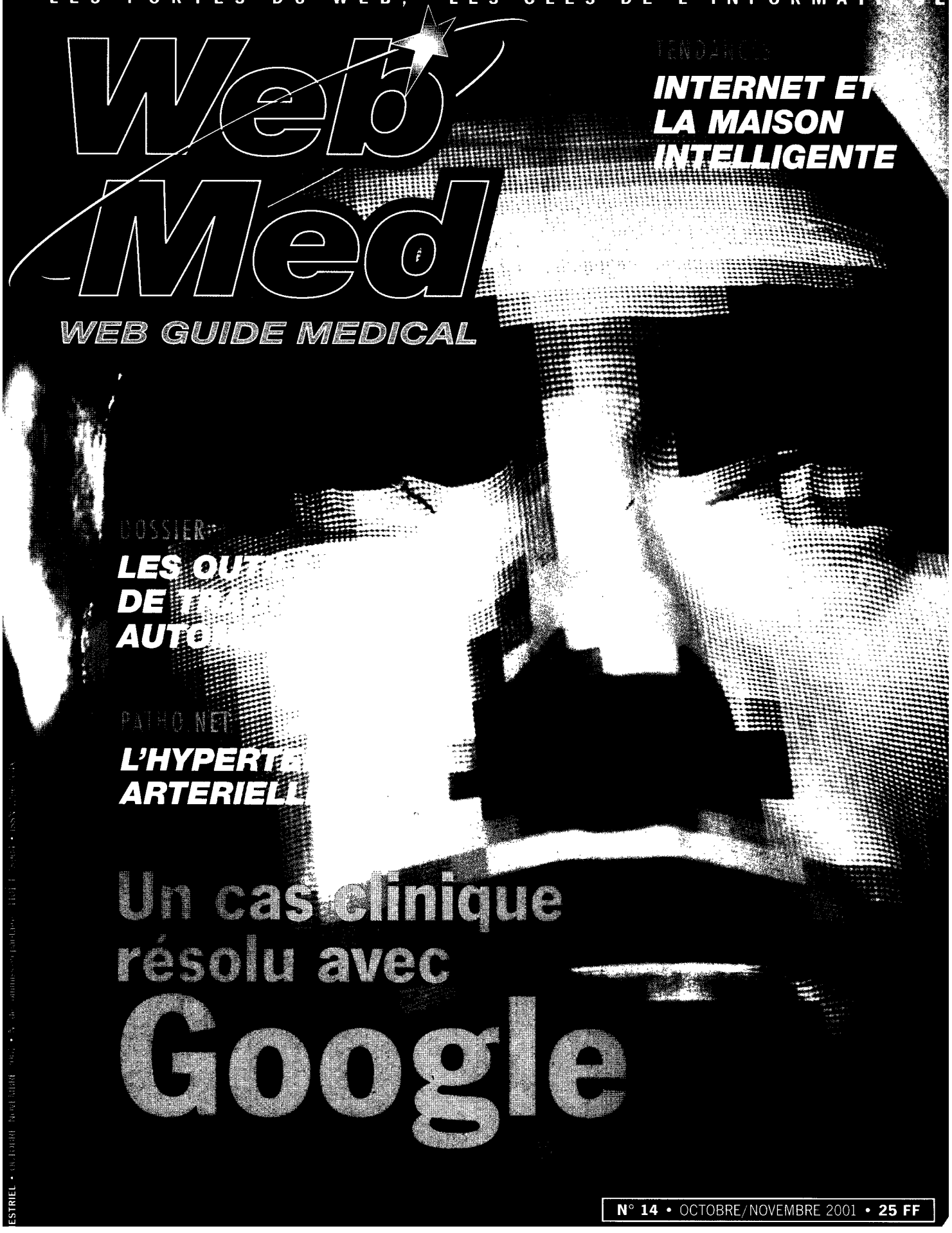
LES PORTES DU WEB, LES CLÉS DE L'INFORMATIQUE

Web Méd

WEB GUIDE MEDICAL



TENDANCES

INTERNET ET LA MAISON INTELLIGENTE

DOSSIER

LES OUT… DE … AUTO…

PATHOLOGIE

L'HYPERT… ARTERIEL…

Un cas clinique résolu avec Google

N° 14 • OCTOBRE/NOVEMBRE 2001 • 25 FF

ESTRIEL • …

Traduction automatique dans l'Internet : panacée ou palliatif ?

