accès libre sans frais de publication.

EDP Sciences publie la plupart de ses articles en HTML, sauf apparemment pour quelques revues mathématiques où seul le PDF est disponible. La traduction automatique en utilisant le navigateur Chrome fonctionne bien mais pour une raison un peu différente de précédemment : les formules mathématiques y sont affichées sous forme d'images.

Pour le moment, seules trois revues du **Centre Mersenne** sont disponibles en HTML (avec un affichage qui utilise MathJax) et donc traduisibles automatiquement, à savoir les sections de biologie, de

chimie et de géosciences des Comptes rendus de l'Académie des sciences.

Les revues du **Muséum national d'histoire naturelle** sont publiées en PDF en accès libre. Mais curieusement, les articles sont aussi disponibles en HTML sur la plateforme BioOne à accès payant.

Il existe d'autres éditeurs en France avec un très petit nombre de revues en sciences dites dures, comme la Société mathématique de France ou Lavoisier (qui a récemment cédé plusieurs revues).

Pour résumer cette section, disons que l'on peut désormais lire en français par milliers des traductions automatiques d'articles dans chaque discipline, de quoi rester assez bien informé.

3. Traduire à partir du français

On peut aussi écrire et publier ses travaux de recherche en français sans craindre un manque de diffusion. Il suffit de déposer quelques traductions automatiques, éventuellement post-éditées, sur un site d'archives comme HAL (https://hal.archives-ouvertes.fr/), en indiquant bien la référence originale. En pratique, du point de vue carriériste, il faut quand même se désintoxiquer du « prestige » des journaux qui n'acceptent pas les manuscrits en français. Il faut trouver une revue dans son domaine qui accepte les manuscrits en français ; voir par exemple la liste de revues http://www.ummisco.ird.fr/perso/bacaer/liens.html.

Expliquons comment traduire automatiquement un article rédigé en français. Si l'article a été rédigé avec « **Word** » ou avec un logiciel du même genre, le plus simple est sans doute de sauvegarder le fichier source en HTML, de l'afficher dans Chrome et de le traduire en entier. On peut alors sauvegarder la traduction en HTML, ce qui permet la post-édition. Pour profiter de la qualité semble-t-il un peu meilleure de DeepL, du moins pour la dizaine de langues disponibles, on peut aussi traduire paragraphe par paragraphe (ou deux par deux) en copiant le texte. La version téléchargée permet de traduire par paquets de 5000

caractères, contre 2000 pour la version en ligne.

Venons-en au cas nettement plus délicat des fichiers écrits en **LaTeX**, comme c'est presque toujours le cas en mathématiques et en physique. La première étape consiste à sauvegarder le fichier LaTeX au format HTML et à modifier le début du fichier pour appeler la bibliothèque MathJax. On peut aller voir sur un exemple comment c'est rédigé, cf.

https://hal.archives-ouvertes.fr/hal-02509142v7/file/2020MMNP.html

(demander au navigateur d'afficher le fichier source). Il faut aussi faire un certain nombre de changements pour le titre, les sections, les listes, la bibliographie, les figures, les références aux équations et à la bibliographie, les accents : il faut remplacer la syntaxe LaTeX par son équivalent en HTML, ce qui peut se faire dans certains cas de manière semi-automatique avec la fonction « remplacer » d'un éditeur de texte. On notera au passage que le format des figures peut nécessiter une conversion, de PostScript en PNG. Il faut aussi écrire \(...\) à la place de $...$, \[...\] à la place de $$...$$ (pour les équations sur une nouvelle ligne), et mettre des espaces avant et après les inégalités strictes > et < .

En écrivant certains passages du texte à l'intérieur d'un environnement mathématique, on peut bloquer la traduction automatique. C'est utile pour la bibliographie, pour les noms des auteurs (la transcription latine d'un nom japonais a peu de chance d'être retraduite en japonais correctement) ou pour la référence à l'article original à placer au début de la traduction.

