

Éléments de réflexion sur l'ambiguïté et le traitement automatique des langues (TAL)

BOUHADIBA Farouk
Université de Mostaganem

L'ambiguïté au sein de structures syntaxiques d'une même langue a toujours posé des problèmes au traducteur. Ces problèmes deviennent encore plus sérieux lorsqu'il est question de Traduction Automatique. En français 'Il m'a parlé de *sa* tendresse' en est un exemple. L'ambiguïté lexicale dans la Traduction Automatique des Langues (TAL, NLP en anglais) pose également de sérieux problèmes aux dictionnaires électroniques.

Nous essayerons de traiter quelques cas d'ambiguïté dans la traduction automatique de l'arabe vers le français et l'anglais. Nous aborderons question de la segmentation concernant la détermination / l'indétermination, la coordination, et celle du mot (unité de sens ou unité fonctionnelle). Quelques aspects de la segmentation des constituants en catégories grammaticales et leur comparaison avec des structures de l'arabe, du français, de l'anglais, ... seront abordés dans cette réflexion sur la langue arabe et le TAL.

La langue arabe appartient à la famille des langues dites sémitiques. Les historiens, tout comme – beaucoup plus récemment – les linguistes s'accordent à retracer ses origines à partir de langues très anciennes telles que l'Akkadien (3000 avant JC) de la vieille Mésopotamie. D'autres langues telles que le Phénicien ou l'Ethiopien (l'Amharique) présentent des caractéristiques linguistiques plus ou moins similaires à celles de l'arabe. Néanmoins, l'Hébreu demeure la langue avec laquelle la langue arabe partage le plus de similitudes

MUTARĠIM n° 11, Jan – Juin 2005

linguistiques, plus particulièrement sur les plans morpho-syntaxique et lexical.

Parmi les particularités (spécificités) de la langue arabe, nous retiendrons dans le cadre de cette présentation son graphisme (écriture de droite à gauche), ses sons (les gutturales et les emphatiques) et surtout sa morphologie (processus de formation d'un mot, affixation, reduplication et autres spécificités morpho-syntaxiques telles que les temps, les modes, voix, etc.).

Pour l'historique, il est à retenir que les premières traces d'écriture de cette langue remontent au IV^{ème} siècle de notre ère. Celles-ci ont été relevées dans la péninsule arabique, mais l'on situe l'utilisation de cette langue dans les échanges entre tribus dès le V^{ème} siècle. Beaucoup plus récemment, les linguistes partagent l'idée que la langue arabe (classique) que nous connaissons actuellement dérive en fait de deux parlers locaux d'Arabie, le Najd et le Hijaz.

De par ses caractéristiques morpho-phonologiques, morpho-syntaxiques et parfois même lexicales, la langue arabe, tout comme les langues dites sémitiques, est basée essentiellement sur une morpho-syntaxe de non-concaténation (non-enchaînement) de ses morphèmes. Elle diffère dans ce sens des langues Indo-européennes (de morphologie essentiellement concaténative ou de non enchaînement des morphèmes) pour lesquelles plusieurs modèles théoriques de description sont disponibles (structuralisme, fonctionnalisme, générativisme, et plus récemment le modèle auto-segmental et autres ...). Elle demeure dans ce sens une langue assez complexe à gérer dans le domaine du Traitement Automatique des Langues (TAL). Parmi ces difficultés, nous citerons les problèmes de la représentation graphique du texte arabe pour un programme de traduction automatique, plus précisément dans les cas de textes dits non vocalisés. Par exemple, une

Le traitement automatique des langues

suite de radicales non voyellisées (non-vocalisées) pourrait être interprétée par la machine de deux ou plusieurs façons différentes. Ainsi, <ktb>, dans cet ordre de radicales, va pousser la machine à chercher grâce à un premier filtre (ou filtrage) le « kernel meaning » de cette racine ou sens générique, global d' « idée d'écrire ». Elle trouvera entre autres /kataba/ 'il a écrit', /kutiba/ 'il a été écrit', /kutub/ 'des livres' etc. Certaines de ces difficultés ont trouvé solution puisqu'il est tout à fait possible de nos jours de procéder au traitement automatique du texte en arabe, de faire appel à des dictionnaires électroniques et vérificateurs d'orthographe, voire même d'envoyer des messages SMS et autres sur le Web en arabe. Tout ceci est révélateur de grands pas franchis depuis les balbutiements des années 70 face à la complexité, voire même l'impossibilité, de procéder au traitement de l'écrit de cette langue par ordinateur.

