Audebert Cl., Gaubert Chr., Jaccarini A., *Linguistique arabe – Programme de traitement par automates de la langue arabe (Tala)*. Dossier du nº 44 (2010) des *Annales islamologiques* de l'Institut français d'archéologie orientale du Caire (Ifao).

Le Caire, 2010.
ISBN: 978-2724705614

Le programme de traitement par automates de la langue arabe que proposent André Jaccarini (chercheur au Cnrs), Claude Audebert (Professeur émérite à l'université de Provence) et Christian Gaubert (chercheur à l'Ifao) présente ce paradoxe extrêmement stimulant d'être doublement original tout en renouant, que ce soit en linguistique ou dans la modélisation mathématique des procédures informatisées, avec des propositions faites il y a plus d'un demi-siècle.

Une première originalité de cette approche est que, contrairement à ce que l'on peut observer dans les tendances actuelles du traitement par automates de langues et de l'ingénierie des langues – développement de dictionnaires électroniques, modélisation de récurrences statistiques sans analyse linguistique, développement d'analyseurs syntaxiques couplés ou non avec des dictionnaires électroniques, construction à partir d'ontologies explicites ou implicites de cartographies sémantiques –, elle se passe pratiquement de lexique, ne s'intéresse en syntaxe qu'à des configurations d'outils grammaticaux dont elle « découvre » les effets au fur et à mesure de son balayage des textes, n'utilise les récurrences statistiques que pour valider *a posteriori* ses résultats et ne se réclame d'aucune ontologie pour décrire la construction du sens.

Une deuxième originalité est qu'elle se moule à son objet au point que celui-ci devient un nouveau paradigme d'analyse. André Jaccarini établit, à notre connaissance pour la première fois, un parallèle entre la procédure, inscrite depuis Sibawayh dans la tradition grammaticale arabe mais également consubstantielle à l'accès de tout arabophone à son lexique, d'*extraction de la racine* (p. 7) et le *théorème de l'énumération universelle* qui définit le *principe de fonctionnement des ordinateurs* (p. 18), c'est-à-dire, selon Alan Turing, l'« *existence d'une machine singulière, dite universelle, capable de simuler le fonctionnement de toute autre machine de Turing* » (p.18). La langue arabe, avec son système de dérivation, la combinatoire de ses racines et les schèmes selon lesquels ces racines se combinent et se transforment, « est, elle-même, considérée comme une machine abstraite » (p.16); dans cette perspective, celui qui la produit ou l'interprète prend des décisions ou émet des hypothèses sur son fonctionnement selon un protocole récurrent et relativement régulier; « il fait du *Calcul* (dans le sens le plus large que revêt ce terme) un objet d'expérience » (p. 23 et sur cette page la note 42). Il est donc, par rapport à la langue, dans la position, assignée par ses concepteurs à l'ordinateur par rapport à son objet.

Cette démarche « purement algorithmique et minimaliste » (p. 6), pour être inédite, n'en est pas moins une reformulation et un prolongement d'observations et de propositions dont on n'a pas su ou pu tirer tout le parti en leur temps. Celles de Joseph Greenberg d'abord.

En 1950, le père fondateur de la typologie linguistique moderne, observe, à partir d'une étude portant sur environ 6 000 racines, que « *dans les triplets R1, R2, R3 constituant les racines en sémitique, non seulement R1 ne peut jamais être identique à R2 mais que les deux premières radicales ne peuvent jamais appartenir au même groupe homorganique (deux laryngales par exemple)* » (p. 10). Cette observation permet d'établir au sein de la combinatoire *a priori* ouverte des racines des règles de compatibilité ou d'incompatibilité qui ouvrent la voie à la reconnaissance et à la prédiction du fonctionnement et des règles de transformation des combinaisons possibles [1]. Dix ans plus tard, Greenberg [2] esquissera la forme de ces règles et le cadre dans lequel il convient de les inscrire pour qu'elles soient opératoires quelle que soit la langue traitée. Ce qui est remarquable dans la démarche de Greenberg est que le modèle qu'il

