

LALIES

39
2019

LANGUE & LITTÉRATURE

lalies

ACTES DES SESSIONS DE LINGUISTIQUE
ET DE LITTÉRATURE

39

LALIES

Actes des sessions de linguistique et de littérature ENS-Clelia

Les sessions de linguistique et de littérature ENS-Clelia, organisées conjointement par l'École normale supérieure et l'association Clelia, offrent chaque année à un public d'enseignants et de chercheurs, principalement en langues anciennes, un programme d'information et de formation continue dans les domaines de la linguistique générale, de la grammaire comparée et de la théorie littéraire.

Pour tous renseignements sur ces rencontres, écrire à :
Clelia, B.P. 192, F-75 226 Paris cedex 05.

*

Depuis 1979, les actes de ces sessions sont publiés dans *Lalies*.

Comité de rédaction du numéro 39

Bérenger BOULAY, Alain CHRISTOL, Pedro DUARTE, Frédérique FLECK,
Victor GYSEMBERGH, Dimitri KASPRZYK, Peggy LECAUDÉ,
Claire LE FEUVRE, Audrey MATHYS, Guillemette MÉROT, Hamidou RICHER.

Voir en fin de volume les sommaires des numéros disponibles.

lalies

ACTES DES SESSIONS DE LINGUISTIQUE
ET DE LITTÉRATURE

39

ÉVIAN-LES-BAINS, 20-24 août 2018

ÉDITIONS NSRUED'ULM

Cet ouvrage a été publié avec le soutien
du département des Sciences de l'Antiquité de l'École normale supérieure de Paris.

En couverture :
caractères en alphabet shavian.

Tous droits de traduction, d'adaptation et de reproduction par tous procédés
réservés pour tous pays.

© Éditions Rue d'Ulm / Presses de l'École normale supérieure, 2019
45, rue d'Ulm, 75230 Paris cedex 05
www.pressens.fr
ISSN 0750-9170
ISBN 978-2-7288-0646-1

PRÉSENTATION

Ce trente-neuvième numéro de la revue *Lalies* réunit les textes issus de la session de linguistique et de littérature de l'association Clélia qui s'est tenue à Évian-les-Bains du 20 au 24 août 2018, grâce à l'aide financière du Labex TransfèrS. Cette session s'est signalée par la qualité des conférences et la richesse des contributions proposées à titre de *varia*, très nombreuses dans ce volume. Je tiens à exprimer toute ma gratitude au département des Sciences de l'Antiquité et à sa directrice, Christine Mauduit, pour le soutien financier qui a permis à *Lalies* 39 de paraître. Mes plus vifs remerciements vont aussi à Laurence Debertrand, des éditions Rue d'Ulm, dont le travail précis et efficace de mise en forme du volume a permis sa parution rapide, à Lucie Marignac, directrice des éditions Rue d'Ulm, pour son soutien indéfectible à la publication de *Lalies*, ainsi qu'à Marie-Hélène Ravenel qui a également contribué à la genèse de ce numéro. Les membres du comité de rédaction de *Lalies* 39 m'ont été d'une aide précieuse par l'acribie et la ponctualité de leurs relectures : qu'ils en soient ici chaleureusement remerciés. Je ne saurais enfin exprimer suffisamment ma reconnaissance à Daniel Petit qui m'a guidée et soutenue tout au long de ce processus d'édition. Je suis extrêmement touchée de la confiance qu'il m'a témoignée en me proposant de prendre en charge la publication de cette revue qu'il a éditée avec le plus grand soin depuis le numéro 20, paru en 2000, et je joins mes remerciements à ceux de tous les membres de l'association Clélia pour le dévouement avec lequel, année après année, il s'est attaché à faire paraître dans les meilleures conditions les actes de nos sessions.

Ce volume 39 s'ouvre sur un article de Thierry Poibeau qui offre une présentation très complète et très instructive du traitement automatique des langues. Après un historique du développement de ce domaine de recherche, et plus particulièrement de ce qui touche à la traduction automatique, il en explique les principaux enjeux en insistant sur les avancées déjà réalisées et sur les difficultés actuellement rencontrées, avant de proposer un tour d'horizon de ses diverses applications. Sa présentation insiste sur les rapports entre traitement automatique des langues et linguistique, non seulement pour ce qui est des questions de syntaxe, de morphosyntaxe ou de sémantique, mais encore pour l'étude des langues de terrain, l'acquisition du langage ou la typologie linguistique.

La présentation du BCMS, le bosniaque-croate-monténégrin-serbe, proposée par Marijana Petrović est centrée sur les problèmes liés au contexte sociopolitique dans lequel cette langue est actuellement employée. À la suite de l'éclatement de la Yougoslavie, les quatre pays qui en sont issus et qui partagent une même langue ont fait de celle-ci un enjeu d'identification nationale en tentant d'établir quatre standards linguistiques différents.

