théorique car ils réveillent la faculté psychique (*qovvat-e nafsānī*) et c'est pourquoi cet auteur recommandait de gifler le malade au visage; on associait ce procédé avec une cautérisation du sommet du crâne. En désespoir de cause, J̌orǰānī propose de suspendre le lycanthrope dans une cage bien close. Mais le praticien fait observer toutefois qu'il s'agit là de protéger le malade contre lui-même autant que de protéger les autres ; on ne le ligotera en effet que «s'il frappe les gens ou s'il se fait du mal dans ses mouvements convulsifs».

A aucun moment la lycanthropie n'est évoquée sur le mode de la diabolisation: le lycanthrope est un malade pour J̌orǰānī, certes, en nosographie, il est classé *dīvāneh* (terme construit à partir du mot *dīv*, démon), mais c'est là seulement une dénomination médicale (il existe quatre folies *dīvānegī*: la manie, la rage, la frénésie et la lycanthropie). Bokhārī classe lui aussi la lycanthropie en *dīvānegī*. Il n'y a plus que l'étymologie pour rattacher la lycanthropie à la démonologie. Il est vrai qu'Avicenne à propos de la mélancolie avait relativisé les choses. [32]

La comparaison, entre l'Orient et l'Occident, des manières d'aborder cette maladie, sur les plans étiologiques et cliniques, démontrent que la médecine scolastique orientale a sû se préserver une grande marge d'autonomie vis-à-vis de la théologie, évitant des interférences qui n'auraient produit que des aberrations. Il semblerait néanmoins, contre toute attente, qu'en Occident la lycanthropie médicale n'ait pris un tour quasi fantasmagorique que tardivement, à la Renaissance, quand la médecine occidentale éprouva la nécessité de se dégager de l'«arabisme». On crut à la métamorphose, ce qui ne fut pas le cas en Orient. L'existence des monstres n'était cependant nulle part contestée, et la raison, d'ailleurs, si elle n'admettait pas le passage d'une espèce à une autre, tolérait le foisonnement des espèces.

32. Avicenne, *Qānūn fi-ṭibb*, Ketāb II, Fann I, Maqāleh IV, Faṣl VIII, éd. Boulaq, Le Caire, Tome II, p. 67.

Luqmān, XII, 2, printemps-été 1996

Seyyed Mohammad MAHMOUDI

L'*Eżāfe* et le traitement automatique de la langue

L'une des perspectives ouvertes par l'informatique, particulièrement fascinante, est la conception de "systèmes" et de "machines" qui comprendraient notre propre langue. En fait, la réalisation d'un tel objectif dont de nombreux informaticiens et spécialistes de la transmission de l'information rêvent maintenant depuis près de trente ans, n'en est encore qu'au stade de la petite enfance. Les outils informatiques sont en voie de perfectionnement, mais les analyses effectuées jusqu'à présent, à cause de la richesse de la langue et de certaines de ses caractéristiques, sont encore loin de la perfection.

L'objectif de cet article réside dans la présentation de l'un des aspects particuliers de la langue persane qui, dans une perspective de l'informatisation, pourrait être à l'origine de nombreuses ambiguïtés morpho-syntaxiques : il s'agit de la particule de liaison d'*eżāfe*.

Après avoir présenté très brièvement un aperçu général du traitement automatique des langues naturelles (TAL) et de certaines spécificités de la langue persane, nous allons essayer de donner une description logico-syntaxique de la particule de liaison d'*eżāfe*, son rôle et ses fonctions dans une analyse automatique du persan.

Le traitement automatique des langues naturelles est l'un des domaines essentiels de l'intelligence artificielle dont les applications se sont considérablement étendues, tendance qui s'accroît continuellement. Ces applications peuvent être rangées en deux grandes catégories:

– celles qui nécessitent une "analyse automatique" de textes en vue d'une reconnaissance partielle ou complète des unités qui le constituent: mots, phrases, concepts. On trouvera ici, par exemple, la correction orthographique et syntaxique, la documentation et la traduction automatique, l'interrogation en langues naturelles de bases de données scientifiques et techniques (interface homme/machine en langage naturel) ou de bases de données à la formation et l'enseignement assisté par ordinateur, le contrôle de systèmes informatiques ou automatiques (commande de robots): dialogue avec un système expert ou un robot;

– celles qui comportent une "génération automatique" de formes linguistiques (construction d'une forme syntaxique ayant un sens), pour un système de traduction automatique, par exemple, ou bien pour formuler des réponses à des questions et l'élaboration automatique de résumés de textes. On peut aussi citer la résolution de problèmes en langue naturelle (programmation d'ordinateur, à la limite).

