La problématique de la traduction automatique

Roland Raoul KOUASSI
Université de Cocody

Résumé

Les constantes mutations internes subies par l'activité de traduction témoigne de son importance dans les relations humaines. Elle continue désormais son développement en tirant profit des évolutions technologiques. Nous parlons de plus en plus de *traduction automatique.* Cette pratique poursuit sa conquête dans les habitudes aussi bien des universitaires que des professionnels de tous les domaines. Dans cet article, nous faisons le point de la traduction automatique, à la lumière des résultats des théories des opérations en sciences du langage. En effet, quelles que soient les méthodes ou procédures utilisées, la traduction demeure une activité linguistique. Il s'agit en effet d'énoncés au départ et à la fin du processus. Nous attirons l'attention des populations intéressées par la traduction automatique sur le fait que, bien que les progrès technologiques proposent des résultats assez satisfaisants par endroits, il demeure de nombreuses insuffisances en raison des micro- et macrophénomènes dynamiques qui font la particularité de l'activité langagière. Nous sommes encore loin de ce jour où les outils informatiques produiront des traductions identiques au travail de l'Homme. Demain, peut-être.

Mots clés: Traduction automatique, linguistique des opérations, statistique, arbre syntaxique, corpus, programmation.

Abstract

The continuous changes occurring within the activity of translation show its necessity in human relations. In these changes translation is taking profit from technological advancements. Machine translation - which is the result of this interrelation - is taking a place increasingly important in the habits of academics as well as professionals of all industries. This article is a review of the situation of machine translation under the scrutiny of the results provided by the contemporary linguistic theories of operations. Indeed, whatever methods or procedures used, translation remains a linguistic activity. It deals with utterances from source to target. We hereby draw the attention of people with interest in machine translation that, although current developments in computer science help to come up with satisfactory results, at times, there still are many inadequacies because of the dynamic micro- and macrophenomena typifying language fact. The day computers will give translation identical to those of Human being has not yet come. Tomorrow, maybe.

Keywords: Machine translation, linguistics of operations, statistics, syntactic tree-diagram, corpus, programming.

Introduction

Les différences entre langues constituent un défi majeur pour les opérateurs de ce “village” devenu planétaire, selon les prédictions de Marshall McLuhan[1]. Il se pose un problème de compréhension des idées et d’optimisation des échanges socioculturels, politiques, économiques, etc. Pour pallier ces problèmes et permettre aux opérateurs de mettre en place un continuum harmonieux des sens et significations, il y a un usage de plus en plus développé et important des pratiques de traduction. La traduction est un processus dynamique à travers lequel un message encodé dans une langue (langue de départ) est décodé puis re-encoder dans une autre langue (langue d’arrivée), en respectant aussi bien le fond que la forme dudit message. Cette activité permet de briser les barrières linguistiques en créant un espace de distribution harmonieux des significations, dans un monde de collaboration et de partenariat.

La pratique de la traduction a connu plusieurs développements qui ont progressivement affinées les techniques et méthodes pour des résultats de plus en plus améliorés. Ces techniques et méthodes s’enseignent dans les universités et écoles de traduction pour permettre au futur traducteur ou interprète d'être s'imprégner des principes du métier.

Toutefois, les bouleversements socio-économiques et géopolitiques vont conduire à la recherche et au développement de nouveaux procédés pour "optimiser" le travail de l’Homme. En effet, les relations humaines qui se mondialise sont désormais caractérisées par la rapidité du flux des informations et l’urgence. On a besoin de millions d’informations, très rapidement. Le mot d’ordre est clair : rapidité et efficience. L’Homme n’est plus assez rapide. On va alors faire appel à d’autres méthodes. L’ordinateur fit alors son entrée dans la sphère, jusque là très humaine de l'activité traduisante. Dans cet article, nous soumettons la traduction automatique à la dynamique des opérations langagières. Il s’agit de mesurer la fécondité de la traduction automatique à l’épreuve des théories linguistiques des opérations.

---

[1] Au début des années 60, McLuhan prédit la fin de la culture du visuel et de l’impression individualiste au profit de ce qu’il a appelé " interdépendance électronique ". Cet état de fait aboutira sur une identité collective au sein d’une nouvelle organisation sociale qu’il a nommée "Village planétaire". (voir *The Gutenberg Galaxy: The Making of Typographic Man*, 1st Ed.: Univ. of Toronto Press, 1962)

1. Traduction automatique : généralités

L'automatique c'est l'ensemble des disciplines scientifiques et des techniques utilisées pour l'automatisation de processus, la conception et l'emploi des systèmes automatiques. La traduction automatique ou TA est un sous-domaine de la linguistique informatique qui travaille à la théorie et pratique de l'utilisation de l'outil informatique, *uniquement*, pour la traduction des textes écrits et oraux d'une langue naturelle à une autre. On l'appelle également traduction assistée par ordinateur (TAO) ou traduction logicielle. C'est un processus basé sur des faits de substitution à partir de systèmes de corpus, de typologie, de reconnaissance d'énoncés ou de reconnaissance vocale, de rapports entre expressions idiomatiques, proverbes, termes et expressions jargonneux, etc., ainsi que l'isolation des anomalies.

Il est généralement admis que deux expériences majeures ont jeté les bases de la traduction automatique dans les années 1950: les expériences de Georgetown et les applications de Birkbeck College (University of London). En 1966, les rapports des l'ALPAC ont ensuite joué un rôle important sur les méthodes de calcul de performance des différents logiciels proposés. Toutefois, retenons qu'en 1629, dans son traitement de la langue universelle, le philosophe René Descartes postulait déjà plusieurs faits de langues représentables par un symbole unique.

2. Les paradigmes de la traduction automatique

La traduction automatique s'est développée autour de trois paradigmes essentiels :

a. *La traduction symbolique*

En traduction symbolique, des experts encodent explicitement leurs connaissances ; un système les utilise ensuite pour traduire un document source.

b. *Le paradigme des règles*

Un trait caractéristique des systèmes basés sur les règles a été la transformation ou 'mappage' des représentations en forme d'arbres étiquetés. Ces systèmes formalisent des règles de réécritures et proposent une série de transformations des arbres: un arbre morphologique est transformé en un arbre syntaxique, un arbre syntaxique en un arbre sémantique, un arbre d'interface du texte source en un arbre équivalent du texte cible, etc.

Un arbre doit satisfaire des conditions précises: posséder une structure particulière et contenir des unités lexicales particulières ou des traits syntactiques ou sémantiques particuliers. En plus, les arbres eux-mêmes sont testés par les règles de formation et de transformation.