Casper de Jonge propose dans sa contribution une typologie des interactions et connexions qui existent entre critique littéraire grecque et poésie latine, un domaine de recherches longtemps négligé. Il étudie ensuite les différents liens qui existent entre le traité sur *La Composition stylistique* de Denys d'Halicarnasse et l'*Art poétique* d'Horace. L'auteur s'intéresse pour finir aux rapports entre la conception du sublime exposée dans le traité *Sur le sublime* de Longin et le passage du livre VI de l'*Énéide* sur le silence de Didon lors de sa rencontre avec Énée aux Enfers, en recourant de surcroît à la théorie moderne

des passages à forte intensité émotionnelle qui présentent bien des points communs avec les moments sublimes de Longin.

La section des *varia* est particulièrement fournie cette année. Dans le premier article, Cécile Tep se livre à une analyse générique de l'ode III, 28 d'Horace appuyée sur des considérations métriques et propose d'y voir un glissement du genre symposiaque au genre de la nénie, qui passe par une inclusion du genre de l'hymne. Marie de Toledo théorise, quant à elle, les enjeux d'un procédé critique, l'autométatextualisation de la difficulté, qui consiste à considérer que le texte commenté s'autodésigne comme difficile, avant d'examiner sa mise en œuvre dans les commentaires de l'*Alexandra* de Lycophron. Le texte d'Alain Christol présente une typologie des ethnonymes et passe en revue différents choix de désignation de peuples, qu'il s'agisse de noms anciens ou de noms plus modernes. Le volume se clôt sur un texte à visée pédagogique qui promeut l'introduction de la critique textuelle dans les cours de langues anciennes. Victor Gysembergh y donne tous les éléments pour se lancer : principes généraux, bibliographie commentée et exemples concrets d'utilisation dans le cadre d'un cours.

Frédérique Fleck

Lalies

École normale supérieure

45 rue d'Ulm

F-75230 Paris CEDEX 05

frederique.fleck@ens.fr

LE TRAITEMENT AUTOMATIQUE DES LANGUES : TENDANCES ET ENJEUX

Thierry POIBEAU¹

RÉSUMÉ

Le traitement automatique des langues (TAL) est un domaine de recherche pluridisciplinaire, à l'intersection de la linguistique et de l'informatique, qui vise à mettre au point des programmes informatiques pour l'analyse des langues par ordinateur. Ce domaine de recherche est apparu au début des années 1950, en même temps que la mise au point des premiers ordinateurs. Il s'est depuis largement diversifié et les applications de traitement automatique des langues se sont imposées dans la vie courante, que ce soit à travers les correcteurs orthographiques, les outils de traduction automatique ou la commande à la voix des téléphones portables ou des enceintes connectées. Au-delà de l'aspect parfois gadget des applications courantes, cet article essaie d'expliquer les difficultés de la tâche et les grandes évolutions du domaine ; il vise, en outre, à dresser un panorama des recherches en cours et des outils disponibles. Cet article est aussi l'occasion d'une réflexion sur la place de la linguistique dans le paysage actuel de la recherche en traitement automatique des langues.

ABSTRACT

Natural language processing (NLP) is a domain of research that aims at producing computer models for the automatic analysis of languages. The domain emerged in the early 1950s, along with the development of the first computers. It has since then evolved a lot and has given birth to popular applications of everyday life: spell checkers, machine translation tools or voice control for mobile phones and connected speakers. Beyond the gadget-like aspect of some common applications, this article tries to explain the difficulties of NLP and the major developments in the field ; it also aims at providing an overview of current research and tools. Finally, in the course of this paper, we will examine to what extent linguistics still play a role in the field.

1. INTRODUCTION : QUELQUES REMARQUES SUR LA SITUATION ACTUELLE DU TRAITEMENT AUTOMATIQUE DES LANGUES ET SUR LA PLACE DE LA LINGUISTIQUE EN SON SEIN

Le traitement automatique des langues (TAL) est un domaine de recherche pluridisciplinaire à l'intersection de la linguistique et de l'informatique². En guise de définition liminaire, on peut dire que le TAL concerne essentiellement l'analyse des langues au moyen d'un ordinateur. Le TAL est aussi connu sous diverses appellations qui traduisent parfois des nuances au sein de ce vaste domaine de recherche : on parle par exemple d'« ingénierie linguistique », quand l'accent est mis sur les aspects pratiques et opérationnels, ou de « linguistique informatique », quand c'est la linguistique qui tient un rôle important dans les recherches. Notre définition est aussi trop restrictive dans la mesure où le TAL

-
1. Cet article a bénéficié de la relecture attentive et bienveillante d'Audrey Mathys, de Pedro Duarte et de Frédérique Fleck. Leurs remarques ont permis d'en améliorer grandement le contenu : qu'ils en soient remerciés ! Toutes les éventuelles erreurs et imprécisions demeurant sont évidemment de mon fait.
 2. Sabah (1988 et 1989), Poibeau (2003 et 2011) et Léon (2015).

recouvre également la « génération de texte », c'est-à-dire la production (et non l'analyse) automatique de textes, thème de recherche populaire aujourd'hui, à l'heure des agents conversationnels et autres gadgets fondés sur une interaction avec l'utilisateur (pour qu'il y ait interaction, il faut prévoir à la fois une étape d'analyse et une étape de génération).