Parmi les nombreuses applications du TAL, on peut accorder une place privilégiée à l'indexation automatique. Cette approche, qui est basée, à priori, sur une analyse linguistique, est féconde et permet une solution solide au repérage du contenu du texte.

L'indexation automatique consiste, comme l'indexation

manuelle, à identifier et à sélectionner dans le document des éléments qui, une fois isolés et stockés dans une base de données, permettront à l'utilisateur de décrire le contenu du document concerné et de retrouver ultérieurement les informations pertinentes à une question donnée. «Ces éléments, ce sont les descripteurs. Dans un système d'information automatisé, les descripteurs sont les syntagmes nominaux des documents constituant le corpus. La solution adoptée pour l'indexation automatique est donc la constitution d'un analyseur morpho-syntaxique permettant d'extraire les syntagmes nominaux» (Le Guern, 1991: 24).

L'indexation automatique s'effectue donc par une analyse documentaire. C'est-à-dire par un processus plus ou moins formalisé qui permettrait «l'extraction du sens des documents» (Gardin, 1974: 120). Autrement dit, l'analyse documentaire devrait aboutir à la représentation du contenu d'un document.

La mise en place d'un système d'indexation automatique nécessite une réflexion préalable sur certaines formalisations théoriques et techniques. Le recours à un modèle adéquat et cohérent est l'une des conditions nécessaires pour toute application linguistique du traitement automatique.

L'élaboration d'un modèle conceptuel, c'est, du même coup, le choix d'une grammaire de référence, c'est-à-dire d'un système générateur de règles permettant le cheminement de toutes les étapes nécessaires au traitement automatique de la langue. Les différentes étapes d'une analyse automatique sont en fait: la segmentation du texte en phrases et en mots; la reconnaissance des formes lexico-morphologiques, et la reconnaissance des formes syntaxiques.

Dans le cadre de cet article nous n'avons pas la possibilité d'épuiser ce sujet, chaque étape d'un traitement automatique nécessitant en fait une étude détaillée et approfondie. Rappelons seulement que notre objectif se limite ici uniquement à la description de l'un des problèmes majeurs du traitement automatique du persan. Dans ce contexte, il n'est peut-être pas inutile de préciser que dans un processus de conception,

la tâche essentielle ne consiste pas toujours à résoudre, mais à définir plutôt les problèmes qui s'y posent.

Quelques spécificités de la langue persane

La langue persane, continuation directe du vieux-perse et du moyen-perse, appartient au groupe aryen de la famille des langues indo-européennes. Le persan contemporain conserve la plupart des traits essentiels de la langue classique. Cependant, durant le millénaire qui sépare le persan contemporain de la langue ancienne, certains changements importants ont été produits au niveau de l'écriture, au niveau lexical et au niveau du style littéraire. Ces changements sont essentiellement dus à l'influence de la langue arabe qui a été pendant longtemps la langue scientifique et de prestige en Iran.

Bien que le persan, au cours de son évolution, se soit largement enrichi, tant au niveau de la néologie qu'au niveau de la littérature, cependant, il est important de souligner que, du fait de la modification de son écriture et d'autres caractéristiques de la langue, le persan ne se prête pas facilement à une procédure d'informatisation.

Nous allons donc présenter un aperçu général des principales caractéristiques du persan qui pourraient être à l'origine de nombreuse ambiguïtés dans l'élaboration d'un analyseur morpho-syntaxique.

L'absence de l'article défini

A l'inverse de nombreuses langues, comme le français, l'arabe ou l'anglais, il n'existe aucun article défini proprement dit en persan. Certains classificateurs sont, cependant, employés devant le substantif pour indiquer le genre et le nombre du nom quantifié.