Un arbre est rejeté s'il n'est pas conforme aux règles grammaticales du niveau en question: morphologie, syntaxe, sémantique, etc. Les grammaires et les règles de transformation déterminent les conditions ou contraintes qui limitent les possibilités de transfert d'un niveau à un autre et, en somme, d'un texte de la langue source à un texte de la langue cible.[2]

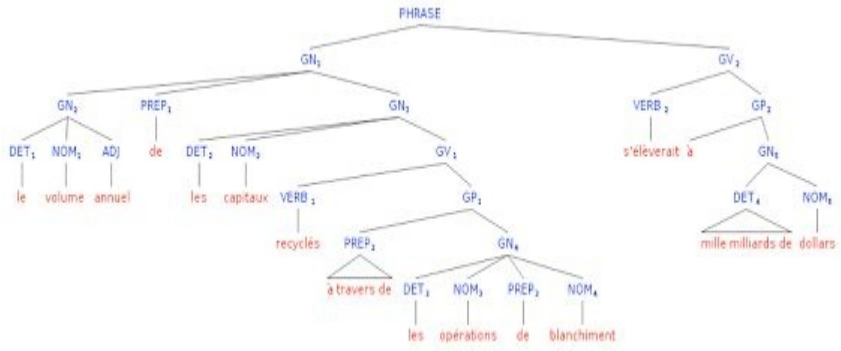


Figure 1: Une représentation arborescente de la phrase *Le volume annuel des capitaux recyclés a travers les opérations de blanchiment s'élèverait a mille milliards de dollars.*

[2] Voir John Hutchins, "Vers une nouvelle époque en traduction automatique", *Troisièmes Journées Scientifiques* LTT, Montréal, 30 septembre 1993

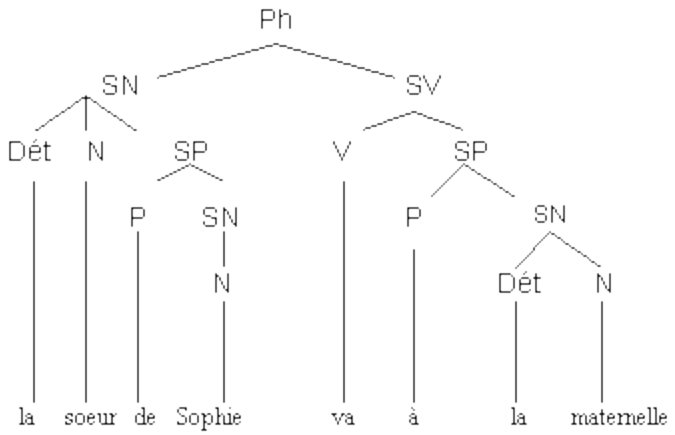


Figure 2: Une représentation arborescente de *La soeur de Sophie va à la maternelle.*

Ces arbres comportent des branches et des nœuds :

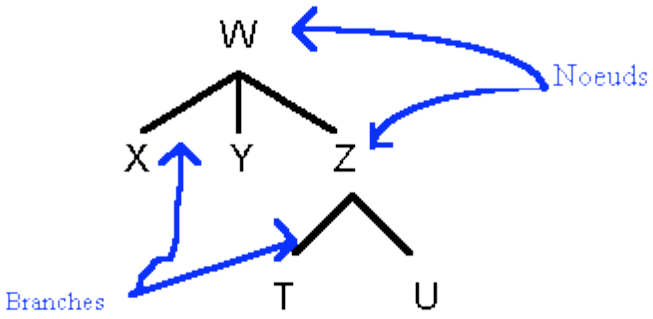


Figure 3 : La composition d'un arbre syntaxique

Cette méthode des règles tire sa source des travaux des linguistes structuralistes (Zellig Harris, Leonard Bloomfield, Charles Hockett, etc.) et linguistes générativistes (Noam A. Chomsky) a qui l'on doit les analyses en constituants immédiats et les règles de réécriture de la formation et transformation des phrases.

c. *Le paradigme statistique*

Le paradigme statistique repose sur l'exploitation de corpus de traduction qui permettent de déterminer la traduction la plus fréquemment utilisée, par les spécialistes, pour une expression donnée.

Cette technique a été vulgarisée par *Google Translate* :

> Notre système adopte une méthode différente: d'une part, nous introduisons dans l'ordinateur des milliards de mots provenant de

> textes monolingues dans la langue cible; d'autre part nous ajoutons des textes mettant en parallèle les deux langues. Ces derniers sont créés à partir d'échantillons de traductions réalisées par des traducteurs professionnels. Nous appliquons ensuite des techniques d'apprentissage statistique pour créer un modèle de traduction. Nous avons obtenu d'excellents résultats dans le cadre d'évaluations réalisées dans ce domaine.[3]

Pour répondre aux insuffisances que présente de plus en plus la méthode des règles, en raison de la complexité du comportement des langues naturelles, certains laboratoires et groupes de recherche s'orientent vers les méthodes statistiques. Les ordinateurs devenant de plus en plus puissants, il est désormais possible de puiser dans les immenses corpus des bases de données informatisées pour réutiliser des fragments de phrases déjà traduites par les traducteurs professionnels. C'est la naissance de la méthode de traduction statistique. Elle est initiée par IBM, mais atteint son apogée dans les années 2000 avec Google qui récupère toutes les traductions existant sur Internet pour construire l'architecture de son outil de traduction *Google Translate*. Cette approche statistique s'appuie sur des corpus bilingues *alignés*. En effet, un lien est créé entre chaque partie du texte de la langue source et la partie correspondante dans la langue cible. Ce lien est généralement créé au niveau de la phrase. Une analyse statistique utilise les redondances existant dans ces corpus pour estimer les paramètres du processus de traduction. « La traduction statistique est possible car les modèles *ad hoc* sont couplées avec des algorithmes de programmation dynamique qui maximisent une fonction de traduction d'une phrase source vers une phrase cible ».[4]

Enfin, l'approche statistique nécessite la définition d'un modèle de traduction qui puisse calculer efficacement les probabilités de traduction entre les mots, les suites de mots et les autres constituants de la phrase.

d. *Les tendances actuelles*

---

[3] http://www.google.fr/intl/fr/help/faq_translation.html#statmt

4 Caroline Lavecchia, Kamel Smaili, David Langlois, « Une alternative aux modèles de traduction statistique d'IBM: Les triggers inter-langues », *LORIA/Speech Group*, Campus scientifique, Juin 2008. Dans cet article ambitieux, les auteurs proposent "approche permettant de construire un modèle de traduction fondé sur les triggers inter-langues (extension des triggers classiques) pour construire [leur] système de traduction statistique. Le concept de triggers est bien connu de la communauté de la modélisation statistique du langage. Facile à mettre en œuvre, il possède une certaine souplesse qui permet de l'appliquer à différents niveaux de lecture de la phrase (mots, genre, nombre, constituants syntaxiques)"

Les logiciels de traduction automatique pratique de plus en plus des méthodes d'intégration de différents paradigmes pour l'efficacité des résultats. Ils combinent les paradigmes pour bénéficier des forces de chacun dans un système bien construit: c'est le modèle hybride. *Systran*, *Google Translate* et *Bing Translator* opèrent un système hybride.

Systran exploite un moteur de traduction hybride qui intègre une analyse statistique à l'analyse sémantico-syntaxique traditionnelle du texte source. Cette approche permet au logiciel de choisir la solution la plus fréquente entre deux propositions du moteur sémantico-syntaxique. En plus, elle intègre un module d'amélioration continue.

Ce moteur hybride permet à Systran de se positionner en leader du marché, à ce jour. Auparavant, la méthode utilisée par les logiciels était fondée sur un système d'analyse sémantico-syntaxique. Le moteur analysait chaque phrase source et créait l'arbre syntaxique permettant de représenter ses composantes et les relations qui les unissent. Puis, chaque expression était traduite en faisant appel à un dictionnaire. Une fois l'arbre entièrement traduit, le logiciel restituait la phrase cible. Le dictionnaire constitue alors un élément central: plus il est complet, meilleur est le résultat. Pourtant, même avec des dictionnaires très fournis, il est presque impossible de produire une phrase cible totalement correcte dans la mesure où le dictionnaire, qui est un recueil de données lexicales aura du mal à rendre compte des mots et expressions contextualisés ou nouveaux.