Il faut dès à présent souligner que ce domaine a subi des mutations extrêmement importantes en quelques années, et que ce qui se fait aujourd'hui sur le plan technique n'a plus grand chose à voir avec ce qui se faisait il y a encore quelque temps de cela³. Du point de vue du grand public, le domaine a longtemps été connu essentiellement à travers les correcteurs orthographiques et les traducteurs automatiques, qui étaient de surcroît souvent de qualité médiocre. À l'inverse, il existe aujourd'hui de plus en plus d'applications visibles, opérationnelles et relativement efficaces. Tous les problèmes ne sont pas résolus et les systèmes automatisés font toujours des erreurs, mais le domaine a connu ces dernières années des réussites indéniables. La qualité des traductions, au moins du français vers l'anglais, devient de plus en plus satisfaisante, on peut maintenant converser avec un agent artificiel de manière relativement efficace (on pensera à Siri d'Apple, à OK Google ou à Alexa d'Amazon), les moteurs de recherche sont de plus en plus précis et corrigent d'eux-mêmes certaines fautes de frappe, suggèrent des alternatives, etc. Pour la première fois depuis quelques années, des applications se répandent et sont utilisées par le grand public.

La situation présente s'explique par des progrès récents extrêmement importants, qui n'ont pas grand-chose à voir avec la linguistique, mais qui sont avant tout techniques. Deux éléments principaux expliquent pour l'essentiel la situation présente (et plus généralement les progrès relativement réguliers depuis 25 ans) : d'une part la masse de données textuelles disponible sur Internet ; d'autre part la puissance de calcul des machines, en constante augmentation. Le développement de ce que l'on appelle l'« Intelligence artificielle » et, au sein de ce domaine, le développement de l'« apprentissage artificiel », a considérablement renouvelé le TAL⁴ : contrairement à ce que l'on pensait il y a encore quelques décennies (et contrairement à ce que pensent encore aujourd'hui certains collègues, probablement), il est extrêmement efficace d'« apprendre » depuis les données. Nous reviendrons sur ce terme : il s'agit d'un abus de langage, l'ordinateur n'apprend rien à proprement parler, mais on a aujourd'hui accès à tant de données qu'il est possible de repérer des régularités et plus globalement des « faits de langue » en quantité suffisante pour obtenir des systèmes efficaces et capables d'analyser et même de produire des phrases variées et complexes, dans une multitude de langues.

On a longtemps pensé que le développement d'un système de traduction automatique, par exemple, nécessitait de formaliser sur ordinateur toute la langue et même les connaissances de sens commun, car celles-ci sont nécessaires pour déterminer le sens en contexte⁵. Pour prendre un exemple simple et un peu artificiel, si on me dit : « L'avocat est entré dans la pièce », je comprends immédiatement que « l'avocat » désigne un humain exerçant un métier lié à la justice, qu'il est probablement le défenseur d'une personne accusée de quelque chose, etc. Je ne vais pas considérer l'option « avocat qui se mange » (avocat fruit) car cette acception serait très étonnante ici : un avocat (fruit) se mange, mais n'entre pas dans une pièce car il ne peut pas se mouvoir. Cela est évident pour tout le monde, c'est même enfoncer une porte ouverte que de le dire. Mais si je développe un système automatique, je vais avoir affaire à un mot (« avocat »), avec au moins deux significations (et peut-être plus, il peut aussi être l'avocat des parties civiles, ce qui est différent du sens le plus commun du mot « avocat » comme « homme de loi, qui défend un accusé » etc.), et il faudra que je conçoive

3. Adda-Decker *et al.* (2014).

4. Gilleron (2019).

5. Poibeau (2017).

des règles capables de prédire le sens en contexte. Or, il s'agit d'une connaissance difficile à modéliser. L'ordinateur, lui, va analyser les contextes d'apparition du mot à partir de très grands corpus, en tirer des régularités (simplement via des cooccurrences, c'est-à-dire des associations de mots apparaissant ensemble) qui permettent une désambiguïsation (c'est-à-dire le choix d'un sens parmi un ensemble de sens) assez efficace, en tout cas plus efficace que ce que ferait n'importe quel humain après un très long travail d'analyse manuelle.