Quoi de plus tentant que de faire traduire automatiquement - et gratuitement - le dernier article à succès publié dans *Nature Medicine*, une page web fournie par un moteur de recherche ou un résumé d'article trouvé dans *Medline* ? Mais à quoi faut-il s'attendre au juste ? Le point sur les grandeurs et les misères de la traduction automatique accessibles sur l'Internet.

S. Vandaele et M.-C. L'Homm

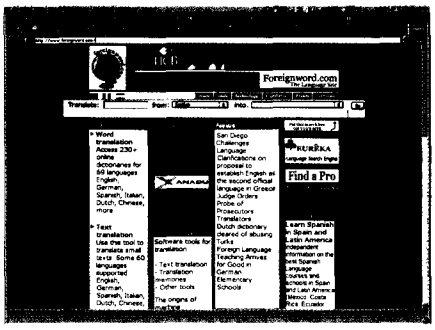
Qui n'a pas rêvé de pouvoir avoir recours à une pierre de Rosette informatique pour décrypter instantanément un texte sybillin écrit dans une autre langue ? Le rêve pourrait sembler devenir réalité, puisque l'on trouve maintenant dans l'Internet un nombre croissant de services de traduction automatique (TA). L'utilisateur soumet un texte rédigé dans une langue, appelée *langue source* (en indiquant une adresse URL ou en copiant le texte dans une fenêtre prévue à cette fin) et récupère la traduction dans la langue qu'il a sélectionnée, appelée *langue cible*. Mais attention : pour des motifs que nous examinerons plus loin, les résultats ne répondent pas nécessairement aux attentes de l'utilisateur. En outre, tous les logiciels proposés n'offrent pas le même niveau de traduction et les particularités de la langue médicale rendent l'emploi de ces logiciels parfois délicat.

Traduction automatique en ligne : accès instantané

Ces services s'adressent à tout le monde (par opposition à d'autres logiciels, commerciaux, conçus spécifiquement pour les traducteurs) et sont offerts, en général, pour permettre aux utilisateurs d'accéder au « contenu » d'un texte rédigé dans une langue étrangère.

Quelles sont les paires de langues proposées ?

Selon les points d'accès et le système utilisé, le nombre de langues source et de langues cible peut être très variable. Certains sites permettent de sélectionner directement une paire de langues, d'autres imposent de choisir la langue source et la langue cible séparément. Pour un même logiciel utilisé, les différents accès n'offrent pas nécessairement le même choix : ainsi, selon qu'on accède à Reverso par l'intermédiaire de la société qui le produit, ou par un site autre, le nombre de paires de langue n'est pas le même.



À quoi faut-il s'attendre ?

L'objectif des systèmes de traduction automatique n'est pas de rendre un texte identique à celui que produirait un traducteur humain. De toute manière, aucun logiciel n'y est encore parvenu. S'il est vrai qu'à l'origine, la recherche en traduction automatique visait l conception d'un système pouvai manipuler des textes de tous genres dans toutes sortes de langues, les objec tifs actuels sont nettement plu modestes : la traduction, même impa faite, doit simplement permettre d comprendre les différentes notion abordées dans un texte. Les fourni seurs de service eux-mêmes sont pru dents : bien qu'un certain optimism soit de mise dans la manière de préser ter les services offerts, il n'en reste pa moins que la plupart prennent la pr caution de prévenir l'utilisateur d l'imperfection du résultat obtenu. P exemple, E-lingo indique que ses tr ductions suffisent pour transmetti l'idée générale d'un message (*gist* Elle recommande à ses clients d'utilis des phrases courtes grammaticalemer correctes, et d'éviter le « jargon » et langage familier. Pour les traductio qui se doivent d'être « proches de perfection », elle les incite à faire app à des traducteurs humains.