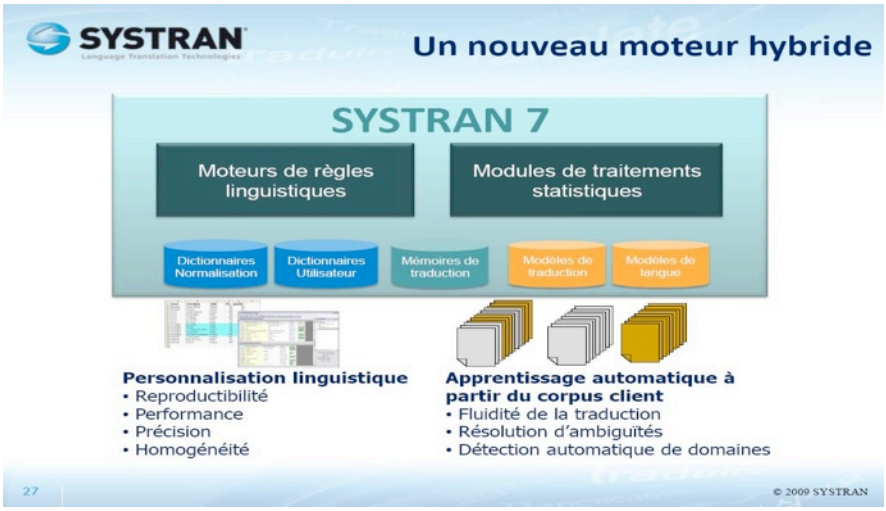


Figure 4 : La base architecturale de *Systran 7*

3. Le problème de la désambiguïsation en traduction automatique

Les ambiguïtés constituent l'un des défis du processus de traduction automatique. Ces ambiguïtés sont principalement d'ordre sémantique et d'ordre syntaxique. En effet, l'abondance des équivalents et les nuances liées à la complexité des structurations syntaxiques ont du mal à trouver des solutions automatiques efficaces.

Exemples:

Tu es malade. *You're sick. You're ill. You're nuts. You're out of your mind. ...*

Je ne mange pas. *I don't eat. I'm not eating. I won't eat.*

De nombreuses recherches sont entreprises pour résoudre ces problèmes d'ambiguïté et aboutir à une désambiguïsation qui puisse permettre à la machine de rendre les significations souhaitées pour la traduction.

La désambiguïsation attribue à chaque unité lexicale une étiquette unique (partie du discours et information morphologique de base) en contexte. Parmi les expériences de désambiguïsation tentées depuis quelques années, les chercheurs ont eu recours à de nombreuses techniques dont la sélection "statistique" des mots traduits de la requête, le calcul basé sur la moyenne relative de la fréquence des termes, l'utilisation de plusieurs critères afin de déterminer le sens d'un mot dans un contexte, y compris les valeurs syntaxique, sémantique et pragmatique, les relations de cooccurrences syntaxiques, le développement des requêtes utilisant la mise en grappes des termes et des documents, etc.

En plus des techniques mentionnées ci-dessus, des outils ont été développés:

a. OpenMind Word Expert

Le *Open Mind Word Expert* (OMWE) est un système qui donne aux usagers la possibilité de pouvoir désambiguïser les mots. Toute personne qui visite le site d'OMWE a la possibilité d'apporter sa contribution aux significations des mots dans

des phrases données. Ainsi OMWE réussit-il à créer un vaste corpus étiqueté pouvant servir à la construction de systèmes de désambiguïsation automatique. Le principal objectif d'OMWE est de développer un logiciel intelligent, en partie grâce à la collecte d'un vaste ensemble de données et de fournir à l'usager une infrastructure ouverte où toutes les idées nouvelles sont les bienvenues. Les données et le logiciel qui en résultent sont ensuite mis à la disposition de tous.

b. Extended WordNet

Extended WordNet est un projet visant à développer un ensemble de ressources pouvant être utilisées sur des versions actuelles et futures de *WordNet*, de manière à perfectionner les applications du traitement linguistique. *WordNet* est une base de données lexicales anglaises adoptée dans de nombreuses applications d'intelligence artificielle et de linguistique informatique. Les éléments de base de *WordNet* sont un ensemble de mots reliés entre eux grâce à leurs relations sémantiques (synonymes, antonymes, etc.). La base *WordNet* est disponible au grand public et est largement utilisée dans la composition de bases de données multilingues. *Extended WordNet* propose des améliorations aux définitions, commentaires, exemples et mots se trouvant déjà dans *WordNet*. Ces glossaires sont ensuite analysés syntaxiquement. Chaque mot reçoit une étiquette de partie du discours et est ensuite relié à d'autres glossaires qui décrivent les concepts associés.

c. Framenet

Le projet *FrameNet* a pour principal objectif la création d'une ressource lexicale anglaise, basée sur la structure sémantique. *FrameNet* vise à documenter les nombreuses possibilités sémantiques et syntaxiques pour chaque mot pour chacune des significations possibles, grâce à des annotations manuelles. Ces annotations sont ensuite extraites et organisées. *FrameNet* étudie les mots, décrit la structure conceptuelle sous-jacente, examine les phrases en utilisant un vaste corpus de textes anglais contemporains contenant ces mots et enregistre les diverses manières dont l'information est exprimée dans ces diverses structures de phrases.

d. XeLDA

TEMIS SA (*TExt MIning Solutions*) a mis au point plusieurs outils pour l'extraction, la désambiguïsation, la catégorisation, la mise en correspondance et la création automatique de terminologie. En ce qui concerne la désambiguïsation, les textes sont soumis à une analyse utilisant différentes catégories de règles. Ces règles identifient chacun des mots d'un texte et attribuent une étiquette morphosyntaxique aidant à la désambiguïsation syntaxique. Parmi les outils conçus par Temis, XeLDA est un moteur linguistique multilingue qui permet de modéliser et standardiser les documents non structurés de façon à en exploiter automatiquement le contenu. Les services offerts par XeLDA incluent l'identification des langues, la segmentation, l'analyse morphologique, la désambiguïsation, l'extraction des phrases nominales, la vérification des contextes des mots dans un dictionnaire, les regroupements morphologiques. XeLDA est disponible pour douze langues spécifiques.

Tous ces outils ont pour objectif de traiter les structures microlinguistiques pour dévoilées les subtilités morphosyntaxiques et sémantiques.

4. Recherches en traitement de la langue naturelle

Les avancées en traduction automatique sont, en grande partie, basées sur les développements de plus en plus importants des recherches en traitement automatique des textes écrits ou oraux. Ces recherches sont fondées sur des sujets tels que les correcteurs d'orthographe, de grammaire ou de style, les systèmes de traduction automatique, les systèmes de dialogue homme-machine, les générateurs multilingues, l’analyse automatique de la morphologie, de la syntaxe, de la sémantique, de la pragmatique ainsi que des relations interactionnelles, l’alignement, l’extraction d'informations, le résumé automatique, les systèmes de questions-réponses, l’inférence textuelle, la génération automatique et la planification de textes, la communication homme-machine en langage naturel, etc. Ces traitements sont fondés, pour la plupart, sur des approches symboliques ou statistiques.

Plusieurs centres de recherche ont ainsi vu le jour pour étudier ces phénomènes et apporter des réponses efficaces. Nous pouvons citer:

- *Le Computing Research Laboratory* (CRL) de l'Université de New Mexico.

Le CRL est un centre à but non lucratif créé en 1983, avec pour objectif principal la recherche dans le domaine de l'intelligence artificielle, la linguistique informatique et l'interaction humaine avec les ordinateurs. Les principales applications développées par le CRL incluent une variété de configurations et de combinaisons de paires de langues pour la traduction automatique, l'extraction de l'information, le repérage d'information, la constitution de résumés, la recherche d'information multilingue, l'acquisition de connaissances, etc.