(1) « It is an obvious, though little noted fact, that the characteristic triconsonantal verb morphemes of Semitic languages (the traditional triliteral roots) do not ordinarily contain identical first and second consonants. On the other hand, a pattern of identical second and third consonants is of frequent occurrence, constitutitng the well-known geminate subtype of Semitic verb. Thus, while sequences such as **mmd* are virtually non-existent in Semitic languages, Arabic *mdd* 'to stretch', *frr* 'to flee' etc. are representative of a common Semitic type. The existence of of this degree of patterning led to the present investigation of the over-all patterning of the triconsonantal verb morphemes of the Semitic languages, particularly Arabic. (…) 1. In the first two positions, not only identical but homorganic consonants are excluded. (…) 2. Homorganic consonants are likewise excluded in positions two and three, though not quite as rigorously as in the first two positions. The rule for positions two and three does not preclude identical consonants, as we have seen, so that it should be rephrased as referring to homorganic but not identical consonants. (…) 3. In positions one and three there is marked, but less rigorous exclusion of homorganic, including identical consonants, than in other combinations of positions. (…)" (Greenberg, « Patterning of Root Morphemes in Semitic », *Word* 6, p. 162).

(2) Greenberg, « A Quantitative Approach to the Morphological Typology of Language », *IJMES* 26/3, p. 178-194.

décrit et qu'il appelle « quantitatif » semble calqué sur la structure algorithmique des langues sémitiques et plus particulièrement au sein de ces langues, sur celle de l'arabe.

Greenberg [3] commence par suivre la perspective d'Edward Sapir qui réinterprète la tétratomie classique entre langues *(a) isolantes; (b) agglutinantes; (c) fusionnelles*; et *(d) symboliques*, en estimant qu'elles désignent « deux, parfois trois techniques » – l'*agglutination*, la *fusion* et l' *agglutination-fusion* qui recouvrent les degrés plus généraux de complexité ascendante *analytique, synthétique* et *polysynthétique* – dans l'association de deux « types de concepts » que toutes les langues sans exception utilisent : « un stock de racines » et « des idées purement relationnelles ». Après avoir rappelé que cette perspective a amené Sapir à classer les langues du monde en quatre *types fondamentaux* : le type simple purement relationnel (A), le type complexe purement relationel (B), le type simple relationnellement mixte (C) et enfin le type complexe relationnellement mixte (D) où l'on trouve l'arabe [4], Greenberg propose quelques modifications à ce classement afin de remplacer « les estimations intuitives fondées sur des impressions générales [5] par une définition des propriétés impliquées dans cette classification en termes de rapport entre deux unités, chacune définie avec suffisamment de rigueur et par le calcul d'un index numérique fondé sur la fréquence relative de ces deux unités dans des segments de textes [6] ». Cinq types de rapports sont proposés par Greenberg pour mener à bien ce *calcul* :

I. Le *degré de synthèse* ou de *complexité globale du mot* : rapport M/W (M pour *morphème* et W pour *mot*) ;

II. Le degré ou *index d'agglutination* : A/J (A pour le *nombre de constructions agglutinées* et J pour le *nombre de jointures* (dans les agglutinations) ;

III. La *présence ou absence de concepts dérivés et de concepts relationnels-concrets* (les morphèmes appartiennent à trois classes : celle des *racines*, celle des *dérivés* et celle des *infléchis* ; chaque mot doit avoir au moins un morphème racine et plusieurs langues en ont plusieurs : ce phénomène manifeste des différences significatives entre les langues. Il est calculé de trois manières : un *index compositionnel* R/W (R pour le nombre de *racines* et W pour le nombre de *mots*) ; un *index dérivationnel* D/W (*rapport des morphèmes dérivés aux mots*) et un *index flexionnel* I/W (*proportion de morphèmes fléchis dans les mots*) ;

IV. L'*ordre des éléments subordonnés au regard de la racine* qui se calcule de deux manières : un *index de préfixation* P/W (rapport des *préfixes* au *nombre de mots*) et un *index de suffixation* S/W (rapport des *suffixes* au *nombre de mots*). C'est ici que Greenberg signale deux phénomènes bien illustrés par les langues sémitiques et notamment l'arabe : le *containment* (une *restriction* ou *rection*, contrôlée par un élément préfixé, de la *désinence suffixée* ou de la *désinence de la flexion*) et l'*intercalation* (une *insertion* où une *portion de l'élément subordonné* précède ou suit la racine alors qu'une *autre portion* lui est *incorporée*).