Selon les cas, il est possible c soumettre des textes collés dans u fenêtre, des URL (pour traduire u page web) ou des courriers électr

niques. Aucune forme d'intervention n'est autorisée pendant le processus de traduction. En outre, l'utilisateur ne peut enrichir le logiciel avant de soumettre sa traduction (certaines formes d'enrichissements, notamment l'ajout

« Le site Foreignword sélectionne les systèmes de traduction pertinents en fonction de la paire de langues choisie, s'y connecte et affiche la traduction dans un cadre. »

de mots inconnus, sont permises dans les logiciels de TA que l'on installe sur micro-ordinateur). La plupart des systèmes testés retournent la traduction d'un texte de 100 à 200 mots en quelques secondes, sauf InterTran™, parfois très lent et pouvant prendre plusieurs minutes avant de retourner une réponse. Systran, sur Altavista Babelfish, annonce qu'il accepte des textes de 800 mots (copié-collé) et des pages web de 5 K, tandis que E-lingo n'a pas traduit plus de 350 mots environ d'un texte soumis comportant 700 mots. Reverso Voila, FreeTranslation, InterTran™ et GPL Translation acceptent des textes d'environ 500 mots et les retournent avec des délais variant de quelques secondes à une minute, tandis qu'ils refusent des textes supérieurs à 1 000 mots (affichage d'un message d'erreur ou informant que la capacité du système est dépassée). La capacité de ces systèmes est donc fortement limitée par le nombre de mots des textes soumis autant dire qu'on ne pourra pas envisager de recourir à ces services pour traduire une thèse...

Afin de donner un aperçu de l'utilité de ces logiciels pour accéder au contenu

de textes médicaux rédigés en langue étrangère, nous leur avons soumis des textes et nous en avons évalué la traduction. Étant donné que la plupart des articles sont écrits en anglais, nous nous sommes bornées à tester la paire anglais-français. Toutefois, selon les particularités lexicales, syntaxiques et grammaticales de chaque langue, il se peut que certains systèmes soient plus performants pour certaines paires de langue que pour d'autres.

« Les fournisseurs de service eux-mêmes sont prudents : la plupart prennent la précaution de prévenir l'utilisateur de l'imperfection du résultat obtenu. »

Trois échantillons de texte ont été soumis : (A) un texte relativement général contenant des expressions idiomatiques de la vie courante ; (B) un résumé d'article médical contenant toutefois relativement peu de termes spécialisés ; (C) un extrait du *Merck Manual* contenant de nombreux termes cliniques.

Quelques problèmes courants observés dans les textes soumis

• Le dictionnaire du logiciel ne renferme pas tous les mots contenus dans le texte : ce problème est crucial lorsqu'on soumet un texte spécialisé comme un texte médical. Ainsi, *pathogenicity, autoclaving* sont reproduits tels quels dans la traduction produite par Reverso. *Spongiform, fatigability, dyslexia* sont laissés tels quels dans la traduction produite par Systran. InterTran™ place les mots qu'il ne connaît pas entre crochets : *[Salmonella] [enteritidis].*

• Le logiciel connaît le mot, mais pas le sens qu'il a en médecine. Par exemple, *disturbance* est rendu par perturbation par Systran, *spongiform* est rendu par en *forme d'éponge* par Reverso.

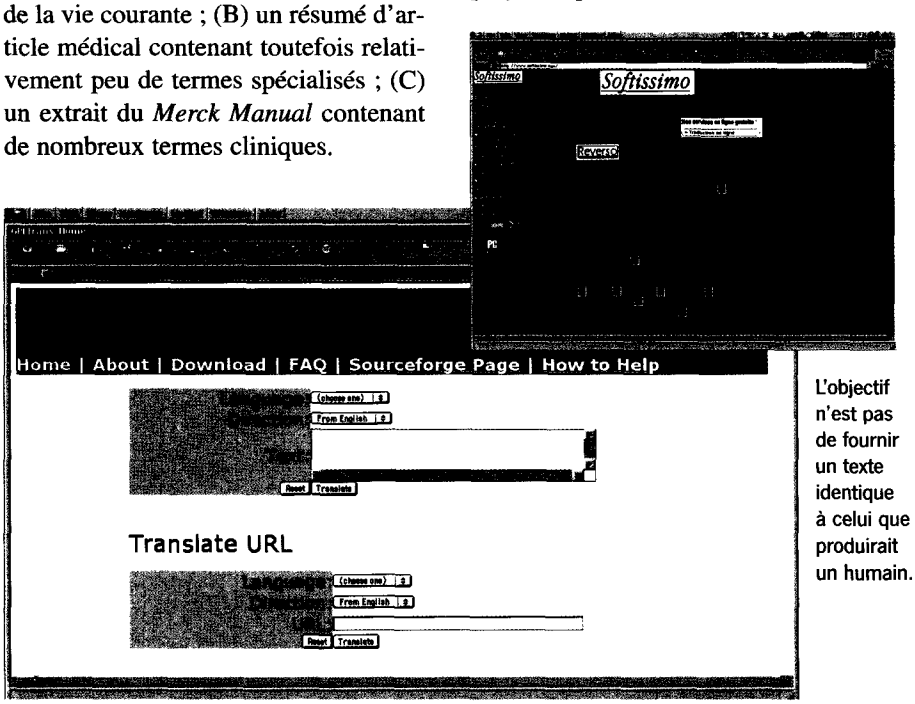
• Un mot a plusieurs sens (mot polysémique) et le logiciel ne choisit pas le bon : *In these cases* est rendu pas *Dans ces caisses* par Systran. À noter que Reverso propose, entre parenthèses, une seconde possibilité pour les mots polysémiques : *In these cases* => *Dans ces cas (affaires).* InterTran™ perm également à l'utilisateur de visualis d'autres possibilités d'équivalents : p exemple, *food* est rendu par *pâtu* mais l'utilisateur peut accéder à *alime* et *nourriture.*