- *Le Center for Intelligent Information Retrieval (CIIR)*

C'est une institution faisant partie de la Fondation Nationale Scientifique de l'Université du Massachusetts. Ce centre fait partie des plus importants laboratoires de recherche d'informations du monde. La principale tâche de CIIR est de développer des outils donnant un accès plus efficace et plus rapide à de larges bases de données textuelles et multimédia. Le CIIR effectue des recherches sur le repérage d'informations monolingues, le filtrage, la détection de sujets, l'indexation et le repérage multimédia, le traitement des images, l'exploration textuelle, le repérage d'information multilingue, etc.

- *Le Johns Hopkins Center for Language and Speech Processing (CLSP)*

Ce centre existe depuis 1992 grâce au soutien du gouvernement américain (NSF, DARPA, DoD). Son principal objectif est de promouvoir la recherche et la formation pour les sciences et les technologies de la langue et du discours. Les recherches du CLSP touchent tous les domaines des sciences et des technologies de la langue et du discours. Parmi ces recherches figurent les études sur la modélisation de la langue, le traitement des langues naturelles, l'acquisition de la langue, etc. Le CLSP étudie également certaines tâches comme la désambiguïsation, l'analyse morphologique

(coréenne), l'étiquetage des parties du discours, etc.

- *Le Computational Linguistics and Information Processing Laboratory (CLIP)* de l'Université du Maryland Institute for Advanced Computer Studies (UMIACS)

Le CLIP est constitué de deux groupes principaux : le *Natural Language Group* et le *Database Group*. Le *Natural Language Group* étudie les nombreux domaines touchant le traitement linguistique multilingue : traduction automatique, détection translinguistique de documents et repérage d'information multilingue. Le *Database Group*, quant à lui, étudie les architectures utilisant de vastes sources de données. Parmi les projets de ce groupe, il y a la structuration des requêtes, la localisation de ressources de qualité grâce à l'utilisation de métadonnées, etc.

- Le *Natural Language Processing Group* de l'Université de Columbia

Ce groupe poursuit des recherches dans plusieurs domaines du traitement des langues naturelles incluant la génération et la production de résumés, la modélisation statistique des langues, les bibliothèques virtuelles, etc. Parmi les principaux projets de ce groupe, nous pouvons citer le projet Climb qui explore et développe des stratégies pour l'extraction automatique de métadonnées, le projet Magic qui est un système d'intelligence multimédia destiné au domaine médical et le projet Tides qui est un projet axé sur la détection de l'information, l'extraction et l'élaboration de résumés. Ce projet porte particulièrement sur le multilinguisme.

- L’association pour le traitement automatique des langues (ATALA)

Cette association, qui existe depuis 1959 est l’une des plus actives en France dans le domaine du Traitement Automatique des Langues (TAL) ainsi que d’autres domaines de la linguistique informatique. À l’origine, cette association était principalement tournée vers la traduction automatique, avec quelques réunions par an. Au fil des ans, cette association s’est amplement structurée :

- elle édite une revue (la revue TAL qui publie trois numéros par an, dûment

sélectionnés par un comité de lecture international),

- elle organise une conférence annuelle (la conférence TALN avec sa session étudiante RÉCITAL),
- elle soutient l'organisation de journées d'études,
- elle maintient une base de données sur les enseignements de TAL en France,
- elle est présente sur internet par un site web et une liste de diffusion. Cette structure lui permet de rassembler et de fédérer tous les acteurs de la communauté du TAL francophone.

- Le *Language Technologies Institute* (LTI) de *Carnegie Mellon University* (CMU)

Le LTI conduit des recherches variées en linguistique informatique, en traduction automatique, en reconnaissance et synthèse vocales, en apprentissage des langues assisté par ordinateur, en intelligence artificielle, etc.

Ce survol rapide qui présente un échantillon des centres de recherche permet d'appréhender l'importance accordée au traitement automatique des langues naturelles. Les énormes ressources allouées, par les institutions aussi bien privées que gouvernementales constituent est indice de l'importance de l'enjeu. La modélisation du langage et la traduction automatique font partie des postes privilégiés de dépenses de budget.

5. ***Les composantes de la traduction automatique***

Le paysage de la traduction automatique est composé de deux parties essentielles: la traduction automatique logicielle et la traduction automatique en ligne.

a. La traduction automatique logicielle

La traduction automatique logicielle, ou traduction hors ligne, est un processus basé sur un logiciel de traduction installé sur un ordinateur. Ces logiciels sont de plus en plus sophistiqués et offrent des possibilités de plus en plus appréciées par les utilisateurs. Ces derniers proviennent de différentes couches de la société: universitaires, professionnels du monde des affaires, étudiants, programmeurs, concepteurs de sites internet, etc.

Ces logiciels permettent, dit-on, de gagner du temps en évitant des recours interminables aux dictionnaires et encyclopédies volumineux.

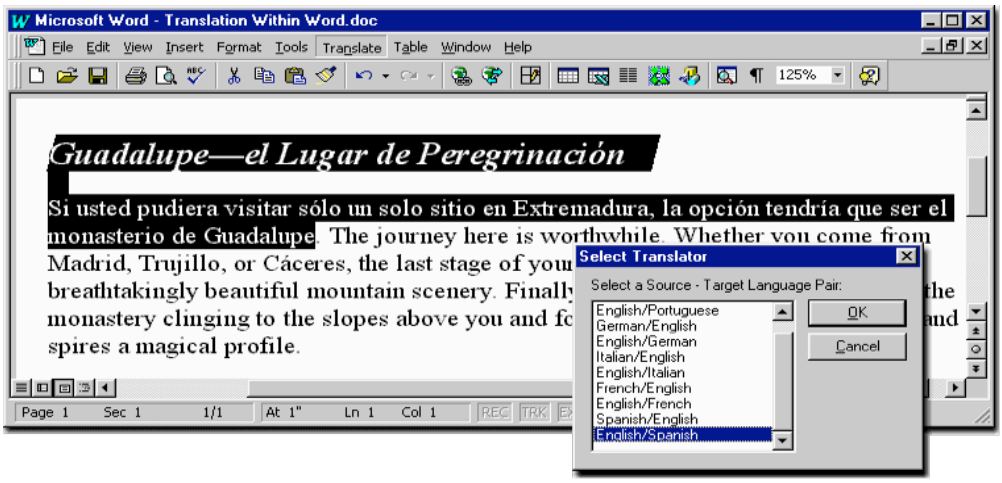
Le domaine de la traduction automatique a évolué très rapidement avec l'édition de logiciels de plus en plus performants. En voici quelques exemples:

- Translate Pro
- Systran Pro
- Power Translator Pro
- Babylon

Ces logiciels sont les plus utilisés.