Cela est particulièrement vrai pour une tâche comme la désambiguïsation sémantique, c'est-à-dire la tâche consistant à trouver le sens d'un mot en contexte parmi un ensemble de sens possibles listés dans un dictionnaire⁶. Il existe 50 000 mots dans un dictionnaire courant (et bien davantage si on prend en compte les mots composés). Il est dès lors impossible pour un humain d'énumérer tous les contextes possibles et de concevoir des règles de désambiguïsation efficaces en pratique : personne ne peut imaginer tous les emplois d'un mot, et même en se fondant sur des dictionnaires et des encyclopédies, la tâche reste quasi impossible car il faut aussi considérer les textes réels, pour voir l'usage réel des mots en contexte. L'ordinateur est au contraire infiniment efficace pour cela : il va enregistrer sans problème, en un instant, tous les contextes possibles pour tous les mots du dictionnaire et, si un programme efficace a été conçu, il va même pouvoir analyser le sens des mots automatiquement en examinant la diversité et la dispersion des contextes observés (plus un mot apparaît dans des contextes variés, plus il a des chances d'être polysémique).

Évidemment, comme on l'a dit, même avec les progrès récents, les systèmes ne sont pas parfaits. Il existe toujours des contextes difficiles ne permettant pas une désambiguïsation aisée des mots : il faut que le système dispose d'assez de données pour inférer des règles efficaces ; il faut que le domaine visé soit proche du domaine qui a servi à mettre au point le système initial, sinon le processus ne va pas fonctionner, etc. Ainsi, la traduction automatique fonctionne relativement correctement pour les langues indo-européennes bien représentées sur Internet. Les systèmes sont nettement moins bons quand il s'agit de traduire vers le chinois, l'arabe ou le japonais : bien que l'on dispose de données et que de nombreuses recherches aient été faites, la structure et la complexité des langues jouent évidemment un rôle, et il est plus difficile de traduire vers des langues distantes que vers des langues proches. On voit cela aussi pour les langues indo-européennes. Il est par exemple plus facile de traduire de l'allemand vers l'anglais que de l'anglais vers l'allemand car on sait relativement bien analyser les mots composés de l'allemand, mais on sait beaucoup moins bien comment produire des mots composés bien formés en allemand.

Il n'empêche : on a aujourd'hui affaire à des masses de données si grandes, au moins pour quelques langues « dominantes » et bien représentées sur Internet, que les approches automatiques sont devenues très efficaces et elles sont même utilisées dans différents cadres opérationnels. La recherche est devenue depuis 25 ans de plus en plus technique, informatique voire mathématique, laissant au second plan (voire au dernier plan) la linguistique : pour le dire très directement, la linguistique a quasiment disparu des conférences de TAL, au moins la linguistique telle qu'on la connaît traditionnellement. Ce que l'on appelait « grammaire formelle » par exemple avait non seulement sa place dans le domaine, mais elle occupait même une place centrale. On parlait alors de l'idée que les données brutes (c'est-à-dire les corpus) étaient trop disparates pour pouvoir être analysées directement. Une étape de normalisation et de formalisation était nécessaire au préalable pour parvenir à une analyse réellement opérationnelle. Les approches récentes, à l'inverse, se fondent directement sur l'analyse de corpus massifs, par des techniques brutes (on voit depuis une dizaine d'années un quasi-monopole des réseaux de neurones,

6. Navigli (2009).

à travers ce que l'on appelle «l'apprentissage profond» ou le «*deep learning*») sans préanalyse du corpus.

Ces aspects opérationnels connaissent des succès indéniables mais ne doivent pas cacher certaines faiblesses. On sait que la distribution du lexique est très hétérogène (c'est ce que montre la loi de Zipf) : quelques mots sont très fréquents (les articles, les prépositions, etc.), mais beaucoup d'autres mots sont très peu fréquents en corpus (même avec un très gros corpus, on trouvera beaucoup d'hapax et plus généralement de mots apparaissant seulement quelques fois) et, entre les deux, une zone réduite de mots relativement fréquents (les noms et les verbes les plus courants par exemple). La langue suit généralement ce modèle⁷ : quelques règles de grammaire s'appliquent quasiment à toutes les phrases (la présence d'un sujet dans la phrase, par exemple, et le fait que celui-ci précède généralement le verbe, etc.) tandis que beaucoup de règles ne s'appliquent que dans de très rares cas. Les mots ont de très nombreux sens rares (d'un simple point de vue statistique) que l'on peut quasiment négliger si on souhaite développer un système opérationnel généraliste (le système fonctionnera bien dans la plupart des cas), etc. C'est ce qui explique à la fois le succès des systèmes mis au point sur le plan opérationnel, et en partie la quasi-disparition des linguistes de ce domaine.

Ainsi, dans les années 1980, la Commission européenne avait lancé un grand programme de recherche transnational autour de la traduction automatique. Le but était de produire des systèmes opérationnels, ou du moins «à l'état de l'art» (c'est-à-dire au meilleur niveau des performances d'alors), entre les principales langues européennes. Quelques années plus tard, de nombreuses thèses avaient été produites sur la négation (ou, plus exactement, sur la portée de la négation) dans différentes langues européennes et sur la façon d'en rendre compte d'une langue à l'autre. Il s'agissait probablement d'un problème linguistique intéressant, mais cela n'avait quasiment aucun intérêt pratique pour la traduction automatique (si l'on cherche à traduire des articles de journaux, la portée de la négation est rarement un problème). Il s'agit là d'une anecdote, mais elle révèle bien, à notre avis, l'écart entre les préoccupations des linguistes et celles des informaticiens.