Sur le plan de la représentation graphique d'un texte, la langue arabe dispose de vingt-huit segments, qui des labiales aux vélaires /b.... k, g/ pourraient être latinisées. Mais le problème demeure, dans une certaine mesure, dans la représentation des consonnes gutturales et emphatiques, dans la représentation de la gémination qui n'apparaît que rarement dans le texte écrit (signe diacritique reconnu tacitement par le natif mais qui ne risque pas d'être reconnu, du moins aussi facilement, par la machine) et de l'allongement vocalique qui se fait dans le texte arabe en ajoutant des consonnes à la syllabe allongée. Là aussi, la machine devrait reconnaître dans quel contexte (*alif*, *waw*, *ya*) jouent leur rôle de consonne dans la chaîne parlée ou écrite et dans quel autre contexte elles représentent des signes (marqueurs d'allongement vocalique).

Sur le plan de la morpho-syntaxe, le mot en arabe ne représente pas nécessairement la même charpente, ni la même formation que le mot dans les langues Indo-européennes. En

partant du principe général que cette langue est basée sur un système de racines (trilitères et quadrilitères essentiellement), celles-ci reposent sur un sens général (en anglais kernel meaning, voir supra) qui est construit en fonction de l'agencement des radicales, représentées par des consonnes seulement alors que la racine dans les langues Indo-européennes est généralement à base syllabique, c'est-à-dire qu'il existe un agencement de consonnes et de voyelles qui constituent la racine dans les langues Indo-européennes. Ainsi, pour la langue arabe, nous dirons que la racine < ڤر > réfère à la notion générale (kernel meaning) de « faire quelque chose, produire, entreprendre, ... » alors que si l'on actionnait le processus d'agencement de cette racine en déplaçant ces radicales au sein de la racine, nous pourrions avoir < ڤر > qui dénote par contre l'idée « de soigner, soulager, traiter ... ». De même, la suite des radicales / ڤ , l, m/ dans cet ordre donne la racine < ڤlm > 'savoir, savant, ... » alors que la série /l, m, ڤ / donne la racine < lmڤ > pour le sens « briller, illuminer, brillance, ... ». Il en est de même pour les racines quadrilitères telles que < Owrt > « révolution, ... » au lieu de < Orwt > « richesse, ... ». Certes, ce ne sont pas n'importe quelles combinaisons de radicales qui vont automatiquement générer une racine nouvelle, donc un sens global nouveau. La langue dispose d'un certain nombre de contraintes imposées par la hiérarchie des sons en arabe. Parmi ces contraintes, nous citerons à titre d'exemple le fait qu'on ne peut avoir d'agrégaats consonantiques à l'initiale alors qu'en français et en anglais cela est possible comme dans *produire*, *traduire*, *produce*, *translate*, respectivement (contraintes syllabiques). La morphologie du mot devient encore plus complexe lorsque des processus de flexion et de dérivation tels que la préfixation, l'infexion, la reduplication et autres sont déclenchés. Il s'avère donc qu' alors que le natif ou néo-natif de cette langue

Le traitement automatique des langues

reconnaît relativement aisément un mot tel que « mustasmah_{un} » comme étant composé du préfixe (instrumental) {mv} suivi du préfixe {istv} qui devient infixé dans ce cas particulier (où v représente une voyelle), puis de la racine <Smfj> « pardonner, accepter, ... », le tout pour rendre le sens équivalent de 'permis' en français ou « allowed » en anglais, la question demeure à savoir si la machine serait capable d'agir de la même façon que le locuteur. Cet exemple très simple pose de prime abord un problème non seulement de reconnaissance du mot par la machine pour lui trouver grâce à ses systèmes experts et ses filtres un équivalent en français ou en anglais, mais surtout cela nécessite une expérimentation soutenue pour arriver à une segmentation de l'arabe qui est loin d'être similaire à la segmentation des langues Indo-européennes. Des propositions sont faites pour modéliser la segmentation du français ou de l'anglais et de l'adapter à l'arabe. C'est le cas du système de compilation morphologique MORPHE¹ et autres, mais qui demeurent encore très récents et sujets à débat.