Dans le contexte de l'analyse automatique, du fait de l'absence de l'article défini, lorsqu'un nom est exposé dans le texte (sans l'usage de la marque de non-défini, *-i, yek,* etc.), deux interprétations sont possibles: le nom est défini ou

non-défini. Cette attitude rend l'analyse difficile, car elle ne permet pas de décrire la structure interne des syntagmes nominaux.

Par exemple, la séquence *dar-e khāne*, peut avoir deux traductions différentes: -[porte de maison] et -[porte de la maison]. La première est un prédicat complexe qui correspond à un syntagme nominal non défini, ouvert à la quantification, tandis que la deuxième n'est pas un prédicat complexe. Elle est en fait un syntagme nominal fermé et défini. Comment peut-on donc individualiser un concept ou un univers de discours, alors qu'on ne dispose pas d'article défini? Cela est une question fondamentale qui mérite une étude indépendante et approfondie.

De la phonologie à l'écriture

Comme toutes les langues du monde, lors d'une communication orale, les consonnes, les voyelles longues et brèves, les pauses sont explicitement prononcées et exprimées en persan. Cependant, certains éléments phonétiques, comme les voyelles, ne sont pas tous écrits.

L'écriture persane ne laisse apparaître en général que le squelette du mot, c'est-à-dire les consonnes, les voyelles longues et les supports du *hamze*. Rarement apparaît le *tašdīd* (redoublement), et jamais les voyelles brèves. De ce fait «l'écriture persane s'avère en quelque sorte un véritable piège du sens» et «reste, avant tout, une écriture étymologique» (Moïnfar, 1991 : 99).

Ainsi, dans la perspective d'une analyse informatique du texte écrit en caractères arabes dont les voyelles brèves sont graphiquement absentes, lorsqu'une séquence de caractères, qui est constituée uniquement des consonnes et des voyelles longues, est saisie au clavier pour l'analyse, l'ordinateur ne peut pas toujours analyser correctement la séquence, car plusieurs interprétations et lectures différentes sont possibles.

Par exemple, «ce que l'on écrit: *DH/*دە , peut être lu: *DeH* (village), *DaH* (dix); *MLK/*ملک peut être lu: *MaLeK* (roi,

souverain), *MaLaK* (ange), *MoLK* (territoire, pays), *MeLK* (propriété)» (De Fouchécour, 1981 : 23).

La lecture en persan consiste donc à ajouter spontanément à ce squelette les signes qui lui manquent, à prononcer syllabe après syllabe en observant la liaison entre les mots et en fonction du contexte dans lequel le mot est employé. Cela suppose la connaissance préalable du vocabulaire, du sens des mots à lire et des règles qui permettent de les combiner correctement, ce qui n'est pas actuellement possible pour l'ordinateur qui traite plutôt la forme des mots que le sens.

L'*eżāfe*

Parmi les voyelles brèves en persan, il faudra accorder une place privilégiée à l'*eżāfe,* qui comme toutes les voyelles brèves ne s'écrit pas, mais se prononce. L'*eżāfe* est, en fait, une particule enclitique de liaison *-e* ou *-ye* (après le mot terminé par une voyelle ou un *hamze*, si la voyelle finale est *e*) qui s'intercale entre deux ou plusieurs mots de façon à former une structure composée, représentant une image bien précise, qui peut être syntaxique, lexicale etc. Sa présence est aussi nécessaire pour éviter une ambiguïté de lecture. L'absence physique de cette voyelle peut, sensiblement, affecter le résultat d'une analyse automatique en persan.

Cette particule qui est nommée aussi l'*iżāfat* ou l'*iżāfa*, est parfois traduite comme "annexion" (Périer, 1901) ou "rapport possessif". Certains auteurs, à tort, l'appellent aussi le "subordonnant désinence" (voir A. Roman, 1993: 110).

En fait, à l'inverse d'une désinence qui est une terminaison variable des mots (par opposition au radical), l'*eżāfe* n'est pas une terminaison variable, elle est avant tout une voyelle brève ou une particule de liaison qui, dans un contexte bien particulier, réunit deux ou plusieurs unités du discours. Par ailleurs, il est aussi important de préciser que, contrairement à un subordonnant qui a pour rôle principal d'être associatif et commutatif (par exemple *va,* "et"), l'*eżāfe* n'est pas

commutatif.[1]

L'*eżāfe*: origine et évolution

Dans l'histoire de la langue persane, le recours à une marque particulière pour la formation des structures composées remonte à une date très ancienne. Au 2^{e} millénaire av. J.C., on découvre déjà dans les inscriptions du vieux-perse des traces et des signes de l'*eżāfe* sous sa forme primitive. Cette structure, au cours de son histoire, comme toute autre partie de la langue, subira une transformation évolutive.

