Chapitre 3.6

Compréhension et multilinguisme

« To translate is to re-express in a second language what has been understood by reading a text. Any purported solution to the problem that does not involve understanding in some sense is, at best, ad hoc… »
Martin Kay, The Proper Place of Men and Machines in Language Translation

3.6.1. Introduction

Comme l'ont montré les chapitres précédents, la compréhension d'énoncés dans une langue est une tâche délicate qui ne fait que se compliquer lorsque plusieurs langues sont en jeu. Certains humains arrivent à comprendre et à s'exprimer en plusieurs langues, quoiqu'il soit très difficile de le faire aussi facilement dans une langue différente de sa langue maternelle. Cette dernière forge les processus de pensée et donne les premiers référents dont il est si difficile de s'abstraire par la suite. Ces différences n'arrivent pas seulement lors de changement de langue ; on n'a qu'à penser aux différences entre les français (de France, du Québec, de Belgique, de Suisse, etc.). On pourrait imaginer que la compréhension est un processus qui se passe à un niveau indépendant de la langue mais cette question des subtiles interactions entre le langage et la compréhension est loin d'être réglée.

Le traitement automatique de la langue a d'ailleurs débuté dans les années 50 avec le traitement multilingue pour des besoins de traduction automatique entre le russe et l'anglais. On pensait procéder de façon analogue au décodage de messages en cryptographie. Cette problématique, poussée par des impératifs politiques et

---

Chapitre rédigé par Guy LAPALME et Elliott MACKLOVITCH.

économiques, a suscité d'importants travaux et espoirs (partiellement déçus) dans le domaine de la traduction automatique ; on pourra consulter le chapitre 7 de [FUC 93] pour un rappel historique en français des travaux précurseurs dans le domaine. Dans ce cas précis, le but n'était pas de produire une compréhension profonde des textes mais plutôt d'engendrer un transcodage dans une autre langue de façon à ce qu'il devienne compréhensible par quelqu'un qui ne connaît pas la langue du message original. Cela pose avec une certaine acuité un problème fondamental et qu'on peut synthétiser de façon polémique avec la question suivante :

Est-il vraiment nécessaire de comprendre pour traduire ?

Les traducteurs humains répondraient unanimement et sans hésitation par l'affirmative. Mais quel est le sens de cette question pour une machine ? Peut-on distinguer différents niveaux de compréhension dont certains seraient intelligibles par un ordinateur ? Et ces niveaux sont-ils suffisants pour permettre à une machine de bien traduire ? Dans cet article, nous allons tenter de donner différentes réponses à cette question, réponses qui nous conduiront à explorer certaines applications du multilinguisme en traduction, en extraction ou en recherche d'information. Nous les illustrerons avec des travaux effectués dans notre laboratoire, le RALI à l'Université de Montréal[1].

Mais d'abord, il convient de noter que le traitement multilingue multiplie les besoins en ressources linguistiques car il faut disposer de lexiques, de bases de données terminologiques, de grammaires et d'ontologies pour chacune des langues à traiter. Non seulement faut-il obtenir ces ressources, il faut en plus les mettre en correspondance sous la forme de lexiques bilingues, de formules de correspondances, de règles de traduction ou de termes d'ontologies partagées.

3.6.2. Traduction

Lorsqu'il est question de multilinguisme, la traduction nous vient immédiatement en tête, car, par sa définition, elle met en relation des énoncés dans deux ou plusieurs langues. D'ailleurs, voici la définition du verbe *traduire* fourni par le Petit Robert : « faire que ce qui était énoncé dans une langue le soit dans une autre, en tendant à l'équivalence sémantique et expressive des deux énoncés. » La traduction, autrement dit, est **une relation d'équivalence sémantique** entre deux énoncés en langues différentes. D'ailleurs, une excellente façon de démontrer notre compréhension d'un concept ou d'un texte est de le traduire en mots dans une autre langue. C'est ce qui fait spontanément tout traducteur lorsqu'on lui demande de nous expli-

[1] Recherche Appliquée en Linguistique Informatique, `http://rali.iro.umontreal.ca`

quer le sens d'une expression en langue étrangère : il nous la traduit dans notre langue maternelle.

3.6.2.1. *Les générations de la traduction automatique*

Comment capturer cette équivalence sémantique dans un système informatisé ? Le schéma traditionnel de la traduction automatique (Figure 3.6.1) – proposé à l'origine par Bernard Vauquois – comprend un niveau d'abstraction commun (appelé interlingua ou langue pivot) aux processus d'analyse de la langue source et de génération dans la langue cible. En principe, cette représentation abstraite est indépendante de toute langue et elle est censée formaliser tout le sens ou le contenu d'un message linguistique. Ainsi, elle est l'aboutissement de la phase d'analyse dans un système de traduction automatique (TA), c.-à-d. l'expression formelle de la compréhension du texte d'entrée, et aussi le point de départ de la phase de génération qui verra à formuler ce contenu dans une autre langue. En pratique, cependant, l'élaboration d'une telle interlingua qui saurait exprimer de façon neutre tous les sens possibles dans toutes les langues humaines représente un défi énorme. On peut même se demander si une telle chose est possible même en nous limitant à un hypothétique sens littéral. Comme le souligne Anne Reboul au chapitre 2.3, les aspects pragmatiques sont fondamentaux dans le traitement de la langue et d'autant plus en traduction. Mais alors comment intégrer la pragmatique dans un interlingua.Afin de contourner les problèmes occasionnés par la définition d'une représentation abstraite, on a souvent recours à un processus de transfert qui s'opère à un niveau inférieur dans le processus d'analyse et qui met en relation des expressions (des mots ou des structures) des langues sources et cibles.