Si l'on était appelé à relever certaines difficultés dans la traduction automatique de l'arabe vers le français ou l'anglais, et en considérant les parties du discours les plus élémentaires telles que l'article, la préposition, la conjonction ou le pronom (les particules) ou les catégories telles que l'adjectif, l'adverbe, le nom, ou le verbe, nous remarquerons d'emblée que la segmentation ne pourrait se faire de façon quasi identique comme ce serait le cas pour le français et l'espagnol et dans une moindre mesure pour le français et l'anglais. Nous discuterons ceux qui nous semblent les plus problématiques dans la reconnaissance des équivalents par rapport au français et à l'anglais, en l'occurrence la détermination/ indétermination, la particule de conjonction de coordination

{waw} et la catégorie du nom, plus précisément les doublets sémantiques en arabe.

En règle générale, la langue arabe diffère du français et de l'anglais principalement sur le plan de sa structure syntaxique, de ses temps et de ses modes. C'est une langue du type beaucoup plus VSO (Verbe-Sujet-Objet) que SVO (Sujet-Verbe-Objet). Elle n'a pas de forme passive, ou du moins elle a une construction du passif où l'agent n'est pas signalé. Elle fonctionne avec trois temps, le perfectif, l'imperfectif et le participe et cinq modes, l'indicatif, l'impératif, le subjonctif, le jussif et l'énergétique. En voici quelques exemples:

a) Structure VSO : kataba al kaatib kitaab lit. Ecrit l'écrivain un livre.

b) Forme passive : kutiba al kitaab lit. Ecrit (participe passé) le livre.

c) Le perfectif: katab(a)

L'imperfectif: jaktub(u)

Le participe : kutib(a)

L'indicatif : taktub(u)

L'impératif: aktub

Le subjonctif : utilisé avec des particules comme dans {lazam, alajka an, ...}

lazam taktub, alajka an taktub lit. Il faut tu écris

Le jussive: utilisé dans des constructions avec la particule de négation {lam}

lam aktub lit. pas j'écris

L'énergétique: utilisé pour exprimer la fermeté dans l'intention

sataktubu / taktubanna (rendus souvent par un impératif en français.

Il est à noter que la langue arabe ne dispose pas d'une forme d'infinitif telle que conçue en français (comme dans « écrire », « finir », « savoir ») ou en anglais (« to write », « to

Le traitement automatique des langues

finish », « to know ». Dans le cas de l'arabe, la troisième personne du masculin singulier du perfectif sert de forme verbale de base pour la conjugaison.

Pour revenir aux particules simples et leur utilisation dans la langue par rapport au français et à l'anglais, il faudrait signaler que l'article défini {al} de l'arabe est unique et invariable. Placé en début de mot auquel il est annexé, il dénote en fait la détermination par opposition à l'indétermination qui n'est pas marquée (Ar. *al* kitaab / kitaab ; Fr. *le* livre / *un* livre ; Ang. *the* book / *a* book). Par contre, le nombre est déterminé par des alternances vocaliques au sein du schème alors que le genre, pour ainsi dire, est marqué par la 'ta marbuta' {ـة} ou marqueur du féminin singulier en arabe. Il est possible, néanmoins, de trouver en arabe des noms masculins avec une 'ta marbuta' tels que Hamzatun (pour Hamza, le nom d'un chevalier arabe ou chevalier du désert). Le féminin pluriel étant constitué par un allongement vocalique suivi d'une {ta}. Ceci ne poserait en apparence pas de problème majeur de traduction automatique de l'arabe vers le français. Les difficultés seraient moindres dans ce cas précis puisqu'il serait possible de regrouper les articles définis (le, la, les) du français par exemple et de leur attribuer l'article {al} moyennant l'utilisation des mécanismes d'alternance pour le genre et le nombre. Ainsi, nous aurons :

Le livre {alkitaab}, la bibliothèque {almaktabatu}

Les livres {alkutubu}, les bibliothèques {almaktabaatu}

A l'indéfini, nous aurons :