La polysémie (multiplicité des sens pc un même mot) constitue un type d'écu qui introduit une ambiguïté dans le te à traduire. C'est un problème courant traduction, que les traducteurs humai résolvent avec plus de facilité que l traducteurs machines, car pour ces d niers, la prise en compte du sens et contexte constitue un défi de taille.

• Le logiciel ne connaît pas toujours d expressions figées contenant plusie mots : Ainsi, *sunny-side-up or over-ea eggs* est rendu par le *côté ensoleillé haut* ou *des oeufs sur-faciles* par Reve et par *ensoleillé-côté-vers le haut* ou *d dessus-facile œuf* par Systran.

• Un mot peut appartenir à plusie catégories grammaticales (verbe nom, par exemple) et le logiciel choi la mauvaise. Par exemple, dans *rec media reports, reports* est traduit p *enregistrer* par Systran alors qu'il aur dû choisir *rapports* dans ce contexte. même, *must bare safe handling instr tions* : *safe* est interprété comme un n par Reverso et traduit par *coffre-fort.* Là aussi, la similitude morphologiq entre un adjectif et un verbe, ou en un adjectif et un nom, crée une am guïté que le traducteur humain réso plus facilement qu'une machine.

• Les prépositions et les locutic prépositives sont mal traduites : c'est une difficulté fort courante, étant dor la multiplicité des possibilités ex tantes. En effet, l'usage d'une prépo tion est généralement contraint par autre mot, un verbe ou un nom. De pl elles font souvent partie d'expressic idiomatiques qui ne sont pas trad sibles mot à mot. L'équivalence en



L'objectif n'est pas de fournir un texte identique à celui que produirait un humain.

prépositions appartenant à des langues différentes n'est pas stricte et les systèmes ne sont pas toujours aptes à décider de la bonne solution. Par exemple, InterTran™ traduit systématiquement la préposition *by* par *près de*, et non par *par* (l'indication de moyen étant pourtant le sens le plus courant de *by*).

• Enfin, en raison des différences importantes existant entre les langues, les logiciels éprouvent généralement des difficultés à produire des phrases correctes dans la langue cible. Remarquons qu'InterTran™ ne fait aucune forme d'analyse de la phrase et ne fait que substituer les mots du texte d'origine par des mots de la langue cible après avoir consulté un dictionnaire :
Restaurants and other retail establishments should continue to follow their local and state health regulations covering food service establishments, many of which have adopted the FDA Food Code.
Restaurant et autre vente au détail établissements devez reprendre à suivre à eux local et état santé réglementaire enveloppe pâture service établissements, un grand nombre de quoi prendre adoptée les FDA Pâture Code.

Le texte traduit est-il compréhensible?