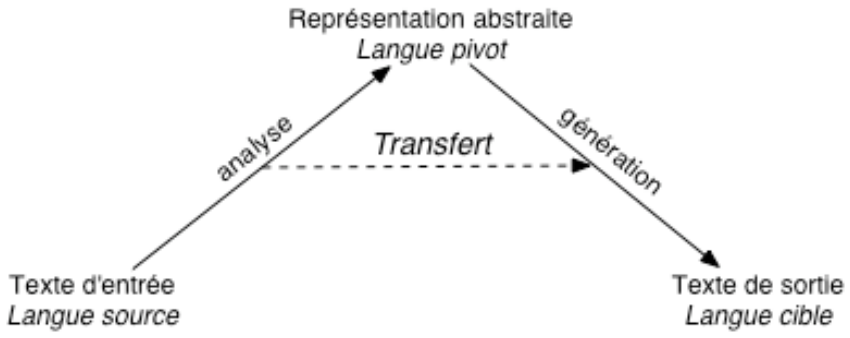

Figure 3.6.1. *Schéma de traduction avec interlingua*

Mais si l'on arrive ainsi à faire le pont entre deux langues naturelles, pourquoi s'acharner à concevoir une interlingua ? Le projet d'un *lingua universalis* remonte

au moins à Leibniz, qui voulait de cette façon minimiser les désaccords entre les hommes, mais ici nous lui donnerons une justification plus pragmatique. Supposons que nous voulions développer un système de TA multilingue. De prime abord, l'approche par interlingua semblerait nous simplifier la tâche, car elle permet de se passer de multiples composantes de transfert. Pour chaque nouvelle langue que nous voudrons ajouter au système, il suffirait d'élaborer les règles d'analyse qui effectueront la correspondance entre les textes de surface et leur représentation abstraite ; et la même chose pour la composante de génération. Dans l'approche de transfert, en revanche, de la même façon, nous aurons besoin de composantes d'analyse et de génération (même si celles-ci n'auraient pas à traiter des structures aussi abstraites) ; mais de plus il faudra définir une composante bilingue de transfert qui est propre à chaque paire de langues – et plus il y aura de langues dans ce système, plus cette tâche de rédiger les modules bilingues sera onéreuse.[2] Cependant, cet avantage théorique dont jouit l'approche par interlingua n'a pas aidé à surmonter les difficultés pratiques qui ont toujours fini par miner les efforts d'élaborer un *lingua universalis* explicite et complet.

Ce qui nous amène à reformuler la question que nous avons posée ci-dessus :
La compréhension profonde est-elle toujours nécessaire à une traduction ?

Autrement dit, ne pourrait-on pas se limiter à une transposition des mots d'une langue à l'autre et reporter la compréhension sur le lecteur de la langue cible ? C'était essentiellement l'approche des systèmes de TA dits « de première génération », c'est-à-dire ceux qui ont été développés pendant les années 50 et le début des années 60. Et, il faut bien l'admettre, cela fonctionne parfois, surtout lorsqu'il s'agit de deux langues proches qui appartiennent à la même famille. Pour transposer les mots de L1 à L2, ces systèmes de première génération comportaient un important dictionnaire bilingue – c'était d'ailleurs là leur principale composante – dont le contenu était appliqué directement à la chaîne de mots en langue source ; venait ensuite une petite composante de règles ad hoc qui s'appliquaient à la chaîne de mots cibles et qui servaient à les réordonner et à les accorder pour qu'ils respectent la grammaire de cette langue. Certes, ces systèmes renfermaient une certaine compréhension des textes sources, mais celle-ci était assez rudimentaire et essentiellement de nature lexicale. Si tout le monde reconnaît que la compréhension lexicale est nécessaire pour traduire, elle est hélas loin d'être suffisante.

> « les dictionnaires… partent d'une abstraction fondamentale : en proposant des équivalents aux mots, ils imposent une torsion parfaitement artificielle à la langue, car ils négligent ce faisant le fait que les mots n'ont de sens que dans un contexte, qu'avec la présence d'autres mots. » [DES 04]

[2] Pour un système qui traite *n* langues, la formule pour calculer le nombre transferts nécessaires est *n x (n-1)*.

Ainsi, les limites de l'approche mot à mot incarnée dans les systèmes de première génération sont vite devenues apparentes. Déjà en 1957, Victor Yngve, de l'université MIT, proposait une approche *indirecte* à la traduction automatique dans laquelle les règles de transfert s'appliqueraient, non pas à une chaîne de mots, mais plutôt à une représentation plus abstraite, résultat d'une analyse syntaxique complète de la phrase. D'après [YNG 57], cette représentation permettrait d'identifier la fonction syntaxique de chaque mot et de chaque syntagme de la phrase et offrirait une base plus fiable pour la désambiguïsation, car justement les mots seraient analysés dans leur contexte phrastique. Voilà un des fondements des systèmes dits de deuxième génération, l'autre étant la modularité, c.-à-d. une séparation nette des connaissances linguistiques et les algorithmes procéduraux qui les appliquent.

Les systèmes de deuxième génération ont dominé les travaux en traduction automatique pendant les années 70 et 80 et, par rapport à leurs prédécesseurs de première génération, ils représentaient indubitablement un progrès, car ils visaient une compréhension plus ambitieuse du texte source. Mais la force de ces systèmes était en même temps une de leurs faiblesses, dans la mesure où ils exigeaient des grammaires d'analyse de large couverture. À l'époque, il n'y avait d'autre choix que de rédiger ces grammaires à la main, et très souvent ils renfermaient de vastes ensembles de règles fort complexes, ce qui ne les rendait pas plus robustes pour autant. Face à une phrase de type non prévue ou contenant un mot inconnu, le système ne produisait souvent aucune analyse, et, par conséquent, aucune traduction. Même dans le meilleur des cas, c'est-à-dire. là où l'analyse réussissait, la compréhension syntaxique sur laquelle ces systèmes reposaient pouvait se révéler insuffisante, car les connaissances nécessaires pour traduire dépassent souvent le niveau purement grammatical. Les exemples sont tellement nombreux qu'il n'est pas nécessaire d'en citer ici. En fait, aux dires de Martin Kay, un des pionniers de la TA,

> « There is nothing that a person could know, or feel, or dream, that could not be crucial for getting a good translation of some text or other. » [Kay 92]

Les systèmes de TA de troisième génération devaient en principe pallier cette insuffisance en intégrant dans leurs représentations de l'information sémantique sur la nature des actants et des actions, mais également de l'information non explicitée dans le texte d'entrée qui provient de connaissances du monde et des fonctions pragmatiques de l'énoncé. Contrairement aux systèmes de deuxième génération, qui utilisent le vocabulaire de la langue cible pour distinguer le sens des expressions de la langue source, ces systèmes de TA « à base de connaissances » visent une représentation formelle et plus abstraite du contenu du l'énoncé, qui saurait en plus supporter des inférences ; voir à ce sujet [NIR 90]. En cela, ces systèmes de troisième génération dépassent même les visées des premières propositions de l'approche par interlingua. Ainsi, il n'est pas surprenant de constater qu'il existe aujourd'hui peu de systèmes de TA opérationnels qui se revendiquent de cette génération ; et les quel-

ques exceptions ont dû être élaborées pendant de nombreuses années et pour des domaines sémantiques bien cernés, p. ex. le système KANT conçu pour la société Caterpillar.