Cet écart s'est sans doute réduit depuis les années 1980 (Internet est «passé par là», de même que le recours aux corpus massifs et aux outils automatiques, même pour les linguistes éloignés du TAL : on peut bien évidemment continuer à faire de la linguistique sans avoir systématiquement recours à l'ordinateur (et c'est heureux !), mais, d'un autre côté, le TAL est devenu tellement technique qu'il n'est pas facile pour un linguiste d'y trouver sa place. Depuis 25 ans les experts prédisent que les performances vont atteindre un «plateau» et que les systèmes, pour progresser, vont devoir intégrer «davantage de linguistique», mais cette promesse a toujours été repoussée jusqu'ici. L'avènement de l'apprentissage profond a encore repoussé cette échéance, les systèmes ayant connu un «bond de performance» par le simple usage de cette technique. Il est toutefois vrai que les chercheurs essaient de plus en plus d'intégrer des «connaissances linguistiques» dans les systèmes, c'est-à-dire en général des ressources mises au point par des linguistes.

Le but de ce texte n'est pas de faire le simple constat d'un divorce entre les deux domaines : d'un côté le TAL, et de l'autre la linguistique. Il s'agit au contraire d'appréhender les avancées du domaine, d'essayer de comprendre la situation actuelle et le type de techniques utilisées en TAL, et enfin de voir comment le TAL peut aujourd'hui servir de «marchepied» à des recherches en linguistique, ou plus généralement en littérature ou en sciences humaines et sociales.

L'exposé commence par un examen rapide des principales difficultés du traitement automatique des langues, afin de mettre en perspective les succès et les échecs des soixante-dix

7. Lebart et Salem (1994).

dernières années (section 2). La section suivante dresse un rapide survol historique du domaine, en se focalisant plus particulièrement sur la traduction automatique, l'application emblématique du TAL (section 3). L'exposé se poursuit avec un panorama des recherches actuelles, de la morphosyntaxe à la sémantique (section 4) et avec une présentation des techniques d'évaluation (section 5). On examine ensuite les applications du TAL à la linguistique (section 6), puis les applications à destination du grand public (section 7) avant de finir par les applications en lettres et sciences sociales, et plus globalement par les liens entre TAL et «humanités numériques» (section 8). L'exposé se termine enfin par une conclusion et une bibliographie qui permettra au lecteur d'aller plus loin sur certains aspects du problème.

2. POURQUOI L'ANALYSE DE LA LANGUE PAR ORDINATEUR EST-ELLE DIFFICILE ? LA QUESTION DE L'AMBIGUÏTÉ

Le TAL est difficile parce que l'ordinateur n'a *a priori* aucune connaissance de la langue. Il faut donc lui indiquer ce qu'est un mot, une phrase, etc. Jusque-là, les choses peuvent sembler relativement simples. En fait, il faut bien voir que dès ce niveau, la langue est complexe et ambiguë. Prenons deux exemples. L'apostrophe marque l'élision d'une lettre entre deux mots, comme dans «l'éléphant» (la lettre finale *e* de l'article défini est élidée devant un mot à initiale vocalique) et donc l'apostrophe peut être considérée comme un séparateur de mots. Cela est vrai en général, mais une séquence comme «aujourd'hui» est généralement considérée comme formant un seul mot, qui possède pourtant une apostrophe (qui ne joue plus alors son rôle de séparateur). Le tiret pose aussi des problèmes redoutables, et peut être soit un séparateur («Rendez-vous, vous êtes cernés!»), soit une partie du mot («J'étais en retard à mon rendez-vous»). Dans le premier exemple, «rendez» est un verbe à l'impératif, tandis que dans le second exemple, «rendez-vous» est un nom dont les parties ne doivent pas être séparées lors de l'analyse (même si, étymologiquement, on peut décomposer le mot, mais ce type d'analyse n'aurait pas de sens ici, dans le cadre d'un système de traitement automatique des langues qui n'a pas à tenir compte de l'étymologie). Le problème est en fait beaucoup plus prégnant : quasiment chaque mot, chaque expression et chaque phrase peuvent être ambigus⁸.

Prenons un exemple : «l'avocat a livré une plaidoirie au vitriol». Chaque mot introduit de nombreuses difficultés pour un ordinateur. Pour un humain, il est par exemple évident que «avocat» désigne un juriste, «a livré» correspond au verbe et que «au vitriol» est une expression figée. Il n'en va pas de même pour un ordinateur : «avocat» peut désigner un fruit ; «livré» peut facilement être identifié comme un verbe mais le sens est ici largement métaphorique : il n'y a pas de livraison à proprement parler dans la phrase. Les compléments prépositionnels posent eux aussi des problèmes redoutables : comment savoir que «au vitriol» est rattaché à «plaidoirie» et non au verbe «livré»? Si on avait eu affaire à la phrase «l'avocat a livré une plaidoirie au palais de justice», le complément «au palais de justice» aurait dû être rattaché au verbe et non au nom «plaidoirie», alors que la structure des deux phrases semble tout à fait comparable de prime abord.