Le résultat global est que, en raison des différences de syntaxe importantes existant entre les langues et de l'absence de correspondance rigoureuse du sens des mots, les logiciels éprouvent des difficultés à produire une traduction correcte. Il est possible de répertorier un certain nombre d'erreurs en traduction, mais celles qui aboutissent à un sens erroné sont les plus graves, qu'elles soient dues au mot choisi (erreurs terminologiques) ou à une structure de phrase déficiente (erreurs grammaticales et syntaxiques).
Les différents logiciels testés retournent, pour un même texte en langue source, des traductions sensiblement différentes, en raison, notamment, des stratégies variables de traitement du mot et de la phrase. Il n'a pas été possible, bien entendu, de porter un jugement direct sur les stratégies employées (connues uniquement des concepteurs). Seul le résultat est accessible. Pour tenter d'objectiver les différences entre les différents logiciels testés, nous avons, sur le texte (B), attribué des points aux erreurs commises par les logiciels, en pondéra davantage les erreurs induisant u ambiguïté ou une erreur au niveau d sens (5 à 7) que les erreurs de syntax et de grammaire (3 à 4) ou de style (sans conséquences majeures sur le se de la phrase. Certains systèmes peuve être écartés d'emblée. Ainsi, GPLTran encore en développement, est très pe efficace (il indique lui-même, selon l textes, que 16 à 17 % environ du tex soumis a été traduit) et InterTran™, q n'effectue pas d'analyse syntaxique, n permet pas d'obtenir de texte compr hensible. FreeTranslation retourne, l aussi, des textes difficiles à interprét Les deux systèmes qui semblent l plus performants pour les échantillo soumis sont Systran et Reverso.
Systran retourne une traduction pl qu'acceptable. Reverso propose d'autr solutions de traduction pour certai termes entre parenthèses, mais il n pas su interpréter correctement le déb de la phrase. IBM Alphaworks comm une faute d'accord et ne traite p correctement le possessif. On note que bien que produisant une traductio acceptable pour la phrase prise e exemple, T-mail obtient un score relat vement bas pour l'ensemble du tex

Termes cliniques du texte (C) en contexte	Traduction française correcte*	REVERSO	SYSTRAN	IBM Alphaworks	e-lingo
Commonly evidenced by self-neglect.	Se traduit habituellement par une négligence corporelle.	Généralement faite la preuve par la négligence de soi.	Généralement démontrée par individu-négligent.	Manifesté communément par le désintérêt de moi.	Manifesté par moi communément - négligence
Some patients complain of easy fatigability, somnolence, insomnia, or other sleep disorders.	Certains patients se plaignent d'une fatigabilité anormale, de somnolence, d'insomnie, ou d'autres troubles du sommeil.	Quelques patients se plaignent de fatigability facile, la somnolence, l'insomnie, ou d'autres troubles du sommeil.	Quelques patients se plaignent du fatigability, du somnolence, de l'insomnie, ou d'autre facile des désordres de sommeil.	Quelques patients se plaignent de le fatigability facile, la somnolence, l'insomnie ou les autres désordres de sommeil.	Quelques malades se plaignen de fatigability facile, somnolence, insomnie, ou aut désordres du sommeil.
Other abnormalities of higher cortical function (eg, aphasia, apraxia, dyslexia, dysgraphia, agnosia, left-right disorientation, unilateral neglect) may occur.	D'autres anomalies des fonctions corticales supérieures, (p. ex. aphasie, apraxie, dyslexie, dysgraphie, agnosie, troubles de la latéralité et désintérêt à l'égard d'un hémicorps) peuvent survenir.	D'autres anomalies d'une fonction corticale plus élevée (par exemple, aphasie, apraxie, dyslexia, dysgraphia, agnosie, désorientation de gauche à droite, négligence unilatérale) peuvent se produire.	Les autres anomalies de la fonction corticale plus haute (par ex., le disorientation, le désintérêt unilatéral l'aphasie, l'apraxia, la dyslexie, dysgraphia, agnosia, la gauche droit) peut se trouver.	Autres caractères anormaux de plus haute fonction corticale (eg, aphasie, apraxia, dyslexie, dysgraphia, agnosia, gauche - bon dépaysement, négligence unilatérale) peut se produire.	Autres caractères anormaux de plus haute fonction corticale (eg, aphasie, apraxia, dyslexie dysgraphia, agnosia, gauche - bon dépaysement, négligenc unilatérale) peut se produire.

* Solutions de traduction proposées dans la version française, Manuel Merck, 2e édition.

soumis. E-lingo et les autres logiciels sont moins performants. Enfin, on soulignera la difficulté posée par la traduction du mot *polysémique responder*, qui est rendu, selon le cas, par *répondants, défendeurs, personnes interrogées, défenseurs*. Il faut souligner toutefois que, selon les textes, la comparaison entre logiciels peut se révéler difficile et que chaque logiciel peut se montrer plus performant pour résoudre certaines difficultés particulières. Ainsi, pour d'autres portions de texte, Systran a été le moins performant, Reverso et e-lingo étant à peu près à égalité.