Les efforts humains et les coûts nécessaires au développement de systèmes de TA ont fini par avoir raison de l'approche rationaliste dans laquelle les connaissances (linguistiques ou autres) sont codées dans des règles rédigées à la main par de nombreux spécialistes hautement qualifiés. Les années 90 ont vu l'émergence d'une nouvelle tendance empirique en traitement automatique des langues (TAL), dont les premiers champions en traduction automatique était l'équipe dirigée par Peter Brown et Fred Jelinek au centre de recherche d'IBM aux Etats-Unis. S'inspirant du succès de l'approche empirique en reconnaissance de la parole, ce groupe rejetait la codification des connaissances linguistiques au moyen de règles déclaratives et prônait à la place une approche probabiliste à la traduction, les paramètres de leur système étant inférés d'énormes corpus parallèles par des méthodes d'apprentissage automatique ; voir [BRO 90]. La performance étonnante de leur système dans les compétitions organisées par DARPA au milieu des années 90 a littéralement révolutionné tout le domaine, au point où l'approche statistique domine complètement la recherche en TA de nos jours ; pour s'en convaincre, il suffit de consulter les actes de n'importe quelle grande conférence en TAL. Nous nous pencherons sur la TA statistique de façon plus détaillée dans la prochaine section, mais revenons pour le moment à la question que nous avons posée au tout début de ce chapitre : *Est-il vraiment nécessaire de comprendre pour traduire ?* Quel genre de compréhension renferme ces systèmes de TA statistique ? La réponse n'est pas du tout évidente, surtout quand on se rappelle que pour Brown et al. toute phrase dans une langue est une traduction possible de toute phrase dans une autre, la seule chose servant à distinguer une traduction complètement farfelue d'une traduction « correcte » étant la probabilité plus élevée que le système assignerait à cette dernière. La compréhension dont fait preuve ces systèmes n'est peut-être pas profonde, dans le sens que leurs paramètres mathématiques ne prétendent rien révéler sur la structure sous-jacente des langues en cause ; et ils sont loin d'être limpides à l'observateur humain. Mais compréhension, il doit y en avoir ; autrement comment expliquer le fait que la performance de ces systèmes tend nettement à s'améliorer lorsqu'on leur alimente de nouvelles données parallèles. Et pour l'heure, cette compréhension semble la être seule qui est en mesure de cerner de façon robuste des objets aussi compliqués et remplis d'exceptions que deux langues humaines dans une relation de traduction.

3.6.2.2. *Modèles probabilistes de traduction*

La disponibilité grandissante de textes sous format électronique, évoquée par Benoît Habert au chapitre 3.2, ainsi que des moyens de stockage et de traitement de plus en plus performants ont permis un développement accéléré de l'approche empi-

rique au cours des dernières années. Comme nous l'avons vu, les modèles statistiques de langue ont l'avantage d'être robustes, c.-à-d. de pouvoir traiter de *vrais* textes et non pas seulement des idéalisations théoriques. Ils peuvent de plus fournir toujours une réponse (parfois même plusieurs) et ils sont en général plus rapides à appliquer que les méthodes symboliques. Il faut toutefois convenir de quelques inconvénients : ces systèmes sont complexes et difficiles à mettre au point par quelqu'un qui n'a pas d'expertise en calcul statistique ; et leurs sorties sont souvent opaques, c.-à-d. difficiles à comprendre en termes de principes intelligibles par un linguiste non statisticien.

Un modèle de langue probabiliste peut être présenté comme une fonction qui donne la probabilité de rencontrer un mot en fonction des *n* mots précédents. Par souci de simplification, on ne considère que le contexte immédiat comme pouvant influencer le prochain mot, car des raisons pratiques imposent des restrictions sur la longueur du texte tenu en ligne de compte. En principe, on aimerait que *n* soit grand mais on aurait alors trop de paramètres à estimer, comme le montre le tableau 3.6.1 pour un vocabulaire de 20 000 mots.

	n	**# de paramètres**	
bigramme	2	$20\,000^2$	400 millions
trigramme	3	$20\,000^3$	8 trillions
4-gramme	4	$20\,000^4$	$1.6 \bullet 10^{17}$

Tableau 3.6.1. *Nombre de paramètres à estimer pour des n-grammes pour un vocabulaire de 20 000 mots.*

Différentes méthodes sont appliquées pour restreindre l'espace des probabilités, chercher des estimateurs à vraisemblance maximale et surtout traiter l'éparpillement des données. Cette approche a déjà montré son utilité dans plusieurs applications dont la reconnaissance de la parole et de caractères, la correction de fautes d'orthographe et les systèmes de traduction automatique. Manning et Schütze [MAN 99] présentent une excellente introduction au domaine. Debili et Souissi [DEB05] étudient le problème d'un autre angle, celui de la taille des règles.

L'approche statistique aux modèles de traduction estime des probabilités de rencontrer des mots de la langue cible étant donné l'historique du texte cible et du texte source. On utilise souvent l'analogie du canal bruité dans lequel entre une phrase cible (*target*) **t** d'une langue *T* ayant T comme vocabulaire. Le canal est si bruité que la phrase source **s** qui en sort est la traduction de **t** dans une autre langue *S* ayant S comme vocabulaire. Le processus de traduction consiste donc, en présence d'une phrase de sortie **s** de la langue S, à trouver **t** dans la langue T qui l'a produite.