La traduction automatique logicielle se fait en trois étapes essentielles, après le démarrage du programme :

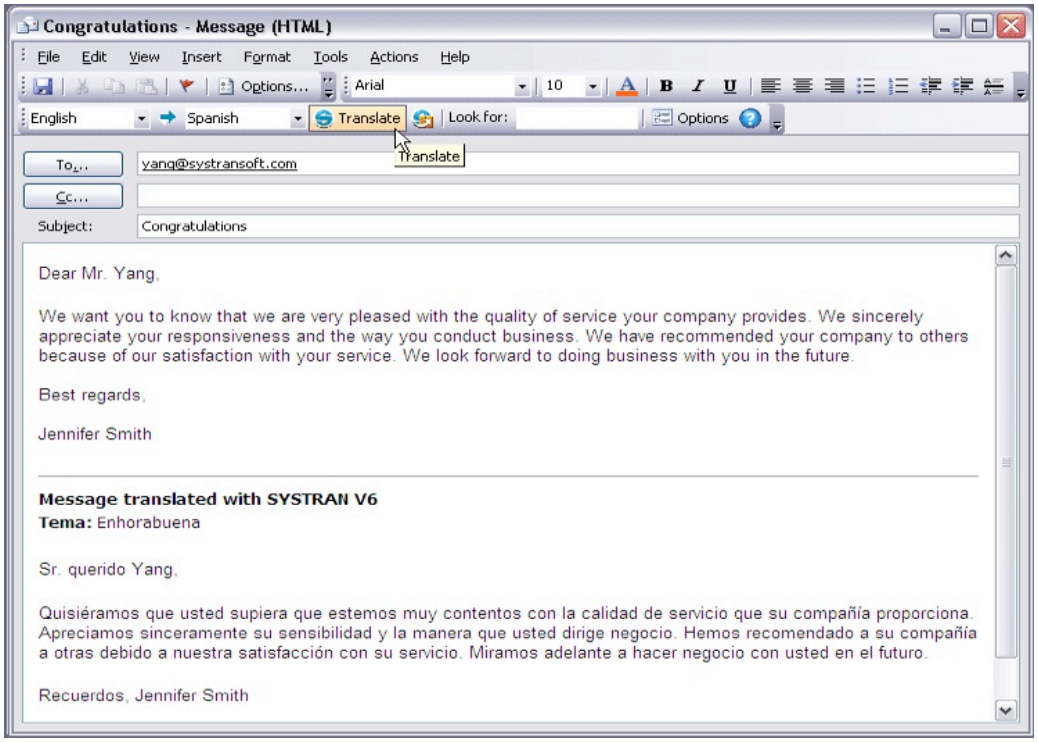
- Choix des langues de départ et d'arrivée (en fonction de la richesse linguistique proposée par le logiciel)



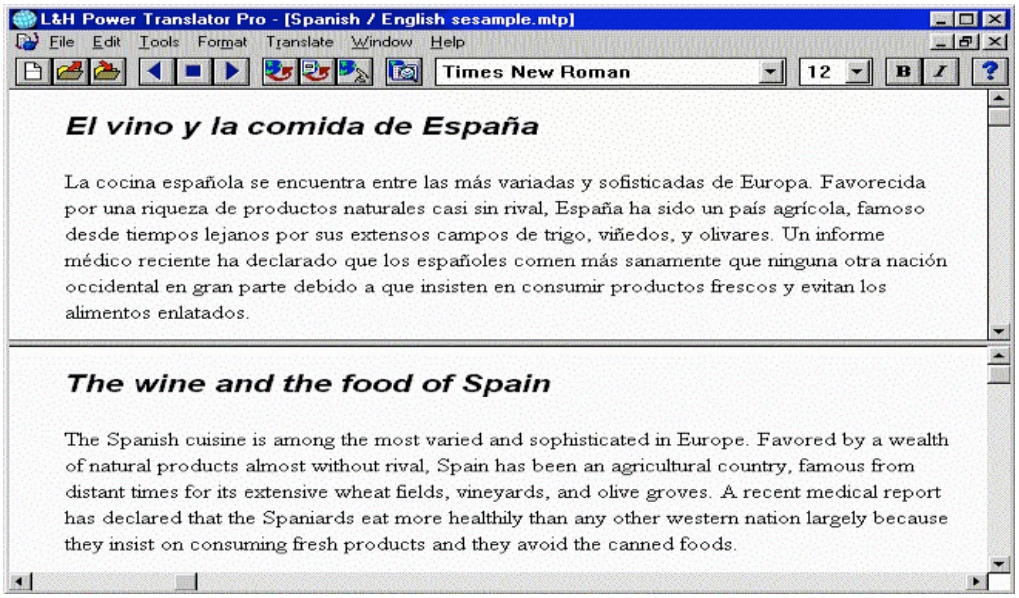
- Entrée du texte à traduire dans une fenêtre, celle du haut ou de gauche. Le texte peut être saisi à partir du clavier ou collé à partir du presse-papier.

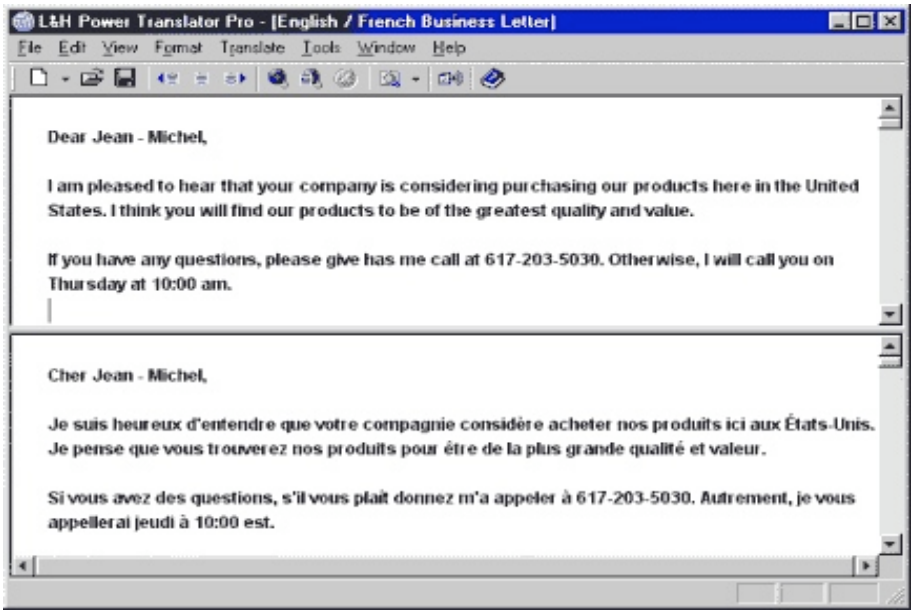
- Clic sur *traduire* ou *translate* et le texte traduit s'affiche dans la fenêtre de bas ou de droite.

Ces logiciels de traduction ont souvent des modules qui s'intègrent à Microsoft Office, à Apple iWork ou autres applications et permettent ainsi de faire des traductions directement à partir d'une fenêtre ouverte de *Word, Pages, Excel, Numbers, PowerPoint, Keynote, Outlook, Mail, Internet Explorer, Safari, Firefox*, etc.



L'illustration ci-dessus présente un message en anglais traduit avec Systran 6 en espagnol. Celle qui apparaît ci-dessous est une traduction par Power Translator Pro de l'espagnol vers l'anglais. Elle est suivie d'une autre traduction, également par Power Translator Pro, d'un énoncé de l'anglais vers le français.





b. *La traduction automatique en ligne*

La traduction automatique en ligne est un service de traduction de textes sur internet. Elle fonctionne fondamentalement de la même façon que la traduction logicielle hors ligne mais elle nécessite une connexion internet. En effet, le logiciel n'est soit pas

installé sur l'ordinateur, soit installé de façon très minimale (à travers, par exemple, les gadgets développés par Microsoft Windows Vista). Ces dernières années, la toile a connu une floraison d'outils permettant aux internautes de traduire instantanément des textes lorsqu'ils/elles font des recherches. Les traducteurs en ligne les plus utilisés à ce jour sont les suivants :

- SYSTRAN NET
- GOOGLE TRANSLATE
- PROMT
- REVERSO
- YAHOO BABEL FISH
- BABYLON
- BING TRANSLATOR

Ces outils, très pratiques permettent à l'internaute de faire traduire des résultats de recherche dans une langue qu'il ou elle comprend. L'utilisateur peut également utiliser le logiciel en ligne comme traducteur de textes divers, ou comme un dictionnaire.