V. Les *outils* utilisés pour relier les mots les uns aux autres qui sont utilisés selon trois mécanismes également calculables : (a) l'*absence de concordance* (index d'*isolation par rapport au noyau* O/N où O se rapporte à l'*ordre* et N au *noyau* ou *nexus*) ; (b) la *flexion* (index de *pureté flexionnelle* Pi/N où Pi désigne l'inflexion pure) ; (c) la *concordance* (index de *concordance* Co/N).

Soit, au total, 10 index de calcul (1960 : 16).

Résumant le principe qui se trouve à l'origine de la combinatoire particulière sur laquelle s'appuie son calcul, Greenberg écrit (c'est nous qui traduisons) : « *Ce qui est spécifié comme lié ou libre n'est pas un morphe* [7] *en tant que tel mais une classe contextuellement déterminée de morphes mutuellement substituables l'un à l'autre. Une telle classe est appelée ici classe de substitution de morphes (MSC)* [8] ». La notion de *morphe* avec les relations contextuelles que sa définition implique permet de reconfigurer les catégorisations grammaticales traditionnelles mais aussi les contraintes lexicales idiosyncrasiques en ne tenant compte que de l'algorithme du par-

(3) *Ibid.*

(4) « As an example of Sapir's total scheme, let us take his classification of Semitic. These languages are assigned as a group to D, i.e., complex-mixed relational languages with all four types of concepts present. They are synthetic. In the aerea of derivational concepts (II) the techniques are listed as *d, c,* in that order, i.e. Symbolic and fusional. In the aerea of mixed-relational concepts (III) the techniques are given as *c, d,* fusional, symbolic. Under IV the technique is *(a),* isolation, i. e., significant word order, the parentheses indicating its weak developement » (*Ibid.*).

(5) « *Intuitive estimates based on overall impressions* » (*Ibid.*).

(6) « *To define each feature involved in this classification in terms of a ratio of two units, each defined with sufficient rigor and by the calculation of a numerical index based on relative frequency of these two units over stretches of text* » (*Ibid.*).

(7) « *Un morceau de matériau morphologique, formé d'une séquence de zéro ou plus phonèmes, considéré* en lui-même, *sans aucune référence à son statut morphologique. Un morphe peut représenter un seul morphème, une séquence de deux ou plusieurs morphèmes, une partie de morphème ou pas de morphème du tout. Par exemple, le nom basque* mendi (=montagne) *forme un locatif pluriel* mendietan (= dans les montagnes) ; *nous pouvons parler de l'occurrence dans ce mot du morphe* –etan *sans nous commettre dans une quelconque analyse de cet élément.* » [C'est nous qui traduisons], (entrée *Morph* de R. L. Trask, 1993, *A Dictionary of Grammatical Terms in Linguistics*, Routledge, London – New York.

(8) Greenberg, « A Quantitative Approach to the Morphological Typology of Language », *IJMES* 26/3, p. 178-194.

cours combinatoire des morphes. Les MSC rendent possibles un calcul de la construction du sens relativement indépendant aussi bien du lexique que des aspects les plus locaux et les plus arbitraires de la grammaire.