La modélisation statistique de la traduction est définie avec l'équation suivante simplifiée à l'aide la loi de Bayes pour séparer les processus de traduction et de génération de la phrase cible.

$$\hat{\mathbf{t}} = \arg\max_{t \in T} p(\mathbf{t} \mid \mathbf{s})$$

$$\hat{\mathbf{t}} = \arg\max_{t \in T} \frac{p(\mathbf{s} \mid \mathbf{t}) \times p(\mathbf{t})}{p(\mathbf{s})}$$

$$\hat{\mathbf{t}} = \arg\max_{t \in T} p(\mathbf{s} \mid \mathbf{t})_{traduction} \times p(\mathbf{t})_{langue}$$

Pour ce processus, il faut d'abord estimer p (**t**) la distribution de la langue cible et p (**s** | **t**) la distribution de la traduction pour ensuite rechercher la phrase $\hat{\mathbf{t}}$ qui maximise le produit de ces probabilités. Les distributions des modèles de langue et de traduction sont estimées à partir d'observations faites sur de grands corpus de traductions alignées au niveau de la phrase. En plus des défis posés par la taille de ces modèles comme nous l'avons signalé précédemment, il faut mettre au point des moyens de les optimiser rapidement en y intégrant des informations fondées sur les alignements de mots entre les phrases. Brown et al [BRO 93] ont présenté une suite de cinq modèles de complexité croissante et qui ont ensuite été améliorés soit par de nouvelles méthodes d'optimisation [BER 96], soit en intégrant des unités plus grandes que les mots individuels [OCH 04].

Cette approche probabiliste est maintenant intégrée dans plusieurs systèmes de traduction automatique, leur principal avantage étant qu'ils peuvent être mis au point beaucoup plus rapidement que les systèmes classiques de TA qui requièrent le développement manuel de règles et de dictionnaires bilingues. En revanche, une fois défini le modèle général d'apprentissage statistique, l'approche probabiliste nous permet de rapidement mettre au point un système de TA pour n'importe quelle paire de langues – même des langues que nous ne comprenons pas – pour autant qu'on ait accès à des corpus parallèles de taille suffisante. La connaissance des langues à traduire est quand même très utile pour l'ajustement de pondérations entre les différents paramètres déduits par les composantes d'un système de traduction automatique. En pratique, les résultats sont souvent grandement améliorés par des pré- et des post-traitements astucieux comme la séparation en unités lexicales appropriées qui sont fortement dépendantes des langues. Comme le montrent les dernières évaluations NIST, les systèmes probabilistes ont réussi après une dizaine d'années à devenir compétitifs avec les meilleurs systèmes de traduction symboliques développés depuis les années 70. Dans notre laboratoire, nous avons effectué le même genre d'exercice en développant en quelques mois une version statistique du système TAUM-Météo [LEP 04]. Le fait qu'on puisse ainsi *apprendre à traduire* à partir de statistiques sur le corpus est sûrement un élément intéressant sur le niveau de compréhension nécessaire pour certaines tâches de traduction. Devrions-nous en déduire

qu'il est possible de traduire *par cœur* ? Tout novice en la matière ou apprenant d'une nouvelle langue effectue souvent ce genre de traduction *machinale* . Il n'est donc pas surprenant qu'un système automatique n'arrive pour l'instant qu'à ce niveau de traduction au premier degré.

3.6.2.3. *Traduction assistée par ordinateur*

Nous avons affirmé ci-dessus que certains systèmes de TA probabilistes ont récemment réussi à dépasser les systèmes de TA classique dans les compétitions organisées par DARPA et NIST. Cependant, il ne faut pas en conclure que le problème de la traduction automatique soit résolu pour autant ; loin de là. Les progrès permis par les nouvelles techniques empiriques se situent surtout au niveau de la mise au point des systèmes. Pour ce qui est de la qualité des sorties, en revanche, les progrès ont été beaucoup moins marqués. En fait, dans la plupart de ses applications de nos jours, la TA est utilisée à des fins d'assimilation d'information, p. ex. par des agences de renseignements militaires, dans un contexte où les exigences de qualité sont souvent moins importantes que la vitesse ou la capacité du traitement. Lorsqu'il s'agit de traduction à des fins de diffusion ou de publication, les exigences de qualité sont forcément plus rigoureuses et les sorties engendrées par les systèmes de TA actuels, quelle que soit leur orientation théorique, sont rarement à la hauteur. Pendant de nombreuses années, on a tenté de convaincre les traducteurs que ces sorties « brutes » pourraient servir de première ébauche que ces derniers n'auraient qu'à corriger et à polir. Mais il faut se rendre à l'évidence : les applications rentables de ce modèle d'interaction sont plutôt l'exception qui confirme la règle, et pour cause. Ce genre de « post-édition » n'a rien à voir avec la révision des traductions humaines, puisque la machine fait des erreurs que ne commettrait aucun traducteur, même débutant. La plupart du temps, les traducteurs trouvent qu'ils peuvent produire un texte de meilleure qualité en moins de temps en s'attaquant directement au texte source.

Mais si la traduction entièrement automatique de haute qualité – ou *fully automatic high quality translation* (FAHQT), pour utiliser l'acronyme célèbre de Bar-Hillel – demeure encore hors de notre portée aujourd'hui, tout comme elle l'était dans les années 50, cela ne veut pas dire que les ordinateurs ne peuvent rien faire pour épauler les traducteurs humains qui doivent produire des traductions de qualité et qui se trouvent de plus en plus débordés. Si nous somme prêts à abandonner l'objectif d'une automatisation complète du processus de la traduction en faveur d'un modèle collaboratif où l'humain demeure le maître d'œuvre et la machine joue un rôle d'adjoint, il y a plusieurs façons pour cette dernière de se rendre utile. Il s'agit ici du paradigme de la traduction *assistée* par ordinateur, plutôt que de la traduction automatique proprement dit. Depuis plusieurs années, le laboratoire RALI s'efforce de développer une nouvelle génération d'outils d'aide à la traduction qui exploitent de

façon originale les techniques de l'approche empiriste à base de corpus. Nous allons maintenant en présenter quelques exemples.