On pourrait objecter à cela que «au vitriol» est une expression figée qui doit être enregistrée comme un tout (c'est-à-dire comme une entrée à part entière) dans le dictionnaire. Cela est probablement vrai mais ne fait ainsi que repousser le problème dans la mesure où cette stratégie revient à augmenter le nombre de mots et d'expressions, ce qui a pour conséquence d'introduire de nouvelles ambiguïtés, et finalement de rendre le problème sans fin.

8. Fuchs (1996).

Un dictionnaire du français courant contient en général entre 50 000 et 100 000 mots (hors noms propres). Quand on considère toutes les formes que l'on trouve effectivement dans les textes (un verbe comme « livrer » correspond en fait à plusieurs dizaines de formes conjuguées : « livrions », « livraient », « livrera », etc.), il est admis qu'il faut multiplier ce chiffre par huit environ en français. À cela, il faut ajouter les noms propres (on trouve des dictionnaires de plusieurs millions de noms propres, la plupart étant homonymes de noms communs, comme « Pierre » qui peut être confondu avec une « pierre », même si l'usage des majuscules limite le problème en français) et les dictionnaires de mots composés (qui peuvent aussi inclure plusieurs dizaines de milliers d'items). Enfin, chaque domaine technique est lui-même susceptible d'inclure de nombreux termes spécifiques, souvent homonymes d'autres mots de la langue.

Les problèmes d'analyse syntaxique s'ajoutent à cela (dans notre exemple ci-dessus, faut-il rattacher « au vitriol » à « plaidoirie » ou au verbe « a livré »?) et l'on voit qu'on a très rapidement affaire à un problème d'explosion combinatoire. La plupart des problèmes peuvent être résolus facilement, de manière locale (par exemple avec des heuristiques du type « comme “plaidoirie” apparaît dans le contexte du mot “avocat”, ce dernier désigne probablement l'homme de loi et non le fruit ») mais d'autres problèmes nécessitent des connaissances plus complexes, difficiles à concevoir de manière exhaustive quand on a affaire à des millions d'items.

Il peut sembler paradoxal que tout cela ne pose aucun problème de compréhension à un humain, qui ne voit même pas qu'il y a ambiguïté (au sens où il faut choisir la bonne étiquette, le bon sens, le bon rattachement de chaque mot pour comprendre la phrase). De fait, cette dimension de la compréhension humaine a longtemps échappé aux concepteurs de systèmes automatiques, tant la compréhension est un phénomène naturel, direct et inconscient pour un humain. Il est d'ailleurs très improbable que le cerveau analyse toutes les possibilités pour chaque mot afin d'obtenir une représentation sémantique pour une phrase donnée : grâce au contexte, le cerveau accède probablement directement à la bonne interprétation, sans même considérer les analyses alternatives. À ce sujet, il a parfois été proposé un parallèle avec le cube de Necker, cette représentation des arêtes d'un cube en perspective cavalière (figure 1).

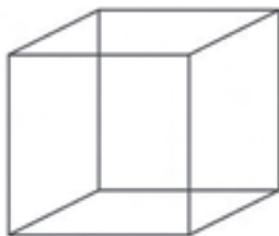


Figure 1. Le cube de Necker, une illusion d'optique publiée par Louis Albert Necker en 1832⁹.

Le dessin est ambigu dans la mesure où rien n'indique *a priori* quelle est la face du cube la plus proéminente. Le cerveau interprète cependant naturellement la figure afin d'en obtenir une représentation valide (on peut voir alternativement deux cubes différents mais ces perceptions ne sont jamais simultanées car le résultat ne serait pas conforme à une représentation valide, ou tout au moins aux conceptions prédéfinies dans le cerveau par observation de la réalité). Les dessins paradoxaux d'Escher fonctionnent aussi sur ce

9. Source de l'illustration : https://fr.wikipedia.org/wiki/Cube_de_Necker#/media/File:Necker_cube.svg

principe, sauf que ces dessins n'offrent pas une vue logique du monde mais à l'inverse contredisent nos conceptions prédéfinies de l'espace.

Nous mentionnons ces exemples pour illustrer le fait que le cerveau interprète et éventuellement modifie les perceptions (visuelles ou auditives) suivant des schémas préconstruits. Ou, plus exactement et suivant les principes de la Gestalt¹⁰, il y a une co-détermination simultanée des parties et de la forme globale : le contexte permet de déterminer le sens des mots qui, eux-mêmes, permettent de déterminer le sens de la phrase et ce de façon simultanée¹¹.