En vieux-perse le signe de l'*eżāfe* était *hya*; quelques siècles plus tard en moyen-perse, ce signe se transforme en une voyelle longue *i*. En persan contemporain l'*eżāfe* subit, pour la dernière fois, une légère modification et deviendra ainsi une voyelle brève *e* (voir aussi Mo'īn, 1984: 8-9). Pour représenter l'*eżāfe* nous utilisons le symbole "z". Exemples:

En vieux-perse:
*kâra-**hya** mana* (mon armée): déterminé + z + déterminant

En pahlavi:
*âp-**i** sêp* (jus de pomme): déterminé + z + déterminant

En persan:
laškar-e man (mon armée): déterminé + z + déterminant

āb-e sīb (jus de pomme): déterminé + z + déterminant

1. Commutatif se dit d'une loi de composition portant sur deux éléments d'un ensemble et dont le résultat ne change pas si on change de place ces deux éléments. Du point de vue logique et mathématique, le commutatif se dit d'une loi de composition interne T pour laquelle aTb=bTa. Par exemple, dans une séquence des mots où *va* joue un rôle commutatif, on peut facilement changer de place des éléments constitutifs, sans qu'il y ait changement du sens et de la nature de la séquence. Ex. *madāres va dānešgāh-hā = dānešgāh-hā va madāres* (les écoles et les universités = les universités et les écoles). Par contre, dans une structure composée (syntaxique ou lexicale) où l'*eżāfe* est employée comme non commutatif, il y a une impossibilité presque absolue de changer de place les composants, tout en gardant la marque de l'*eżāfe* entre les éléments. Ex. *āb-e sard* (l'eau froide) ≠ *sard-e āb* (non-sens).

Rappelons que la plupart des grammairiens traditionnels, lorsqu'ils évoquent "les cas" en persan, emploient souvent l'abréviation "*eżāfe*" pour désigner le cas génitif (*ḥālat-e eżāfe*). Pour ne pas confondre *l'eżāfe* et le cas génitif, nous précisons que tout au long de ce travail, dans la présentation terminologique de la particule de liaison *eżāfe,* nous utilisons "l'*eżāfe*" pour désigner *kasre-ye eżāfe* qui est en fait différent du cas génitif.

Les différentes fonctions de l'*eżāfe*: où intervient-il?

Bien que la question de l'*eżāfe* laisse sous entendre que cette particule de liaison contribue uniquement à la formation de certaines structures composées, il est important de souligner qu'elle remplit aussi d'autres tâches et d'autres fonctions, surtout dans la construction de nombreuses compositions lexicales, syntaxiques et logiques, etc. On peut ainsi résumer les différentes fonctions de l'*eżāfe*:

Elément combinatoire binaire

Dans toutes les langues du monde, au fur et à mesure de leur évolution, l'homme a su, en fonction de ses besoins, inventer systématiquement des entités, lorsqu'il a été capable de concevoir une combinatoire. Du fait, aucune "invention" n'est possible que par la mise en œuvre d'une combinatoire de composantes (extrait de A. Roman, 1993: 103).

Ainsi on a inventé l'*eżāfe* en persan pour construire des combinatoires de composantes au sein d'une langue susceptible de devenir un système à part entière. L'*eżāfe* est donc devenu l'élément indispensable de toute formation complexe de la langue.

Marque de l'incomplétude

Le rôle essentiel de l'*eżāfe* est surtout de marquer l'incomplétude. Autrement dit la présence de cette particule montre que la séquence (nom ou syntagme) n'est pas terminée, qu'elle

a donc un ou plusieurs compléments. Par conséquent l'*eżāfe* est aussi une "marque de l'incomplétude" (Le Guern) et de la continuité de la séquence.