3.6.2.4. TransSearch

TransSearch[3] est un concordancier bilingue (figure 3.6.3), c.-à-d. un système d'interrogation de base de données qui permet à un utilisateur de soumettre des requêtes à une archive de textes traduits afin d'y retrouver des solutions toutes faites à ses problèmes de traduction.

[3] http://www.tsrali.com

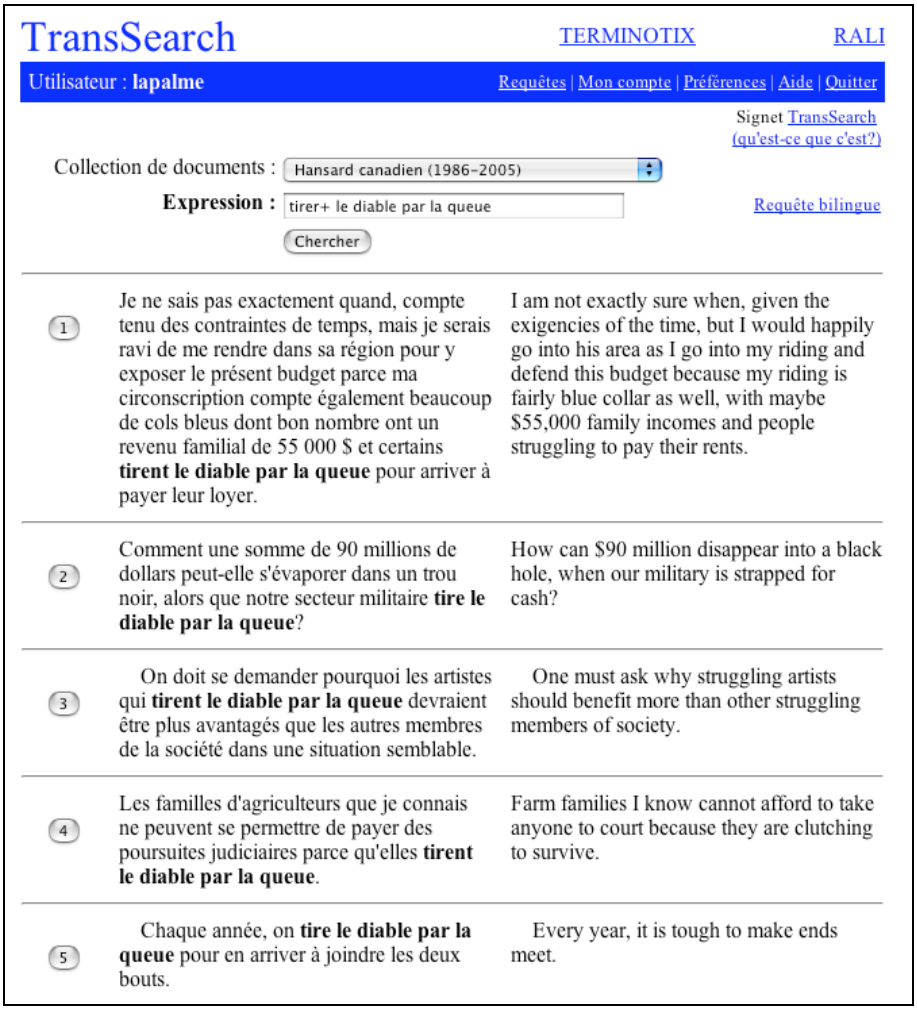

Figure 3.6.3 : Recherche dans un concordancier bilingue de différentes traductions de l'expression *tirer le diable par la queue* dans des textes en français et affichage des phrases correspondantes en anglais.

Plus précisément, cette archive est une mémoire de traduction constituée de groupes de documents qui sont des traductions mutuelles, et dans laquelle les liens phrastiques qui existent entre les traductions sont consignés de façon explicite. Dans la plupart des mémoires de traduction commerciales, cette base bi-textuelle est assortie d'un programme supplémentaire qui effectue une recherche systématique de chaque phrase source dans un nouveau texte à traduire ; s'il en trouve une copie dans la base, le programme indique la phrase cible qui lui est associée. Ici encore la *compréhension* associée à ce mode de fonctionnement reste minimale de la part du

système : la recherche est effectuée au niveau du traitement des chaînes de caractères sauf pour la reconnaissance d'entités particulières telles que les expressions numériques ou les dates qui sont ignorées lors des comparaisons de chaînes. Cependant, hormis les mises à jour et certains types de manuels techniques, la répétition intégrale de phrases complètes est un phénomène assez rare dans la plupart des textes. D'où l'intérêt d'un concordancier interactif comme TransSearch. Dans ce système, c'est à l'utilisateur d'initier une recherche, mais il n'est pas contraint de soumettre une phrase complète ; il peut interroger la mémoire pour y trouver la traduction d'un mot, d'un terme ou d'une expression de longueur variée. Comme le démontre l'exemple dans la Figure 3.6.3., TransSearch est en mesure de repérer toutes les formes fléchies de l'expression en question ; et chaque résultat est affiché vis-à-vis de la phrase qui contient sa traduction. L'utilisateur n'a qu'à se servir.

Le modèle de traduction qui soutient ce système est très simple, voire simpliste. La seule chose qu'il *comprend* des paires de textes qui composent son archive est le fait qu'il dispose d'une table indiquant qu'une certaine phrase *x* en langue L1 est la traduction d'une autre phrase *y* en langue L2. En revanche, le programme d'alignement automatique qui crée cette base bi-textuelle est capable d'apparier des millions de phrases avec très peu d'erreurs et sans aucune intervention humaine. D'ailleurs, la taille des ses bases et la qualité des traductions sont sans conteste ce que les utilisateurs apprécient le plus dans TransSearch. En date d'avril 2005, le système leur donnait accès aux bases de données suivantes, toutes deux en anglais et français :

- le Journal des débats à la Chambre des communes du Parlement canadien (Hansard) depuis avril 1986, environ 252 millions de mots ; et
- des décisions de la Cour suprême du Canada, de la Cour fédérale du Canada et de la Cour canadienne de l'impôt, à partir de 1986, environ 96 millions de mots.