6. **Dynamique des langues et traduction automatique**

La traduction automatique pose le problème des universaux du langage et celui des principes et paramètres de fonctionnement interne inhérents à chaque langue naturelle. Deux faits essentiels sont à relever ici: la dynamique de la métalangue naturelle et la dynamique du contexte d'énonciation.

a. La dynamique de la métalangue naturelle

La métalangue naturelle est le discours de la langue sur elle-même. Elle exprime les microphénomènes abstraits du travail *dans* la langue naturelle. La langue se replie ainsi sur elle-même pour rendre compte de son propre fonctionnement. La métalangue naturelle est le fondement de tout utilisation de la langue en contexte. La saisie de

cette métalangue permet d'expliquer les expressions ou effets de sens, qui ne sont que des *phénomènes*.[5]

La traduction automatique est basée fondamentalement sur des approches qui s'organisent autour du fait de langue en tant que phénomène, c'est-à-dire en tant que fait réalisé ou produit. On établit des correspondances entres mots et expressions, on écrits des algorithmes de traitement informatique pour programmer ces correspondances. L'ordinateur produira la réponse attendue lorsque le stimulus qui lui est associé est enclenché. Ce va et vient entre stimuli et réponses est le résultat de programmation des règles, des probabilités et des correspondances. On calculera par exemple les règles de concordance de temps grammaticaux, les accords au sein du groupe nominal, les accords au sein du groupe verbal, les positions dans l'ordre syntagmatique, les classes paradigmatiques, les niveaux discursifs... C'est un travail important qui permet d'aboutir à des résultats plus ou moins satisfaisants. A l'aide des statistiques, on modélisera les fréquences de *meilleures traductions* à partir de corpus de plus en plus riches et variés. Toutefois, le fait que les analyses soient focalisées sur des données ou corpus tirés de la chaîne linéaire des énoncés pose le problème de l'adéquation des règles et des programmes. Cet qui est à modéliser est le passage du *signifié de puissance*[6] au signifié d'effet, les opérations de mise en discours.

b. *La dynamique du contexte énonciatif*

Un texte est le résultat d'une énonciation. Cette énonciation est possible par des mécanismes dynamiques architecturés par un sujet énonçant dans une situation ou contexte d'énonciation. Tout énoncé produit est donc *nouveau* dans la mesure où les paramètres spatio-temporels sont modifiés à chaque instance. Ce qui sous-tend que la base de calcul du modèle informatique est constamment caduque. Cela suppose donc des mises à jour permanentes pour une efficacité toujours conservée ou améliorée. Systran a mis en place ce type de module qui s'autorégule pour une amélioration

[5] Phénomènes dans la tradition kantienne ou heideggérienne. Le phénomène est ce par quoi un fait se dévoile ou se fait voir. Cette apparition ou parution peut *cacher certains éléments* de l'être en-soi.

[6] Cette expression est empruntée à l'éminent linguiste Gustave Guillaume, le père de théorie linguistique de la psychomécanique du langage.

continue des textes traduits. Mais ce module fonctionne toujours à partir de corpus de phrases et expressions dépouillées de leurs contextes d'énonciation. Le modèle interprétera difficilement les sens.

c. Le traitement d'énoncés par quelques moteurs de traduction automatique

Nous avons testé quelques logiciels de traduction en ligne en leur soumettant les mêmes textes à traduire. Les résultats sont exposés ci-dessous[7] :

Enoncé 1:

He came home, walking like a snake! He was drunk! Smelling like a brewery!

Il est retourné chez lui, à pied comme un serpent! Il était ivre! Smelling comme une brasserie! (Google Translate)

Il est venu à la maison, marchant comme un serpent ! Il était ivre ! Sentir comme une brasserie ! (Systran)

Il est venu à la maison marchant comme un serpent ! il était ivre ! sentir comme une brasserie ! (Yahoo! Babel Fish)

Il est venu à la maison marchant comme un serpent! Il a été ivre! Sentant comme une brasserie! (Reverso Traduction)

Il est venu à la maison en marchant comme un serpent! il était bu! sentir comme une brasserie! (Prompt Translation)

Il est arrivé chez, marche comme un serpent ! Il était ivre ! Sentir comme une brasserie ! (Bing Translator)

Dans ces résultats, nous constatons que le temps grammatical de *came* a plus ou moins été respecté. Toutefois, aucun des moteurs n'a utilisé la notion lexicale de *rentrer (à la maison).* Bing Translator a utilisé *arriver,* mais après *chez,* rien d'autre. Dans le reste des traductions, il y a beaucoup de cacophonie et de non sens.

Enoncé 2:

The animal was cooking outside.

L'animal a été en dehors de la cuisine. (Google Translator)

L'animal faisait cuire dehors. (Systran Net)

[7] Les énoncés à traduire sont en gras et les traductions proposées en italique.

L'animal faisait cuire dehors. (Yahoo Babel Fish)

L'animal cuisinait à l'extérieur. (Reverso Traduction)

L'animal a été de cuisson à l'extérieur. (Bing Translator)

Ici, à part Reverso, le reste est inacceptable. Toutefois, il faut relever l'ambiguïté dans cet énoncé; ce qui peut poser problème à la réponse de Reverso, qui est pourtant grammaticalement correcte. L'animal cuisine-t-il? Ou est-il cuisiné? Ces cas sont difficilement gérés par le moteur de traduction automatique.

Enoncé 3:

I think you should see a doctor, brother.

Je pense que vous devriez voir un médecin, d'un frère. (Google Translator)

Je pense que vous devriez voir un docteur, frère. (Systran Net)

Je pense que vous devriez voir un docteur, le frère. (Reverso Traduction)

Je crois que vous devriez voir un docteur, le frère. (Prompt Translator)

Je pense que vous devriez voir un docteur, frère. (Yahoo ! Babel Fish)

Je pense que vous devriez voir un médecin, frère. (Bing Translator)

Cet exemple pose un problème culturel, un problème de contexte: *Mon frere.* Cet qui appelle automatiquement le *tu* et non le *vous.*

Nous donnons ci-dessous quelques exemples de traduction du français à l'espagnol.

Le logiciel Apertium donne les résultats suivants :

Enoncé 4:

When he finally decided to quit, the company had lost $100,000.

Cuándo él finalmente decidido a quit, la empresa había perdido $100,000.

Nous pouvons relever l'utilisation du pronom *el* qui pourrait creer un glissement semantique (insistance sur le sujet grammatical, ce que ne fait l'enonciateur de l'enonce source). Il y a egalement le *decided* qui est mal rendu. Le verb *quit* n'a pas trouvee son equivalent, le traducteur le reproduit donc. Et enfin le graphie du chiffre.

Enoncé 5:

Le téléphone sonnait mais elle refusait de décrocher.

El teléfono sonaba pero negaba a décrocher.

Décrocher est repris ici, alors qu'il y a *coger*…

Enoncé 6:

Toutefois, le fait que les analyses soient focalisées sur des données ou corpus tirés de la chaine linéaire des énoncés pose le problème de l'adéquation des règles.