3. UN DOMAINE EMBLÉMATIQUE DU TAL : LA TRADUCTION AUTOMATIQUE

Dès les débuts du traitement automatique des langues, juste après la seconde guerre mondiale, deux domaines liés au traitement automatique des langues naturelles sont apparus comme étant d'une importance fondamentale : la recherche d'information et la traduction automatique¹². Le premier domaine visait à retrouver des informations parmi de grandes bases documentaires ou de grands ensembles d'archives, et à gérer des index complexes : le domaine donnera naissance aux bases de données, entre autres choses. L'autre secteur est celui de la traduction automatique : dans le contexte de la Guerre froide, traduire, du russe vers l'anglais en particulier, est très vite apparu comme un enjeu prioritaire¹³. Au-delà des aspects applicatifs immédiats, ce domaine de recherche a aussi été jugé important car il met en jeu toute la complexité du langage : la nécessité de comprendre le texte source, de le représenter formellement et de produire (on dit «générer») un texte comme résultat, dans la langue cible.

La période de l'après-guerre se nourrit en fait d'un ensemble de réflexions sur la langue, sur ses particularités et sur ce que devrait viser un programme d'intelligence artificielle, c'est-à-dire un programme qui imiterait l'intelligence humaine. À côté de la traduction, le dialogue humain-machine est aussi étudié en détail car si une machine savait dialoguer, cela veut dire qu'elle pourrait – un peu de la même façon que pour la traduction automatique – comprendre, analyser ce qui est dit et produire des énoncés en retour pour alimenter la conversation de façon pertinente. Turing en particulier imagine différents dispositifs, dont le célèbre test de Turing¹⁴ : le test est réussi si un humain, dialoguant par écran interposé, est incapable de dire si son interlocuteur est un humain ou une machine (ou, alternativement, si la personne, dialoguant avec deux interlocuteurs – un humain et une machine –, est incapable de discerner l'humain de la machine) ... La période de l'après-guerre fourmille de ces discussions sur ce qu'est l'intelligence, sur ce qu'il est possible de reproduire avec un ordinateur et sur la manière dont on peut le simuler, le tester, l'évaluer.

Assez rapidement on se rend compte du caractère ambivalent du dialogue humain-machine. Ce domaine restera toujours un domaine important du traitement automatique des langues, mais celui-ci ne nécessite pas obligatoirement une analyse poussée : il est en effet possible d'instaurer un dialogue assez réaliste sans compréhension du contenu

10. Guillaume (1979).

11. Victorri et Fuchs (1996). À l'inverse, la publicité joue fréquemment sur le maintien d'une ambiguïté volontaire de sens (par exemple dans un slogan comme «Il a Free, il a tout compris»). Beaucoup de personnes ne voient d'ailleurs pas le double sens spontanément, ce qui montre qu'en situation normale, le cerveau sélectionne naturellement un sens particulier et évite au maximum la double analyse.

12. Léon (2015).

13. Poibeau (2017).

14. Turing (1950).

textuel, aussi étrange que cela puisse paraître¹⁵. Ainsi, Weizenbaum développe dans les années 1960 Eliza, un système de dialogue qui repère des structures conversationnelles typiques et génère des questions réalistes simplement sur la base de ce qui a déjà été dit (par exemple, à partir de « Je n'ai pas aimé ce film », l'ordinateur reconnaît le patron « Je n'ai pas aimé X » avec « X = ce film » et peut proposer très simplement une question comme « Pourquoi n'avez-vous pas aimé ce film ? » pour relancer le dialogue). La traduction ne peut pas se contenter de techniques aussi pauvres, même si une traduction mot à mot peut servir à développer des systèmes opérationnels à moindre coût. Pour obtenir une traduction de bonne qualité, il est nécessaire d'avoir une compréhension en profondeur du texte à traduire – c'est du moins ce que pensaient les premiers acteurs du domaine.

3.1. Les années 1940 : les débuts du TAL

Comme on l'a signalé, la traduction automatique est vue comme une application clé dès l'immédiat après-guerre, avant même l'apparition des premiers ordinateurs. Ce désir d'automatisation et de formalisation du langage faisait écho à de très nombreuses recherches et réflexions menées dès les siècles précédents par nombre de savants et de penseurs, comme Leibniz ou Descartes.

L'après-guerre est marquée à la fois par un besoin – traduire du russe vers l'anglais – et par des avancées majeures qui ont eu lieu pendant la guerre dans le domaine de la cryptographie. La cryptographie vise à traduire dans un langage compréhensible (typiquement, une langue humaine), un message codé suivant une procédure complexe. C'est à cette époque que Claude Shannon et Warren Weaver, deux savants américains qui ont travaillé dans le domaine de la cryptographie pendant la guerre, puis dans ce qui deviendra l'informatique, proposent leur célèbre schéma de la communication (1949), tel que présenté dans la figure suivante (figure 2).