Dans des archives d'une telle envergure, il est rare qu'un utilisateur ne trouve aucune solution à son problème de traduction.

3.6.2.5. *TransType*

Alors que TransSearch est relativement simple du point de vue technologique, TransType [EST 04] est un outil d'aide à la traduction beaucoup plus ambitieux qui mise sur l'interactivité. Le système (figure 3.6.4) observe le traducteur qui tape son texte cible (partie droite) correspondant au texte original (partie gauche) et propose des extensions à la traduction. Le traducteur peut les accepter telles quelles, les modifier, ou ne pas en tenir compte en continuant simplement à taper. Après chaque caractère entré ou éliminé par le traducteur, TransType recalcule ses prédictions sur la suite du texte cible et propose une nouvelle complétion.

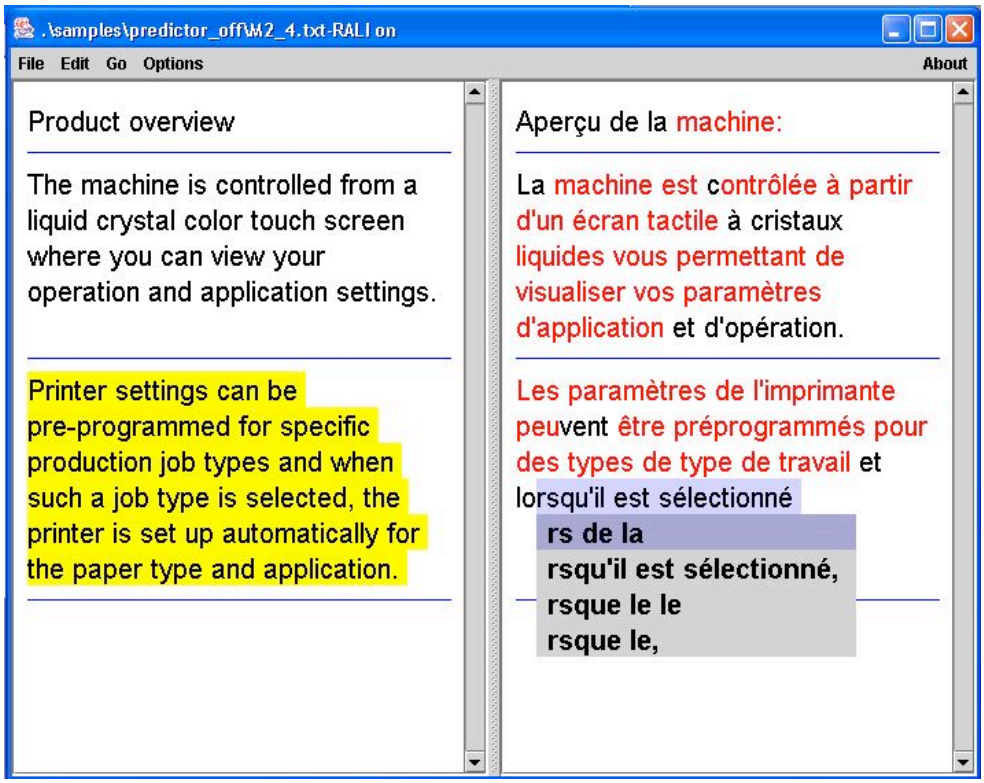

Figure 3.6.4. Cliché d'écran de TransType lors d'une séance de traduction du texte source dans la partie gauche. Le traducteur tape dans la partie droite et le système propose des extensions à l'aide d'un menu. On indique ici en rouge (ils apparaissent en gris sur la reproduction en noir et blanc) les caractères suggérés par le système et acceptés par le traducteur.

Un tel système peut aider le traducteur de différentes façons :

- lorsque le traducteur vise un certain texte cible, les complétions du système peuvent en accélérer la saisie ;
- si le traducteur n'est pas habile au clavier, l'interface vocale du système lui permet de sélectionner des complétions et les insérer dans le texte cible vocalement ;
- lorsque le traducteur cherche une formulation adéquate, les propositions du système servent de suggestions.

L'originalité de ce système réside dans le mode d'interaction entre l'utilisateur et le moteur de traduction automatique qui est enchâssé dans TransType. Traditionnellement, les systèmes de TA interactifs font appel à l'utilisateur afin de les aider à analyser le texte source ; une fois le texte source désambiguïsé grâce à l'intervention de l'humain, le système reprend le contrôle et produit sa traduction automatique-

ment. Cela est peut-être justifié dans certaines circonstances, par exemple lorsque l'utilisateur ne maîtrise pas la langue cible ; mais généralement, les traducteurs n'apprécient pas cette approche parce qu'elle demande des compétences (en linguistique formelle) qui ne sont pas forcément les leurs. Dans TransType, en revanche, l'objet des interactions entre l'utilisateur et le système est le texte cible, ce qui rejoint plus naturellement les compétences d'un traducteur. Qui plus est, le traducteur demeure le maître d'oeuvre du processus, c'est lui qui a le dernier mot. Si les complétions proposées par le système ne correspondent pas au texte cible qu'il a en tête, il est toujours libre de les laisser de côté et de taper sa propre traduction.