Un paramètre essentiel : le vocabulaire spécialisé

Un paramètre important de la traduction en langue de spécialité est lié à la richesse des dictionnaires qui alimentent le logiciel.
La gratuité a ses revers : en effet, contrairement à un système dont on a fait l'acquisition, il est impossible d'améliorer les systèmes proposés sur Internet. À l'aide du texte (C), tiré du *Merck Manual*, qui comporte de nombreux termes cliniques, nous avons mis à l'épreuve Systran, Reverso, IBM Alphaworks et e-lingo.

Il est facile de constater que les termes et expressions médicales sont rendues de manière très inégales et approximatives. Différents facteurs rendent la traduction délicate : (1) les dictionnaires alimentant les systèmes ne sont pas assez riches en terminologie médicale : les logiciels tendent alors à laisser le mot tel quel dans la traduction (*fatigability ; dysgraphia ; etc.*) ; (2) comme on l'a déjà fait remarquer, certains mots acquièrent, dans un contexte médical, un sens particulier qu'ils n'ont pas en langue générale (*easy* dans *easy fatigability*) ; (3) il n'y a pas de correspondance mot-à-mot entre le français et l'anglais (*troubles de la latéralité*, pour *left-right disorientation* ; *désintérêt à l'égard d'un hémicorps*, pour *unilateral neglect*). Là encore, bien que très imparfaits, Reverso et Systran donnent des résultats supérieurs à ceux de leurs concurrents.

Conclusion

De nombreux facteurs interviennent dans la lisibilité du texte produit en langue cible. D'une part, celle-ci résulte de la performance intrinsèque du logiciel à traiter les ambiguïtés lexicales et syntaxiques de la langue source. Retenons que les systèmes de traduction automatique offerts sur l'Internet ne sont pas de qualité égale : Reverso et Systran paraissent les plus performants. D'autre part, la lisibilité du texte traduit dépend de la nature du texte à traduire. Plus celui-ci contient de structures ambiguës, de terme et complexes, moins le texte sera lisible. L'obstacle majeur en cine est donc lié à la nature spéc et complexe de la terminologie. en fin de compte, la connaissa domaine par le lecteur qui pourr un rôle déterminant : celui-ci se fois capable de saisir à demi- contenu, bien que, parfois le r reste totalement incompréhensib Dans tous les cas, on ne saurait s ses systèmes les yeux ferm risques d'erreurs de sens, qui imprévisibles, ne pouvant jama écartés. En fin de compte, s'il e sible de savoir, très grossièrem quoi parle le texte, il ne fait doute que son traitement appro ne permet pas d'être certain de prétation qui doit lui être attribu

Les auteurs

Sylvie Vandaele enseigne la traduction médicale dans le Département de linguistique et de traduction de l'Université de Montréal à titre de professeure adjointe. Faisant appel, dans ses cours, à la mise en ligne de matériel pédagogique pour les étudiants, elle s'intéresse tout particulièrement aux ressources que recèle l'Internet en matière de médecine et de traduction.

Marie-Claude L'Homme, professeure agrégée dans le Département de linguistique et de traduction de l'Université de Montréal, est chargée de la formation aux outils informatiques destinés aux traducteurs. Elle est spécialisée dans le domaine de la terminologie et s'intéresse plus particulièrement à la conception de dictionnaires en format électronique.

Références des textes

• Texte général
FDA corrects erroneous media reports safety guidelines. Press Office, Food and Drug Administration, 15 juin www.cfsan.fda.gov/~lrd/fpeggs2.html

• Résumé article médical
Vally H, de Klerk N, Thompson PJ. Asthma induced by alcoholic drinks : a new food allergy questionnaire. *Aust N Z J Public Health* 1999 ; 23 : 5

• Extrait du Merck Manual
Prion disease. The Merck Manual on CD-ROM, (17th edition), 1999, Merck & Co. Inc., Chap 162 p 1300 et

• Version française
Maladie de Creutzfeldt-Jakob. Manuel Merck de diagnostic et thérape (2e édition, traduction de la 16e éditio américaine) 1992, Merck Research Laboratories, Chap. 1, p. 197 et suiv.

Des pages essentielles

- Pour en savoir plus sur la traduction automatique et les outils d'aide à la traduction (outils, systèmes, articles) : www.foreignword.com/fr/Technology/technology.htm
- Melby A. Why can't a computer translate more like a person ? 1995 Barker Lecture, www.tlt.org/theory/ barker.html
- Pour accéder à différents systèmes de traduction simultanément : www.foreignword.com/Tools/transnow.htm