Une démarche qui trouve sa justification, comme le reconnaît d'ailleurs Greenberg [9], dans un article de 1944, trop technique pour être présenté ici et dont la prise en compte aurait pu faire gagner un demi-siècle à la phonologie, de Zellig S. Harris. Celui qu'il faudra bien reconnaître un jour comme le plus grand linguiste du XX^e siècle [10] y démontre comment (c'est nous qui traduisons) « (…) *plusieurs faits de langue peuvent être découverts et décrits par une seule opération : l'analyse du discours en constituants simultanés. Des séquences automatiques de traits phonétiques produisent les intonations, les accents de mots et ainsi de suite* [11]. »

Les auteurs du programme TALA vont, dans ce contexte, s'appuyer sur un certain nombre de principes et de constatations mathématisables :

1. Dans un sous-ensemble important de la langue arabe dont on exclut la morphologie *non saine* (*mo^ctalla*) toute permutation, dans une phrase, des triplets consonantiques obtenus à partir de filtres déduits par inversion d'une liste donnée d'opérateurs appelés schèmes n'affecte que faiblement la grammaticalité de la phrase ;

2. Pour les mots graphiques – lexèmes agglutinés à d'autres éléments – les règles de concaténation – segmentation – demeurent invariantes par changement de racines (p. 8) ce qui autorise l'usage d'un *langage Turing complet* où toute séquence de langage peut être considérée comme une application d'un segment initial de *N* dans lui-même et autorise du coup la reconnaissance de ce langage par un automate fini puisqu'on définit ainsi une *congruence* sur ce langage dont l'ensemble des classes est fini (p. 33). Autrement dit, la *racine* est définie relativement aux opérations abstraites du calcul et non dans l'absolu. (p. 8). Le lien entre les propriétés structurales et algébriques de l'arabe et les automates et transducteurs passe par la construction de *classes de congruence syntaxique*. La relation de congruence formalise d'une certaine manière les propriétés distributionnelles de la langue (p. 31-32). C'est là le cœur du système ;

3. La morphologie générale de l'arabe est obtenue par *déformation* d'un système originel *stable*. Les productions étrangères aux règles de compatibilité phonologique sont *lissées*. Les *règles de déformation* qui ont le statut de *règles phonologiques* peuvent se définir comme simple *effet de bord d'un automate sous-jacent* ;

4. Le recours au contexte peut se substituer au recours au lexique comme l'avaient notamment montré les auteurs dans (Audebert C., Jaccarini A., 1986, *À la recherche du khabar*) et (Gaubert Chr., *Stratégies et règles minimales pour un traitement automatique de l'arabe*, 300 p., sous presse) ;

5. Les filtres et circuits correcteurs opèrent *a posteriori* mais on ne s'interdit pas d'adjoindre un *micro-lexique* fait de listes de racines attestées et de schèmes ;

6. On privilégie dans le décodage de l'arabe les éléments qui, dans le flux de l'information, font le plus baisser le niveau d'*entropie* au sens défini dans la *Théorie mathématique de la communication* de Shannon & Weaver, c'est-à-dire ceux qui comme les mots-outils de la grammaire, « les invariants lexicaux de la projection du langage sur son squelette » (n.17, p. 15) que les auteurs appellent *tokens* [12], induisent les contraintes les plus fortes et constituent des *révélateurs de structures* ;

7. La classe des langages engendrés ou reconnus par la catégorie la plus simple d'automates, les *automates à états finis* – sans piles de mémoire – coïncidant, comme le démontre le théorème de Kleene, avec la classe des langages rationnels (réguliers), il est possible d'établir une correspondance entre des expressions régulières, c'est-à-dire des formules appartenant à un langage qu'on peut définir récursivement par des propriétés de clôture et des automates finis, c'est-à-dire des programmes.

Ils aboutissent ainsi à « *une grammaire formelle ayant pour vocabulaire de base des lettres et une suite d'environ 300 mots figés* » (p. 21).

L'ambiguïté dans le codage comme dans le décodage est comme on le sait constitutive des langues naturelles. S'il est toujours possible, par une

(9) « *For the basic ideas of the MSC and the derivational sequence, I am largely indebted to the stimulus of the writings of Zellig S. Harris and Rulon S. Wells.* » (*Ibid.*, note 3).