Ce type d'interactivité pose un certain nombre de défis techniques. D'abord, contrairement à des systèmes de TA classiques, TransType doit tenir compte non seulement du texte source mais également du préfixe, ou de la partie du texte cible que l'utilisateur a déjà tapé. De plus, il doit calculer plusieurs complétions, pour tenir compte du fait que différents traducteurs proposeront différents textes cibles, tous corrects. De surcroît, le système doit pouvoir fonctionner en temps réel, de façon suffisamment rapide pour suivre l'utilisateur pendant qu'il tape. Ces raisons justifient l'utilisation d'un système de TA stochastique au sein de TransType. Finalement, le système doit afficher ses prédictions d'une manière conviviale, sans gêner le traducteur, tout en permettant l'insertion facile de ses complétions. TransType ne *comprend* donc pas l'entrée du traducteur, il se borne à calcuer très rapidement une suite de mots qui serait une suite probable selon les valeurs de probabilités conservées dans les modèles de langue et de traduction.[4]

3.6.2.6. *Vérification automatique de traduction*

Dans les cabinets de traduction, la tâche de révision est normalement confiée à des traducteurs chevronnés, vraisemblablement parce que la détection et la correction d'erreurs de fond ou de style exigent plus d'expérience et une meilleure compréhension que celles que possèdent les néophytes. De façon analogue, on peut supposer que la vérification *automatique* de traduction est une tâche encore plus difficile que la traduction automatique elle-même. Cela dit, certains éléments de la révision sont constitués de tâches plutôt mécaniques et minutieuses, et si l'on pouvait les confier à un système informatisé, les réviseurs en seraient très reconnaissants. C'est précisément le but du système TransCheck [JUT 00], un outil conçu pour aider les traducteurs à déceler automatiquement certaines erreurs de traduction et à faire respecter certaines normes linguistiques au sein d'un grand service.

---

[4] Une démo en ligne du système TransType est disponible à
`http://rali.iro.umontreal.ca/Transtype2/Demo/index.fr.html`

Comment fonctionne cet outil d'aide à la révision ? Avant d'effectuer ses vérifications, TransCheck doit d'abord aligner les phrases du texte source et du texte cible. Ensuite, le système TransCheck vérifie les correspondances entre les paires de segments ainsi appariés pour s'assurer qu'elles ne contiennent pas de correspondances interdites et qu'elles respectent certaines correspondances obligatoires. Et c'est précisément cet aspect *bi-textuel* des vérifications qui distingue TransCheck d'autres types de vérificateurs, que ce soit d'orthographe ou de grammaire. Étant monolingues, ces derniers ne peuvent détecter aucune erreur de traduction, même parmi les plus flagrantes, car la traduction, comme nous l'avons déjà vu, met en relation *deux* textes.

Voici quelques exemples de types d'erreurs que TransCheck peut déceler :

omissions : il arrive plus souvent qu'on ne le penserait qu'un traducteur omette par accident une phrase, voire un paragraphe entier. Ce problème peut arriver, par exemple, suite à une fausse manipulation du copier/coller. En exploitant ses algorithmes d'alignement, TransCheck peut détecter de telles omissions.

expressions numériques : la retranscription incorrecte d'un numéro, d'un sigle alphanumérique, d'une date ou d'une somme d'argent est une autre erreur gênante que TransCheck permet d'éviter, d'autant plus que la vérification d'un texte comportant de nombreux chiffres est une tâche particulièrement fastidieuse.

interférences provenant de la langue source : TransCheck se sert d'un antidictionnaire de correspondances interdites pour déceler les faux-amis, les calques et les emprunts non autorisés.

cohérence terminologique : certains clients fournissent leur propre terminologie que le service de traduction doit respecter dans les textes livrés. Lorsque ces textes sont volumineux et qu'il faut les repartir entre plusieurs traducteurs, la cohérence terminologique peut devenir problématique. Là encore, TransCheck pourrait s'avérer utile, en signalant les écarts par rapport à la terminologie souhaitée.

TransCheck ne peut déceler toutes les erreurs de traduction, p. ex. les glissements de sens subtils ne sont pas détectés au moyen de ces mécanismes plutôt rudimentaires. Mais les vérificateurs d'orthographe non plus n'arrivent pas détecter toutes les erreurs dans un texte monolingue, ce qui ne les empêchent pas de rendre de fiers services aux utilisateurs. Et au fur et à mesure que notre compréhension de la traduction progresse, nous saurons étendre la portée de cette vérification automatique pour y inclure d'autres types d'erreurs. La *compréhension* dans TransCheck est le résultat des règles de vérifications formulées par les concepteurs et qui sont appliquées par le système dans le contexte de leur utilisation. Comme pour un détecteur de fautes d'orthographe monolingue, la décision finale revient au traducteur qui lui *comprend* la correspondance entre les textes source et cible.

3.6.3. Recherche d'information

Avec les moteurs de recherche sur le web, tout le monde connaît ce processus de recherche de documents à partir de quelques mots clés. Les moteurs permettant de retrouver des documents pertinents dans la même langue que les mots de la requête sont maintenant bien établis. Les moteurs cherchent les documents contenant ces mots et les trient par ordre décroissant de pertinence fondée sur des mesures comme la fréquence d'occurrence des termes de la requête dans le document ou en fonction du nombre de documents faisant référence à ce document.

Or, certains usagers sont parfois disposés à lire des documents rédigés dans une langue autre que leur langue maternelle, même si cette dernière demeure leur langue de choix pour formuler les termes de leur requêtes. Par exemple, comment peut-on trouver des documents écrits en chinois avec une requête en anglais ou français ? La réponse semble assez évidente : pour ce genre de recherche d'information *translinguistique*, il faut traduire la requête initiale dans la langue des documents. D'accord, mais comment ? Les systèmes de TA classique n'offrent pas nécessairement la meilleure solution, car ils ont été conçus pour traiter des phrases complètes et, face aux mots isolés d'une requête en RI, ils auront souvent de la difficulté à traduire ou à désambiguïser les différents sens d'un même terme. Par exemple, le mot *drug* en anglais peut signifier *drogue* ou *médicament* en français, selon le contexte. Supposons que nous nous intéressons aux traitements contre la déficience cognitive. Si on soumet une requête composée de deux termes, *drug* et *cognitive deficiency*, comment s'assurer que les résultats que nous obtiendrons portent sur le sens *médicament* ? Ici, la réponse n'est pas du tout évidente mais, chose certaine, les systèmes de TA classique feront un choix entre les deux, et pas toujours le bon, puisqu'ils sont programmés pour produire une et une seule traduction.