Un livre {kitaabun}, une bibliothèque {maktabatun}

Des livres {kutubun}, des bibliothèques {maktabaatun}

En faisant abstraction des cas assez fréquents de pluriels irréguliers (brisés) en arabe auxquels il est fait appel à d'autres mécanismes de la langue, la question des équivalences concernant la détermination / l'indétermination entre l'arabe et

le français ne paraît pas soulever de problèmes majeurs dans des cas de formations isolées. Le problème est tout autre lorsque cet article {al} se trouve dans une structure syntaxique, représentée par une phrase telle que //mamnuu9 al tadxiin// qui serait équivalente à 'Interdit de fumer' ou au classique « No smoking » de l'anglais où l'on constate une absence d'article dans la traduction française ou anglaise et une présence d'article dans le syntagme de l'arabe. L'article défini de l'arabe ne peut être l'équivalent de 'le, la les' dans certains cas et l'on est obligé de faire appel à une particule en français ou en anglais. Ainsi, //kajfa al ʕaal // serait traduite par « Comment vas-tu / allez vous ? » ou bien « How are you? » (beaucoup plus que « How do you do ») entre autres que par *« comment le temps » que nous retrouvons dans plusieurs versions software de traducteur automatiques sur le marché.

La particule de coordination {waw} est également source d'ambiguïté dans la traduction automatique de l'arabe vers d'autres langues. Elle n'est pas forcément traduite par un équivalent tel que le {et} français ou bien le {and} anglais. Elle permet de relier des phrases sans pour autant impliquer une relation logique entre les phrases en question. La fréquence de son usage en arabe ainsi que les différents sens qu'elle recouvre (*wāw al 9atf*, *wāw al ma9iyya*, *wāw al qasam*, *wāw rub*, connue sous le nom de *waw rubba*, *wāw al faSl* etc.) sont source de problème en TAL pour l'arabe. Par conséquent, il est parfois difficile de lui attribuer un équivalent (en termes de coordination) dans la langue française ou la langue anglaise. A titre indicatif, dans des structures syndétiques de base de la langue arabe, telles que: {*jalastu wa jalasati ssayidatu warda*} (Gibrān Khalil Gibrān, 1949 : 112), la coordination par *wāw* exprime la simultanéité mais surtout l'indépendance des deux actions régies par deux acteurs

Le traitement automatique des langues

séparés (Moi par rapport à Warda). Dans ce cas simple, le *wāw* peut être traduit par la copule de coordination « et » du français pour donner une traduction telle que : « Nous nous sommes assis Madame Warda **et** moi ». Cependant, une structure similaire contenant *wāw* exprimant la simultanéité et l'indépendance telle que {*in tarakūhu halaka wa halakū*} (Mustafa Sādiq al Rafī'i, 1951 : 6, 5) ne peut se traduire en français par « s'ils le laissent (l'abandonnent), il périra **et** ils périront », mais beaucoup plus par « s'ils l'abandonnent ils périront tous » où l'on observe l'absence du « et » dans la traduction vers le français. En anglais, ce serait la copule « so » qui rendrait la traduction vers l'anglais, pour donner quelque chose comme : « If they leave him, he will perish and so they will ».

En langue arabe, un mot est constitué à la base par une racine composée essentiellement de deux (atypique), trois (trilitère), quatre (quadrilitère) consonnes (voir supra) au sein desquelles sont insérées des voyelles communément connues sous le nom de voyelles thématiques. Cette racine avec les voyelles insérées va constituer un « schème » ou un « thème ». Ainsi, pour reprendre l'exemple le plus utilisé pour illustrer cette composition morphologique du mot en arabe, à la racine <ktb> qui représente le sens global d' « écrire », des voyelles thématiques vont être insérées pour construire des schèmes tels que : /*kataba*/ 'il a écrit', /*kutiba*/ 'écrit masc. sing. passif', /*kitāb*/ 'livre' etc. Il est possible de faire appel à des processus de dérivation en arabe (affixation) pour générer des mots tels que /*maktaba*/ 'librairie, bibliothèque', /*kutaib*/ 'petit livre' (diminutif de *kitāb*), /*iktataba*/ 'correspondre', etc.