Nous rappelons que cette remarque a été par ailleurs soulignée par Moḥammad Mo'īn:

> Le nom est soit complet, et donc il n'a pas besoin d'autre(s) mot(s) 'pour s'affirmer', comme *dars* (leçon), *ketāb* (livre), etc.; ou il n'est pas complet, il a donc besoin d'autres mots pour achever son sens, comme *dars-e emrūz* (la leçon d'aujourd'hui), *ketāb-e Moḥammad* (le livre de Mohammad), etc. Le nom qui a le complément sera ainsi appelé *możāf* et le complément est appelé *możāf-un ilayh* (Mo'īn, 1984: 6).

Dans toute construction syntagmatique où le rapport est "déterminé + déterminant", l'*eżāfe* est donc employé pour lier, d'une part, les composants et marquer, d'autre part, la présence de déterminant: "déterminé + z + déterminant".

Notons qu'en vieux-perse, lorsque l'ordre était inversé, "déterminant + z + déterminé", il n'y avait pratiquement pas l'usage de la marque de l'*eżāfe* (*hya*).

Ex. *kâra* ***hya*** *mana* (armée de moi) = *mana kâra* (mon armée).

Formation des composés syntaxiques

Grâce à la particule de liaison, on peut former divers composés syntaxiques. Son rôle essentiel consiste ici à réunir toutes les unités constitutives d'une structure syntaxique. Par exemple, dans la plupart des syntagmes nominaux, où le "déterminé" précède le "déterminant", l'*eżāfe* (z) est employé pour lier les composants.

SN ⟶ déterminé + z + déterminant
bahār-e khoǰaste (printemps heureux).

Dans le langage littéraire, et parfois dans le langage familier, surtout avec un déterminant adjectival, l'ordre pourrait être inversé et l'*eżāfe* supprimé: déterminant + déterminé.

khoǰaste bahār (heureux printemps).

Formation des composés lexicaux

Le recours à l'*eżāfe* pour la formation des mots composés et les locutions prépositionnelles est aussi très fréquent en persan. Dans ce contexte, l'*eżāfe* s'intercale entre les mots de la composition, de façon à n'en former qu'un seul. La présence de l'*eżāfe* marque alors la combinaison et facilite ainsi la prononciation. Notons que les mots composés et les locutions diverses ne font pas tous l'usage de la particule d'*eżāfe*. Voici quelques exemples:

– Mot composé.

Avec *eżāfe: qābel* (adj.) + *e* + *baḫt* ⟶ *qābel-e baḫt* (adj.): discutable

Sans *eżāfe: rāh* (nom) + *āhan* (nom) ⟶ *rāh-āhan* (nom): chemin de fer

– Les locutions prépositionnelles.

Charles-Henri De Fouchécour (1981: 61-74) a recensé plus de 133 locutions prépositionnelles dont 115 sont accompagnées d'une marque de l'*eżāfe*. Ex:

Avec *eżāfe: be-muǰeb-e* (selon ···)

Sans *eżāfe: banā-bar* (selon ···)

Il ne faut pas confondre les mots composés et les composés syntaxiques. En construisant des mots composés (avec ou sans *eżāfe),* chaque élément de la combinaison perd effectivement son indépendance et l'ensemble constitue une unité singulière et indépendante sur le plan du sens. En revanche, dans une combinaison syntaxique les éléments constitutifs gardent leur indépendance et l'ensemble représente un rapport syntaxique. Rappelons que, dans le contexte d'une analyse automatique, la distinction d'un composé syntaxique et lexical est très difficile pour l'analyseur qui ne connaît pas généralement le contexte et l'aspect logique du discours.

Construction partitive

Les classificateurs exprimant une classe, une unité de mesure, une contenance, un nombre, un genre ou une espèce ne

sont pratiquement pas suivis de l'*eżāfe*. Sa présence éventuelle marque en fait une construction partitive ou une représentation anaphorique où le classificateur remplace l'objet quantifié précédemment cité. Ex:

now'-e sevvom (troisième sorte). Dans cet exemple on se réfère à un objet que l'on connaît préalablement.

Hors des expressions de quantité, la présence de l'*eżāfe* marque une détermination. Ex:

*šomāre-**ye** haft* (numéro sept) (De Fouchécour, 1981: 111).