Une autre possibilité pour traduire la requête consisterait à utiliser des modèles de traduction statistiques (voir section 3.6.2.2), construits automatiquement à partir d'un ensemble de textes parallèles. Ces modèles de traduction ne sont pas limités à un équivalent par terme ; au contraire, pour chaque mot de la requête ils fourniront tout un ensemble d'équivalents probables. Cet ensemble de requêtes serait alors repris par le moteur de recherche qui exploiterait ses critères habituels afin de repérer les documents les plus pertinents. Dans l'exemple donné ci-dessus, il y aura de bonnes chances à ce que le système trouve plus de documents en français qui contiennent à la fois *déficience cognitive* et *médicament* que *déficience cognitive* et *drogue*. Cette méthode se révèle économique, car elle ne requiert aucune intervention manuelle, et l'on obtient de meilleures performances qu'avec un système de traduction automatique classique [CLE 03].

3.6.4. Extraction d'information

L'extraction d'information est un raffinement de la recherche d'information en ce sens qu'on cherche à extraire des informations précises d'un corpus de documents et non seulement d'obtenir des références à des documents. Par exemple, on voudrait connaître les différents aspects importants d'un événement : personnes impliquées, période de temps, principaux enjeux. Et ici aussi, il peut y avoir un volet translinguistique dans le cas où on voudrait combiner des informations provenant de documents dans des langues différentes.

Un autre contexte d'application est lorsqu'on demande à un système de répondre à une question comme *Dans quelle ville est située la tour Eiffel ?* Le système analyse la question pour en déterminer le type de question et il lance une recherche d'information dans une base documentaire à partir des termes de la question pour identifier des passages les plus susceptibles de contenir la réponse. Ces passages sont ensuite analysés pour en extraire des entités sémantiques particulières (noms de personnes, dates,…). En fonction du type de réponse recherchée (date, lieu…) l'entité la plus probable est extraite et est retournée comme réponse à l'utilisateur.

Et bien sûr, les informations ciblées peuvent se trouver dans des documents rédigés dans une langue différente de la requête. Dans ce cas-ci, il faudra non seulement traduire la requête dans la langue des documents, mais aussi traduire la réponse extraite des documents dans la langue de l'usager. Ici encore, les modèles statistiques de traduction se révèlent fort utiles, permettant de traduire rapidement les termes de la requête et de les combiner avec le moteur de réponse automatique.

3.6.5. Conclusion

Ce chapitre a discuté de la compréhension par rapport à la traduction à travers différentes applications d'aide à la traduction qui, chacune à leur manière, contournent le problème. Le résultat du processus de traduction étant un autre texte plutôt qu'une série d'actions, il est donc possible de prendre certains raccourcis dans la compréhension pour agir au niveau des *réflexes langagiers* pour reporter le problème au lecteur du texte cible.

Pour conclure, nous aimerions reformuler une dernière fois la question que nous avons posée au tout début de ce chapitre :

> *Comment peut-on évaluer la compréhension multilingue d'une machine autrement que par sa manifestation dans une tâche intelligente dont la traduction est peut-être la plus simple à mesurer ?*

Posée de cette façon, la question rappelle les débats qui ont fait rage dans les années 50 sur la capacité de penser de ces ordinateurs numériques qui venaient de faire leur apparition et qui s'imposaient dans de plus en plus de domaines. À cette époque, Alan Turing ne croyait pas que la question était posée de façon sensée, ce qui l'a amené à proposer le fameux jeu de simulation qui porte son nom. C'est dans ce même esprit du test de Turing que nous serions tentés de répondre à notre question reformulée ci-dessus. Dans la mesure où une machine arrive à bien traduire un ensemble de textes, nous serions tout à fait disposés à lui attribuer la compréhension dont les traducteurs estiment essentielle à cette opération.

3.6.4. Bibliographie

[BER 96] Berger, A., Della Pietra S., and Della Pietra V., « A Maximum Entropy Approach to Natural Language Processing », Computational Linguistics, 22 (1), 1996.

[BRO 93] Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R., « The mathematics of statistical machine translation : Parameter estimation », Computational Linguistics, 19 (2), pp. 263-311, 1993

[CLE 03] CLEF 2003, « Comparative Evaluation of Multilingual Information Access Systems » Fourth Workshop of the Cross-Language Evaluation Forum, Trondheim, Norway, August 2003. Revised papers. Lecture Notes in Computer Science 3237, Carol Peters, Julio Gonzalo, Martin Braschler, Michael Kluck (Eds.). Springer 2004

[DEB 05] Debili, F., Souissi E., « Y a-t-il une taille optimale des règles de succession intervenant dans l'étiquetage optimal », TALN 05, Dourdan, juin 2005, Vol 1, p 363-372.

[DES 04] Descamps P. « Les trahisons de la traduction », Science & Vie (hors série), no.227, juin 2004.

[EST 04] Esteban J., Valderrábanos J.. and Lapalme, G.. « TransType2 - An Innovative Computer-Assisted Translation System ». The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, p. 94--97, Barcelona, Spain, jul 2004 Association for Computational Linguistics

[FUC 93] Fuchs C, « Linguistique et traitements automatiques des langues », Paris, Hachette (Coll. Langue, Linguistique, Communication), 304 pages.

[HUT 92] Hutchins W. J., Somers H., 1992. « An introduction to Machine Translation », chapter 12, pages 207 – 220. Academic Press.

[JUT 00] Jutras J.-M., « An Automatic Reviser : The TransCheck System ». Applied Natural Language Processing 2000, Seattle, p 127-134.

[KAY 80] Kay Martin, « The Proper Place of Men and Machines in Language Translation », reproduit dans Machine Translation, vol. 12 (1997).

[KAY 92] Kay Martin, Forward to « An Introduction to Machine Translation », [HUT 92].

[LEP 04] Leplus T., Langlais P., Lapalme G.. «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, sep 2004 Springer

[NIR 90] Nirenburg S, Goodman K, 1990. « Treatment of Meaning in MT Systems », Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation, Austin, TX, 1990.

[OCH 04] Och F.J., Ney H. « The Alignment Template Approach to Statistical Machine Translation », Computational Linguistics, 30 (4), p 417-419, Dec 2004.

[MAN 99] Manning C. , Schütze H., « Foundations of Statistical Natural Language Processing », MIT Press. Cambridge, MA : May 1999.

[YNG 57] Yngve V.H. « A framework for syntactic translation », Mechanical Translation 4 (3), p. 59-65.