

“La traduction automatique doit son succès aux statistiques

24

Thierry Poibeau

est directeur de recherche
au CNRS et directeur adjoint
du laboratoire Lattice
(Langues, textes, traitements
informatiques et cognition).



Longtemps tournés en dérision,
les systèmes de traduction automatique
ont accompli d'incroyables progrès en
quelques années seulement. Leur secret ?
Puissance de calcul et statistiques.

**La traduction automatique est-elle
un domaine de recherche ancien ?
Ou son essor est-il récent, lié à celui
d'internet ?**

Les deux sont vrais. La traduction automatique s'est énormément développée ces dernières années, et je vous expliquerai pourquoi. Mais l'idée d'automatiser le processus de traduction a toujours fasciné. Des philosophes comme Descartes et Leibniz s'interrogeaient déjà sur la possibilité de formaliser le langage pour le dénuier de toute ambiguïté, et donc le rendre facilement traduisible. Pendant la Seconde Guerre mondiale, grâce au déchiffrement automatisé des messages cryptés allemands, les Alliés ont développé des connaissances quantifiées sur les langues, l'organisation ou la fréquence d'emploi de telle ou telle lettre. Après 1945, les pionniers de l'informatique, qui avaient souvent eux-mêmes travaillé pour la défense, ont imaginé appliquer les mêmes techniques à la traduction, en faisant des comptages sur les mots, et non plus sur les lettres. D'autant qu'avec le début de la guerre froide, les Américains se sont penchés sur les grandes masses de textes en russe, concernant par exemple les avancées scientifiques et techniques des Soviétiques, avec l'idée de les indexer et d'y chercher des informations. Les besoins en traduction automatique sont devenus évidents.

Sur quelles difficultés a-t-on buté ?

Le problème, bien connu, est celui de l'ambiguïté du langage naturel : il y a plusieurs sens par mot. Et beaucoup de mots, de l'ordre de 50 000 dans un dictionnaire usuel du français, 400 000 quand on prend toutes les formes possibles des mots, et jusqu'à 1 million ou plus si on y ajoute les noms propres, les mots composés, les termes techniques... L'ambiguïté existe aussi au niveau des phrases. Si vous écrivez : « J'ai vu Clara avec les jumelles », voulez-vous dire que vous avez regardé Clara à l'aide de jumelles ? Ou bien que c'était elle qui portait des jumelles ? Ou bien encore qu'elle était accompagnée de... deux filles ? La somme des possibilités conduit à une explosion combinatoire qu'à l'époque il aurait fallu coder à la main, sans recours à l'électronique : c'était absolument impossible. En réalité, pour arriver à déchiffrer les mots, il faut percevoir l'intention d'emploi derrière, laquelle ne se trouve pas dans le texte, mais dans le contexte. Et le contexte, l'ordinateur ne sait pas ce que c'est. Pour le prédire, il lui serait nécessaire d'avoir, sur le monde, des connaissances de sens commun, ce qu'il n'a pas.

**Qu'entendez-vous par « connaissance
de sens commun » ?**

Ce n'est pas simple de répondre, car les chercheurs en discutent sans toujours le définir. Je vous donne deux exemples : « On peut manger