Dans une construction partitive, lorsqu'on désigne une partie d'un tout, le numéral et le classificateur peuvent être suivis par une préposition *az* (de) ou d'une marque d'*eżāfe* pour préciser la sélection d'une partie de choses quantifiées ou qualifiées à l'intérieur de l'ensemble.

[numéral] + [classificateur] + *az* + [complément défini pluriel]: *do tā **az** baččeh-hā* (deux des enfants).

[numéral] + [classificateur] + *z* + [complément défini pluriel]: *do tā-**ye** ānhā* (deux d'entre eux).

Devant une telle situation le choix et la détermination des catégories morphologiques pertinentes sont très difficiles pour l'analyseur. Faut-il considérer cette représentation comme un ensemble de prédéterminant ou un SN entier? Cette question révèle en fait les difficultés qu'on peut rencontrer lors d'une analyse automatique de la langue.

Formation des relations variées

Tout en reliant les différents éléments constitutifs d'une structure composée, la présence de l'*eżāfe* montre aussi l'existence de différentes relations qui sont produites au sein du texte. Nous rappelons que l'*eżāfe,* en tant que tel, n'a aucune valeur grammaticale ni de sens en soi, son existence devient significative dès qu'il est exposé dans une situation particulière. De même, l'*eżāfe* «n'indique rien quant à la nature de la relation qui unit le déterminé et le déterminant. Celle-ci ressort seulement de la signification des termes en présence du contexte» (Lazard, 1957: 63).

Voici les principales relations qui pourraient se produire dans un texte par l'intermédiaire de l'*eżāfe:*

– Détermination

Dans un syntagme nominal dont le schéma de construction est "déterminé + z + déterminant", le rôle principal de l'*eżāfe* consiste essentiellement à marquer la détermination nominale. Dans certains cas, l'*eżāfe* -e remplit toutes les fonctions de la préposition "de" en français.

SN ⟶ déterminé + z + déterminant(S):

bahār-e āzādī (le printemps de liberté).

– Qualification

Lorsque l'adjectif est employé comme épithète, il qualifie le substantif et est lié à celui-ci par l'*eżāfe*. Dans ce cas l'adjectif est appelé **qualifiant** et le nom qui le précède nommé **qualifié.** On peut donc adopter le rapport suivant dont le schéma de construction est:

qualifié + z + qualifiant.

L'adjectif, tout en qualifiant le nom, peut aussi déterminer. On peut alors adopter le rapport suivant dont le schéma de construction est:

déterminé + z + déterminant:

dāstān-e tārīkhī (récit historique).

Lorsque l'adjectif est employé comme attribut, il est relié au sujet par un verbe d'état et il ne fait pas usage de l'*eżāfe*.

bahār zībā ast (le printemps est beau)

Dans le contexte du TAL, l'absence physique de l'*eżāfe* entraîne une confusion redoutable entre l'attribut et l'épithète. La distinction entre ces deux derniers relève d'une connaissance préalable du sens et du contexte, ce qui n'est pas très facile actuellement pour l'ordinateur.

– Appartenance

Lorsque l'*eżāfe* est intercalé entre un substantif et d'autres catégories de discours, il marque aussi un rapport d'appartenance, au sens réel ou figuré, avec le déterminant:

tamaddon-e šarq (la civilisation de l'Orient).

– Spécification

Lorsque le deuxième élément d'une composition spécifie le premier, l'*eżāfe* est intercalé pour marquer le spécifiant:

šahr-e Tehrān: la ville de Téhéran.

Dans certains cas l'ordre est inversé et l'*eżāfe* supprimé:

Īrān zamīn (le territoire d'Iran), *Gīlān šahr* (la ville de Guilan).

– Causal

Dans les expressions causales l'*eżāfe* joue le rôle de préposition "de":

sabab-e bīmārī ("la" cause de maladie).

– Métaphorique

L'*eżāfe* s'emploie aussi entre les termes d'une expression ou d'un ensemble de procédés susceptible de présenter un art du discours. Certaines expressions métaphoriques par exemple font l'usage de l'*eżāfe* pour lier les constituants:

ašk-e temsāḥ (larmes de crocodile), *dast-e rūzegār* (la main du temps).