Selon la tradition linguistique occidentale, le mot est un signe qui peut être employé dans une forme grammaticale et qui est constitué de deux parties: une forme acoustique ou

graphique et un sens. Le mot doit donc être considéré comme une unité sémantique, phonologique et grammaticale. Cependant, des unités plus petites que le mot remplissent ces conditions : il s'agit des morphèmes qui ont également des propriétés sémantiques, phonologiques et grammaticales. Les relations entre les morphèmes lexicaux et grammaticaux diffèrent d'une langue à l'autre. Par exemple :

Ar. {zahra}{tun} (~ wardatun) {sa^hīr}{(a)tun} (~ indiquant la variation dans certains usages) (deux entrées lexicales (mots) , quatre morphèmes)

Fr. {petit}{e} {fleur} (deux entrées lexicales, trois morphèmes) (petit- m. lexical, -e m. gram., fleur – m. lexical) (fleurette ne s'utilisant rarement dans des structures telles que : « *fleurettes du brocoli, du chou-fleur* »)

Esp. {poca} (~ pequeña) {flor} (deux entrées lexicales , deux morphèmes)

It. fiorellino (une entrée lexicale, deux morphèmes) (fior – m. lexical, -ellino m. gram.)

Ang. {little} {flower} (deux entrées lexicales , deux morphèmes)

All. {Wenig} {Blume} (deux entrées lexicales, deux morphèmes)

NB. (Ces traductions ont été prises de traducteurs automatiques, sauf pour l'arabe)

Il est clair que la machine ne peut analyser de façon systématique des langues en faisant appel aux mêmes schémas de partition du matériel de langue puisque chaque langue présente en elle-même des caractéristiques qui lui sont propres, parfois similaires à d'autres langues de la même famille et ceux à tous les niveaux d'analyse (morphologiques, syntaxiques, lexicales, ...). En fait, le découpage du syntagme

Le traitement automatique des langues

en parties du discours n'est pas le même selon les langues. Par exemple, dans le syntagme arabe {man huwa al mutar im} (5 mots dont 4 morphèmes (man, huwa, al, mu-) et un substantif dérivé de la racine verbale <tr >). Dans le syntagme français {Qui (Quel) est le traducteur} (4 mots, trois morphèmes (qui, le, -eur).

De fait, la tâche du TAL n'est pas aussi simple que l'on pourrait le croire. Des problèmes surgissent non seulement au niveau des équivalences sur le plan sémantique mais aussi et surtout sur les plans de la segmentation (qui plus est lorsque les langues en question sont d'apparement génétique différent). Cette ambiguïté au niveau sémantique nous paraît beaucoup moindre que l'ambiguïté rencontrée au niveau syntaxique et en phraséologie lorsqu'il s'agit de Traduction Automatique. Ceci, bien sûr, en plus des traits culturels propres à chaque langue et les problèmes de la transposition culturelle qui risquent de surgir. Nous essayerons de traiter quelques aspects de la segmentation de l'arabe par rapport à d'autres langues (plus particulièrement le français et l'anglais), de la phraséologie et les problèmes qui en résultent, ainsi que des divergences à plusieurs niveaux que nous avons rencontrés et que nous rencontrons dans nos travaux de recherche sur l'élaboration d'un traducteur automatique (arabe/ français/ anglais).

Notes :

Le système Morphe a été conçu pour traiter un type de morphologie concaténative et il a été adapté à l'arabe (Leavitt, 1994. MORPHE: a Morphological Rule Compiler. Technical Report, CMU-CMT-94-MEMO).

Références :

- James Dickins, Sandor Hervey, Ian Higgins 2002. Thinking Arabic Translation: a course in translation method: Arabic to English. Routledge, London and New York.
- McCarthy, J. 1979. *Formal Problems in Semitic Phonology and Morphology*. Doctoral Dissertation, MIT, Cambridge, Massachussets.
- McCarthy, J. 1981. A Prosodic Theory of Nonconcatenative morphology'. In *Linguistic Inquiry*, vol. 12:373.
- Souidi, A.et Cavalli-Sforza, V. 2001 Interfacing an Arabic Morphology and Sentence Generation System with an English-to-Arabic Knowledge-based MachineTranslation System. In CNRST, PROTARS III, USA.
- Wright R. 1981. *A Grammar of the Arabic Language*, 3rd Edition, CUP. London.