Sin embargo, el hecho que los análisis sean focalizadas sobre datos o corpus tirados de de la chaine lineal de los énoncés plantea el problema de la adecuación de las reglas.

Toutefois, le fait que les analyses soient focalisées sur des données ou corpus tirés de la chaîne linéaire des énoncés pose le problème de l'adéquation des règles.

Sin embargo, el hecho que los análisis sean focalizadas sobre datos o corpus tirados de del canal lineal de los énoncés plantea el problema de la adecuación de las reglas.

Ces traductions proposées par Apertium pose plusieurs problèmes: un problème de détermination nominale (de del) et un problème de choix lexical (chaine et canal). La difference lexicale a été le résultat de la représentation orthographique: l'accent circonflexe sur le *i* de *chaîne*.

A titre de comparaison, ce dernier paragraphe est traité ci-dessous par Systran Net, Google Translate, Yahoo! Babel Fish, Prompt, Babylon et Bing Translator comme

suit:

No obstante, el hecho de que los análisis estén concentrados sobre datos o corpus extraídos de la cadena lineal de las declaraciones plantea el problema de la adecuación de las normas. (Systran)

Sin embargo, el hecho de que las pruebas se centran en los datos o corpus de la cadena lineal de las declaraciones se plantea la cuestión de la adecuación de las normas. (Google Translate)

No obstante, el hecho de que los análisis estén concentrados sobre datos o corpus extraídos de la cadena lineal de las declaraciones plantea el problema de la adecuación de las normas. (Yahoo! Babel Fish)

No obstante, el hecho que los análisis sean enfocados sobre datos o cuerpos tirados de la cadena lineal de los enunciados plantea el problema de la adecuación de las reglas. (Prompt Translator)

Avec *Chaine* écrit sans accent circonflexe, Prompt propose *chaine.* Mais avec l'accent circonflexe, il donne *cadena.*

Sin embargo, el hecho de que los análisis se centraron en datos o corpus extraídas de la cadena lineal de los enunciados plantea el problema de l idoneidad de las normas. (Babylon Translator)

Sin embargo, el hecho de que los análisis se concentre en datos o corpus de la cadena de lineal de declaraciones es el problema de la adecuación de las normas. (Bing Translator[8])

De tous les traducteurs ci-dessus, Babylon est le seul à proposer une traduction satisfaisante pour le mot *énoncé.* Toutefois, la graphie de *l'* qui détermine *adéquation* pose problème.

[8] Bing Translator est le traducteur de Microsoft (originellement Window Live Translator).

Apres ce bref survol de quelques traducteurs électroniques, nous pouvons faire les commentaires suivants: tout d'abord, ils ne proposent pas tous les mêmes résultats. Ensuite, il y des problèmes de correspondances typographiques ($100,000 ci-dessus), des équivalences de représentation de déterminants et de choix lexicaux.

La performance des traducteurs électroniques dépend ainsi de plusieurs paramètres :

- la richesse et l'accessibilité des ressources lexicales,
- les terminologies et glossaires,
- les systèmes d'exploitation sous lesquels fonctionnent les ordinateurs utilisés,
- les navigateurs internet,
- la performance de l'ordinateur,
- les algorithmes,
- les architectures de traduction,
- les relations de structures sémantiques, morphosyntaxiques et génétiques qu'entretiennent langue de départ (LD) et langue d'arrivée (LA),
- la structure des rapports culturels entre les locuteurs de LD et LA,
- la dynamique des mises à jour des modules de traductions et des contenus linguistiques et pragmatiques,
- les méthodes et procédures de traitement des contenus de sens et significations des textes à traduire,
- etc.

Ces paramètres posent trois faits majeurs :

- la culture dynamique du logiciel : en effet, le logiciel doit être "hypercultivée" (d'où le développement des corpus et autres ressources lexicales générales et spécialisées). Cette culture doit être dynamique en

suivant l'évolution des créations morphosyntaxiques, avec, par exemples des liens vers les bases de données et corpus les plus importants du web.

- la réactivité du logiciel: le logiciel doit pouvoir répondre à tous les changements dans la structure du texte et proposer des réponses adéquates

- le traitement du sens et de la signification: ce point est l'un des plus importants de l'activité traduisante. Le traducteur traduit du sens et des significations[9]. Le logiciel doit être capable de calculer les sens – ce qui suppose la prise en compte non seulement des phénomènes de dénotation mais également les faits de connotation.

Pour un meilleur résultat de l'activité traduisante automatique, Yahoo! Babel Fish donnent des comportements à avoir :

> Trois conseils pour une bonne traduction[10] :
>
> 1. Veillez à respecter les règles de grammaire, orthographe et ponctuation pour une qualité de traduction optimale.
> 2. Après avoir traduit un texte, cliquez sur le bouton "Rechercher ce texte sur le Web" afin de lancer une recherche utilisant le résultat de la traduction comme mots-clés.
> 3. Comparez la page traduite à l'originale en cliquant sur "Voir la version originale".

Il est vrai que ces conseils ne règlent pas les difficultés que pose la traduction automatique. Toutefois, ils montrent qu'il y a des comportements spécifiques à avoir pour une traduction améliorée. Les techniques d'évaluation aideront, peut-être, à aller plus loin dans la bataille pour l'amélioration des résultats des traducteurs logiciels et des traducteurs en ligne.

7. L'évaluation de la traduction automatique

---

[9] Voir Peter Newmark, *A Textbook of Translation*, Hertfordshire, Prentice Hall, 1988

[10] http://fr.babelfish.yahoo.com/

C’est convaincus des difficultés inhérentes à l’activité traduisante que plusieurs centres et groupes de recherche ont mis en place des techniques d’évaluation pour non seulement améliorer la performance de logiciels de traduction automatique mais également de guider les consommateurs dans leur prise de décision d’investissement ou d’utilisation.

Il existe plusieurs méthodes d’évaluation dont :

- METEOR
- NIST
- BLEU
- …

Ces méthodes permettent d’attribuer des scores de qualité aux traductions proposées par les différents logiciels. La plus utilisée des ces méthodes est la métrique BLEU (*Bilingual Evaluation Understudy*), complétée par NIST, suivie de METEOR (*Metric for Evaluation of Translation with Explicit ORdering*). La qualité des traductions est estimée automatiquement par la métrique existante.

L’efficacité de la métrique BLEU repose sur la comparaison de la sortie du traducteur avec les traductions dites de référence. Cette métrique mesure le recouvrement lexical de la phrase traduite avec une ou plusieurs phrases données comme références de la traduction. Le score BLEU varie de 0 à 1 et, étant un score de précision, il est d’autant meilleur qu’il est grand.[11]

Un exemple de l’importance de l’évaluation en traduction automatique est fourni par le projet CESTA initié par ELDA / ELRA (*Evaluations and Language Resources Distribution Agency / European Language Resources Association*). En effet, ELDA a

[11] Voir Marion Potet, Laboratoire d’informatique de Grenoble, équipe GETALP, “Méta-moteur de traduction automatique : proposition d’une métrique pour le classement de traductions", RECITAL 2009, Senlis, 24–26 juin 2009

mené une *Campagne d'évaluation des systèmes de traduction automatique* (CESTA) (2003-2006). Cette campagne avait les objectifs suivants :

- Définir un protocole fiable pour l'évaluation de la TA
    - mesures de qualité nécessitant des juges humains
    - mesures de qualité automatiques
        - Évaluer des systèmes de TA
    - industriels et académiques
    - traduisant de l'anglais et de l'arabe vers le français
    - dans plusieurs domaines et conditions d'utilisation
        - Mettre à disposition de la communauté des ressources et des outils pour l'évaluation de la TA

Le principe est le suivant: mesurer la qualité d'un texte traduit en comparant celui-ci à une ou plusieurs traductions de référence.