Une fois tout ce travail terminé, ce qui peut prendre une heure pour un article d'une vingtaine de pages, on peut faire la traduction automatique en utilisant le navigateur Chrome, comme indiqué précédemment. De nombreux exemples se trouvent sur le site

http://www.ummisco.ird.fr/perso/bacaer/,

où la plupart des articles sont disponibles en une dizaine de langues, indiquées par [ar, de, es, it, ja, nl, pt, ru, zh], en plus de la version

française en PDF et en HTML. Avec la version HTML, on peut aussi traduire soi-même automatiquement avec Chrome dans d'autres langues, par exemple en anglais.

Pour améliorer la qualité de la traduction, quel que soit le format du fichier source, on peut relire la traduction dans une ou deux langues que l'on connaît, parmi l'anglais, l'allemand, l'espagnol, l'italien, etc. On peut légèrement modifier le fichier source jusqu'à ce que la traduction d'une phrase soit correcte. Ceci s'appelle de la **pré-édition**. Par exemple en mathématiques, on se rend compte qu'il vaut mieux écrire « opérateur linéaire » que simplement « opérateur », qui est parfois mal traduit. Plus généralement, il faut éviter les pronoms « il » ou « elle » qui renvoient à une phrase précédente, car le traducteur fonctionne phrase par phrase. Or dans certaines langues comme l'allemand, le genre est rarement le même qu'en français. Il arrive aussi que le traducteur automatique bégaie. Il traduit plusieurs fois la même phrase. En coupant les phrases très longues en deux, ou en coupant les paragraphes très longs en deux, on parvient en général à éviter ce bégaiement.

Le traducteur automatique a du mal avec les phrases qui comportent des formules mathématiques en leur milieu : il a tendance à traiter les deux extrémités comme indépendantes. Une astuce consiste à transformer certains éléments de LaTeX vers HTML. Par exemple, on écrira « si M est une matrice positive » et non « si \(M\) est une matrice positive », car ce dernier isole « si » qui peut se retrouver traduit par « yes » en anglais! De même, on écrira « si β est un coefficient négatif » et non « si \(\beta\) est un coefficient négatif ». Une autre technique consiste à modifier un peu l'ordre des mots : « l'intégrale converge si \(m > 0\) » pose moins de problèmes que « si \(m > 0\), l'intégrale converge ». Il n'y a pas besoin de retenir ces bizarreries. Il suffit de constater ce qui se passe et d'y remédier au coup par coup.

Le traducteur automatique de Chrome utilise l'anglais comme langue pivot. Ceci peut conduire à quelques erreurs. Par exemple, « puisque » devient « seit » (depuis) en allemand au lieu de « da » car le

traducteur passe par l'anglais « since » qui a plusieurs sens. On préférera alors « parce que » qui donne « because » et « weil ».

Une fois cette pré-édition opérée, c'est la qualité de la traduction automatique dans toutes les langues qui est sans doute améliorée. On pourra donc distinguer deux fichiers sources : le véritable fichier source sans changement d'une part, et le fichier source légèrement modifié pour améliorer la traduction d'autre part. On peut dès lors s'aventurer à traduire dans des langues que l'on ne peut pas relire. Pour des langues très éloignées du français, comme le japonais, il est peut-être plus sage de se contenter d'une traduction en anglais, qui pose peu de problèmes.

Pour l'arabe, l'instruction DIR="AUTO" (direction automatique) permet bien d'obtenir l'alignement à droite du texte. On remarque aussi que plusieurs « langues de France », dites aussi langues régionales, le catalan, le basque et le corse, figurent dans la liste des langues proposées par Google.

Pour corriger certaines erreurs de traduction qui subsistent malgré le travail de pré-édition, on peut enregistrer la traduction de Chrome comme un fichier HTML, et utiliser la fonction « recherche » de l'éditeur de texte pour trouver les erreurs, car ce fichier est difficile à lire (surtout si MathJax a remplacé chaque passage écrit en LaTeX par des dizaines de lignes). Après les corrections, on ouvre le fichier dans un navigateur et on sauvegarde éventuellement en imprimant vers un PDF. Attention : dans le fichier HTML, l'appel à MathJax doit préciser une cible en **SVG** pour que les formules s'affichent correctement :

```
<script type="text/javascript" id="MathJax-script" async
  src="https://cdn.jsdelivr.net/npm/mathjax@3/es5/tex-svg.js">
```