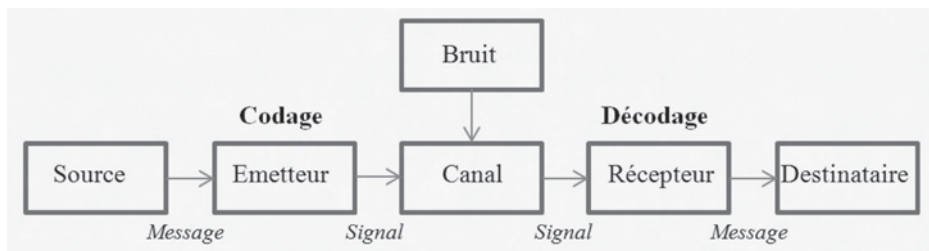


Figure 2. Le schéma de la communication d'après Shannon et Weaver¹⁶.

Ici, une source veut émettre un message. Celui-ci doit être encodé par l'émetteur : il peut s'agir par exemple de la traduction en morse d'un message en langue naturelle, ou tout simplement de la mise en mots d'idées que le locuteur souhaite émettre (si l'on imagine qu'il peut exister des idées abstraites dans la tête du locuteur, indépendamment de sa langue de communication) – le modèle était à l'origine essentiellement un modèle abstrait à des fins d'analyse du signal, mais il a depuis été largement appliqué à la communication humaine. Le message est ensuite transmis par un canal qui peut être « bruité », parce que le message circule dans un environnement soumis à divers aléas (bruit de fond, affaiblissement du signal en fonction de la distance, etc.). Ce message est enfin intercepté par un récepteur qui doit le décoder (transformer le message en morse ou en un message en langue naturelle, ou

15. Landragin (2013).

16. Source de l'illustration : Picard (1992).

interpréter des mots pour en inférer des idées). Le schéma de Shannon a donc une portée très large, il sera largement repris et reste encore largement étudié, en introduction aux formations traitant aussi bien de théorie de la communication que du traitement du signal.

À la même période, et avant que ce schéma ne s'impose, une idée similaire fait son chemin pour ce qui concerne la traduction automatique : ne pourrait-on pas considérer un texte en langue étrangère comme un message codé, qu'il faudrait analyser et traduire dans la langue cible, l'anglais par exemple ? Ce sont les échanges entre trois savants qui vont faire avancer ce domaine : Shannon et Weaver, dont nous avons déjà parlé, et Norbert Wiener, le père de la cybernétique, cette science qui vise, dans l'après-guerre, à donner une vision unifiée de l'automatique, de l'électronique et de la théorie de l'information¹⁷. Comme on l'a dit, la traduction automatique peut être vue comme un processus d'encodage / décodage. C'est ce que propose Weaver dès 1947, dans un échange avec Wiener :

On peut naturellement se demander si le problème de la traduction ne pourrait pas être considéré comme un problème de cryptographie. Quand je regarde un article en russe, je me dis : « Ceci a été écrit en anglais mais a été encodé avec des symboles étranges. Je vais maintenant procéder à son décodage. »

Wiener est de son côté extrêmement prudent, et a sans doute de bonnes intuitions concernant la difficulté du processus. Il a bien conscience que les mots ne se correspondent pas directement d'une langue à l'autre et que le problème de la traduction ne peut être résolu aussi simplement :

En ce qui concerne la traduction automatique, j'ai peur que les frontières des mots dans les différentes langues soient trop vagues [...] pour rendre l'idée d'un système quasi automatique de traduction possible.

On voit bien, à travers les réserves de Wiener, la mise en évidence de problèmes qui seront effectivement des obstacles majeurs pour la traduction automatique : les mots n'ont pas d'équivalents directs d'une langue à l'autre, il existe de nombreuses expressions figées ou idiomatiques qui ne peuvent pas se traduire littéralement, etc.

Weaver poursuivra malgré tout ses travaux et rédigera en 1949 un court texte intitulé « *Translation* », mais plus connu sous le simple nom de *Memorandum*, qu'il fera circuler auprès de la communauté scientifique de l'époque. Weaver a en fait « une double casquette » : à côté de ses travaux scientifiques, celui-ci est aussi responsable de programmes de recherche pour les organismes scientifiques américains et doit déterminer les domaines scientifiques prometteurs qu'il faut financer. C'est ce qui explique le succès important de ce petit texte relativement informel : c'est à partir de ce *Memorandum* que la recherche en traduction automatique va démarrer, se structurer et se développer aux États-Unis. Dans son texte, Weaver n'ignore pas complètement les remarques et les réserves de Wiener, mais celles-ci sont tout de même mises au second plan, alors qu'il s'agit de critiques fondamentales qui reviendront avec force 15 ans plus tard.

Dans son *Memorandum*, Weaver met en avant quatre idées à explorer pour le traitement automatique des langues, et en particulier pour la traduction automatique. Ces quatre idées sont les suivantes :

1. L'analyse du contexte d'apparition des mots doit permettre de déterminer leur sens. L'étendue du contexte à prendre en compte doit varier suivant la nature du mot mais aussi suivant le type et le sujet général du texte, si ces éléments sont connus.

17. Sur toute cette période, outre les références déjà citées, voir les livres de référence de Hutchins (1986), Hutchins et Somers (1992) et Hutchins (2000).