- tester la fiabilité de plusieurs de ces métriques, pour les traductions vers le français
        - Mesures employées dans CESTA
    - **BLEU** : *Bilingual Evaluation Understudy* (Papineni et al. 2001)
        - moyenne pondérée du nombre de mots en commun, du nombre de bigrammes en commun, etc. (n-grammes avec n = 1, 2, 3, ou 4)
        - fiabilité inconnue pour des langues cible à morphologie riche
    - **NIST** (Doddington, 2002)
        - variante de BLEU: gain d'information et pénalités selon la taille

- **WNM** : *Weighted n-gram metric* (Babych & Hartley 2004)
  - pondère les comparaisons de n-grammes selon leur fréquence
  - autorise une certaine variation dans la traduction
- **X-Score** (Rajman & Hartley, 2001)
  - analyse la grammaticalité du texte traduit en comparant la distribution morpho-syntaxique du texte avec un corpus de référence
  - mesure expérimentale implémentée par l'ELDA pour CESTA
- **D-Score** (Rajman & Hartley, 2001)
  - analyse de la préservation du contenu sémantique en comparant la représentation sémantique vectorielle du texte traduit avec celle d'un texte de référence
  - mesure expérimentale implémentée par l'ELDA pour CESTA
- Distances d'édition de chaînes de caractères (Leusch et al., 2003)
  - **mWER**: *Multi-reference Word Error Rate*
  - **mPER**: *Multi-reference Position-independant Word Error Rate*

- l'évaluation humaine des systèmes (référence de la qualité)
- la méta-évaluation des métriques automatiques
  - en comparant leurs scores avec ceux des juges humains
- Développement par l'ELDA d'une interface pour l'évaluation humaine en ligne, via http
  - Scores d'adéquation (sémantique) et de fluidité
    - échelle de 1 à 5

- chaque segment est évalué par deux juges différents
- les segments sont présentés aléatoirement

Les conclusions de ce programme sont présentées ci-dessous :

> Les résultats de CESTA ont été discutés au cours d'un atelier final et publié dans le rapport ci-dessus.
>
> L'évaluation humaine a permis de mettre en évidence la proximité de S2 avec la référence humaine (afin de conserver l'anonymat, nous parlons des systèmes S1 à S6). Les systèmes S1, S3 et S4 sont quant à eux assez proches les uns des autres. Le système S3 arrive toutefois en seconde position suivi du système S1, puis S4. Le système S5 est très éloigné des scores de ce groupe. Les résultats du système S6 sont assez bons, même si la référence humaine est largement au-dessus. Les résultats surprenants de S2 (supérieur à la référence humaine pour l'adéquation) sont en cours d'étude, puisqu'il s'est avéré qu'ont été inclus certains fichiers du corpus de test de Santé Canada pour l'entraînement du système.
>
> En ce qui concerne les métriques automatiques, les mesures statistiques (BLEU, NIST, WNM) obtiennent de relativement bons résultats, puisqu'ils sont en deçà de ceux de la première campagne. Toutefois, les corrélations sont d'un niveau acceptable, meilleures pour l'adéquation que pour la fluidité. On retrouve avec ces métriques la structure de l'évaluation humaine, avec le système S2 largement au-dessus, un groupe de trois systèmes assez proches (S1, S3, S4), et un dernier système plus loin (S5). Le classement est légèrement différent. Les résultats obtenus pour la métrique BLEU et les deux scores humains sont résumés ci-dessous :

Systèmes	BLEU	Fluidité	Adéquation
System 1-EN	37.8	54.7	59.6
System 2-EN	89.6	82.1	88.2
System 3-EN	38.4	57.5	60.9
System 4-EN	39.8	54.3	55.7
System 5-EN	33.9	32.0	46.0
System 1-AR	42.3	51.9	42.6

8. Le futur de la traduction automatique

Traduire est un processus dynamique qui consiste en *« verter a otra lengua (lengua terminal - LT) el significado de un texto [escrito en una otra lengua (lengua original - LO)] en el sentido pretendido por el autor »*[12]

[12] Peter Newmark, *Manual de traducción,* Madrid, Cátedra, 2004, p.19

La traduction est donc une activité : l'activité traduisante. Elle est cognitive, sociocognitive, linguistique, stylistique, etc.

Comme l'écrivait Hutchins, « le système de traduction complètement automatique qui produira les textes idiomatiques comparables aux traductions humaines n'est qu'un rêve pour ceux qui n'en ont pas l'expérience. Toutefois, dans les conditions propices, ces systèmes, loin d'être parfaits, peuvent être utilisés avec profit et succès.»[13] L'activité traduisante est, donc, a ce jour sous l'*entière* domination de l'Homme. C'est une activité véritablement (socio-)cognitive que la machine aura beaucoup de mal à intégrer totalement. Le moteur *WorldLingo* dans l'avertissement ci-dessous ne nous contredit pas:

> "En raison des subtilités du langage humain et de l'existence possible de nombreuses traductions et interprétations différentes de certains mots et phrases, les traductions automatiques présentent des limites. WorldLingo décline toute responsabilité concernant l'exactitude de la traduction et les plaintes relatives à votre utilisation des services de traduction WorldLingo."[14]

Conclusion

Le défi de la traduction automatique réside en une question: comment construire une architecture informatique capable de *comprendre toutes les nuances de significations* des énoncés produits dans une langue source et proposer une version traduite satisfaisante, comme le ferait un traducteur professionnel efficace ? Nous avons vu que beaucoup reste à faire même s'il y a, à ce jour, des résultats très encourageants. L'activité langagière est dynamique. Les effets de sens sont nombreux et déroutants. La situation d'énonciation est l'un des éléments clés du déchiffrage sémantico-pragmatique des textes à traduire. Cette raison explique les informations préalables recherchées par le traducteur ou interprète professionnel avant d'aborder le texte d'un domaine particulier. Le traducteur professionnel opère un exercice cognitif important de croisement entre les paramètres de l'énonciateur ou « auteur » du texte, des paramètres spatiotemporels, ainsi que de toutes les données pragmatiques en

---

13 John Hutchins, 1993

14 www.worldlingo.com/fr/products_services/computer_translation.html

présence. Un système informatique n'aurait-il pas du mal à suivre ce parcours quelque peu alambiqué? Peut-être qu'à force dynamisme et de fécondité dans la recherche…

Bibliographie

Adamczewski, Henri, *Grammaire linguistique de l'anglais,* Paris, Armand Colin, 1982.

Hutchins, John, "Vers une nouvelle époque en traduction automatique", *Troisièmes Journées Scientifiques LTT, Montréal, 30 septembre 1993*

Lavecchia, Caroline, Kamel Smaili et David Langlois, « Une alternative aux modèles de traduction statistique d'IBM: Les triggers inter-langues », *LORIA/Speech Group*, Campus scientifique, Juin 2008.

Newmark, Peter, *A Textbook of Translation*, Hertfordshire, Prentice Hall, 1988

Newmark, Peter, *Manual de traducción,* Madrid, Cátedra, 2004

Potet, Marion, "Méta-moteur de traduction automatique : proposition d'une métrique pour le classement de traductions", RECITAL 2009, Senlis, 24–26 juin 2009

WorldLingo, www.worldlingo.com/fr/products_services/computer_translation.html, 10 juin 2009.