(10) Cf. Notre introduction à: Harris, Zellig S.„ *La langue et l'information* [Trad. par Amr Helmy Ibrahim & Claire Martinot de *Language and Information*, avec une introduction sur l'œuvre de Harris par Amr Helmy Ibrahim], CRL, Paris, 1988, 98 p.

(11) Harris, Zellig S., « Simultaneous components in Phonology », *Language* 20, 1944, p. 205. Réédité dans *Readings in Lingusitics I & II* (Eric P. Hamp, Martin Joos, Fred W. Householder & Robert Austerlitz eds.), 1995, The University of Chicago Press, Chicago, p. 67.

(12) « *mots structurants sur que l'on ne peut réduire à des racines et qui n'obéissent donc pas à la morphologie de l'arabe. (…) mots-outils qui comprennent en outre des caractères qui s'agglutinent en début de mot et recouvrent en fait trois prépositions (bi, ka, li) et quelques conjonctions de coordination et de subordination. (…) entraînent des attentes.* » (p. 39). Claude Audebert étudie plus particulièrement dans le dossier les « *tokens* de discours ». Le logiciel *Kawâkib* élaboré et présenté dans le dossier par Christian Gaubert gère « 300 mots-outils (…) incluant toutes les informations de concaténation possibles afin de limiter les bruits » (p. 55).

procédure analytique qui n'est d'ailleurs pas toujours très naturelle, de lever une ambiguïté, les énoncés courants comportent toujours une part plus ou moins importante d'ambiguïté qu'il est difficile de lever. Mais il est également vrai que si dans les interactions verbales quotidiennes, les ambiguïtés, vite dissipées par les paramètres situationnels et la coprésence des énonciateurs, sont le plus souvent éphémères et de peu de conséquence, elles peuvent, à l'écrit, atteindre des degrés qui bloquent purement et simplement une analyse qui se voudrait automatique ou semi-automatique. Le programme *Tala* ne prétend pas en être exempté et prévoit à cet effet deux procédures : celle du *backtracking* (*retour en arrière* ou *rétrobalayage*) et celle de l'adjonction d'un analyseur syntaxique. Les deux étant d'ailleurs dans les faits couplées.

Le produit de cette triple et longue exploration de la langue arabe, des outils de la description linguistique et de ceux de la représentation mathématico-informatique des langues naturelles est le logiciel *Kawâkib* qui peut rendre de précieux services à pratiquement tous les publics, dans la récupération de l'information, la caractérisation des textes et le filtrage sémantique.

Conçu et développé par Christian Gaubert à partir du logiciel expérimental *Sarfiyya* conçu il y a une vingtaine d'années par André Jaccarini, ce logiciel est une application web dont le code source est en accès libre et qui peut être déployée instantanément sans installation particulière sur n'importe quel poste utilisateur moderne. Il permet l'analyse et le tri des racines d'un texte – pour le moment il reconnaît 7 200 racines trilitères – identifie les *tokens* – 300 –, peut comparer des racines, émettre des hypothèses sur un caractère manquant ou illisible ou rechercher un jeu quelconque de consonances.

Pour le moment, la version publique http://ifao.egnet.net/kawakib dans ses deux versions arabe et française – une version anglaise est en préparation – s'adresse surtout aux étudiants de l'arabe mais la version professionnelle, réservée pour le moment à l'équipe qui l'a conçue et à ses partenaires, contient toutes les fonctions présentées dans le dossier et est applicable à n'importe quel texte sans limite de longueur. Elle devrait être en accès public dans un proche avenir.

Amr Helmy Ibrahim
Université de Franche-Comté – Université Paris-Sorbonne

BCAI 27 (2012) Audebert Cl., Gaubert Chr., Jaccarini A.: Linguistique arabe – Programme de traitement par automates de la langue arabe (Tala). Dossier du nº 44 (2010) des Annales islamologiques de l'Institut français d'archéologie orientale du Caire (Ifao)., recensé par A. Helmy Ibrahim