Cette **post-édition** ne permet évidemment que d'améliorer les traductions dans les langues que l'on connaît. Mais pour un article avec plusieurs coauteurs, on peut profiter des connaissances de chacun : l'un connaîtra l'espagnol, l'autre l'allemand, etc. Par ailleurs, des lecteurs ou des amis peuvent signaler les erreurs de traduction dans leur langue. Les auteurs peuvent alors les corriger, au moins pour les langues avec un

alphabet latin. Pour les autres langues (russe, arabe, chinois...), on y arrive aussi avec un convertisseur en **Unicode**, tel que

http://mylanguages.org/converter.php

Les traductions peuvent être déposées soit sur sa page personnelle, soit en fichiers annexes dans HAL. Dans ce dernier cas, les fichiers sont bien recensés dans des moteurs de recherche spécialisés comme « Google Scholar ».

4. Traduire vers le français

Devant le peu d'empressement des collègues à se lancer dans la traduction automatique de leurs articles, on a essayé de traduire des articles publiés sous une licence CC-BY (*Creative Commons*) :

http://www.ummisco.ird.fr/perso/bacaer/traductions/traductions.html

Ce site présente les traductions en français des articles publiés en anglais à partir de 2020 dans « Comptes Rendus. Chimie », « Comptes Rendus. Biologies » et « Comptes Rendus. Géoscience ». Figurent aussi des traductions de certains articles publiés par EDP Sciences, PLOS, MDPI, *Eurosurveillance*, *Science, Scientific reports*, etc., tous sous licence CC-BY. Certaines traductions sont en PDF non modifiables. D'autres existent aussi en HTML modifiable. Seules quelques traductions ont été post-éditées.

5. Conclusion

La traduction automatique d'articles dans les sciences dites dures est donc possible. La qualité de cette traduction ne peut aller qu'en s'améliorant.

On pourrait au minimum songer à publier en français et à faire une traduction automatique en anglais que l'on vérifie. Ce serait déjà mieux que de ne proposer que l'anglais. C'est ce que fait le physicien Yvan

Castin, cf. par exemple

https://hal.archives-ouvertes.fr/hal-02196152v2 .

Reste à changer les mentalités et vaincre les principaux complexes, notamment le « complexe du colonisateur » et le « complexe du colonisé » (Maurer, 2008). Car bien des scientifiques semblent se satisfaire ou semblent inconscients de l'évolution actuelle vers une uniformisation par l'anglais. Souhaitent-ils, comme ce n'est pas rare chez leurs collègues biologistes (cf. Bacaër, 2019), que toutes les sociétés savantes françaises interdisent à terme les manuscrits en français dans leurs propres revues ? Souhaitent-ils que l'Académie des sciences interdise les manuscrits en français dans ses « Comptes rendus », comme ce fut le cas pendant quelques années avec la section consacrée aux géosciences ? Souhaitent-ils que tous les cours de « master 2 » (voire de « master 1 ») en français soient interdits, comme c'est le cas avec le néerlandais dans la plupart des universités aux Pays-Bas dès la licence et comme c'est en projet au niveau « master » avec l'allemand à l'Université technique de Munich et avec l'italien à l'École polytechnique de Milan ? Souhaitent-ils que les exposés en français soient interdits dans les conférences qui se tiennent en France (actuellement, ceci serait contraire à la législation), dans les séminaires universitaires en France ? La traduction automatique n'est-elle pas un moyen d'inverser cette tendance, en permettant notamment de publier les textes originaux en français et d'envoyer des traductions aux étrangers non francophones ?

Bibliographie

Bacaër, N., 2019, Quelques aspects de la disparition du français dans la recherche scientifique, *Francophonie et innovation à l'université* 1, p. 16-27.

Maurer B., 2008, Pour de nouvelles représentations du français dans la modernité, in : *L'avenir du français*, Éditions des archives contemporaines, Paris, p. 139-141.