L'aspect logique de l'*eżāfe*

Du point de vue logique, certains auteurs comme Jean Paul Metzger et Pierre Dupont proposent une comparaison entre l'*eżāfe* et l'opérateur étoile * en français:

La notion de prédicat complexe qui a été proposée par Alain Berrendonner (Berrendonner, 1978: 475-480), est le résultat d'une "opération *". L'opérateur * est un opérateur binaire qui permet de construire une expression prédicative complexe à partir d'un couple (ou plus) de prédicats élémentaires:

prédicat 1 + prédicat 2 ⟶ prédicat

chien + noir ⟶ chien*noir

L'opérateur *, constitutif de prédicats complexes, doit être tenu pour non commutatif et non associatif. Si, en effet, a , b

et c sont des prédicats élémentaires, nous n'avons pas le loisir d'admettre des équivalences comme:

$$(a*b) = (b*a)$$
$$(a*b)*c = a*(b*c)$$

Une autre observation pertinente qui concerne l'opérateur binaire * est qu'il n'apparaît jamais en surface sous la forme d'un morphème: les éléments constitutifs des prédicats complexes se trouvent simplement juxtaposés en structure superficielle, après avoir éventuellement subi, d'ailleurs, certaines transformations de permutation.

Enfin, il faudra observer que l'opérateur binaire * ne peut pas s'appliquer indifféremment à n'importe quel couple de prédicats: son emploi est limité par certaines restrictions de cooccurrence, qui, font, par exemple, que [maison * à (moi)] N est un nom complexe bien formé, mais que [à moi * maison] N n'en est pas un.

Selon Pierre Dupont (Dupont, 1983: 393), l'opérateur *, avec une restriction sur les arguments possibles de cet opérateur, ne peut agir que sur des noms communs, des adjectifs ou des prépositions, mais en aucun cas des verbes.

Après avoir présenté une brève description de certaines caractéristiques de l'opérateur étoile (*) et son rôle dans la constitution de prédicats complexes, nous allons maintenant essayer de comparer rapidement l'*eżāfe* et cet opérateur.

1. Contrairement à l'opérateur * qui n'apparaît jamais en surface, ni sous une forme orale, ni sous une forme écrite, l'*eżāfe* est en fait un élément phonétique qui se prononce toujours oralement, et devra apparaître normalement en surface du texte écrit graphiquement.

2. Comme nous l'avons constaté, l'emploi de l'opérateur * est, en fait, limité. Il ne peut pas s'appliquer indifféremment à n'importe quel couple de prédicats. L'*eżāfe* ne connaît pas, en revanche, ces limites. Il peut intervenir entre les couples de prédicats au sein d'un SN. Il peut aussi s'appliquer entre

n'importe quels éléments constitutifs d'un SN (nom, adjectif, participe passé, etc.).

3. L'opérateur * intervient entre deux expressions de niveau prédicatif et constitue une combinaison "interne" du concept actualisé comme une unité continue, c'est-à-dire une unité non définie; tandis que l'*eżāfe* peut intervenir entre les éléments de n'importe quel SN de nature différente (définie ou non définie). Son existence au sein d'un SN n'a aucun effet sur la nature (fermée ou ouverte à la quantification) d'un SN. En fait, en langue persane, la frontière entre un SN défini et non défini, du fait de l'absence de l'article défini, est très floue et ambiguë.

En conclusion, nous constatons que, dans la mesure où l'*eżāfe* remplit la fonction d'une préposition "de", il est en fait un opérateur binaire logique, comparable à l'opérateur * en français. Cependant nous précisons que l'intervention de l'*eżāfe* ne se limite pas seulement à la formation des prédicats complexes; elle est aussi indispensable à la construction de nombreux composés lexico-sémantiques.

L'*eżāfe* et la reconnaissance automatique des SN

La reconnaissance des syntagmes nominaux comme des formes syntaxiques est au centre des applications de grande envergure, puisqu'elle apparaît comme préalable et indispensable à l'interprétation sémantique des phrases. Aussi, interviendra-t-elle en documentation automatique (interrogation de banques de données, etc.), en enseignement assisté par ordinateur et plus généralement dans tous les systèmes de communication homme-machine utilisant une langue naturelle (Carré, 1991: 156).

A travers le syntagme nominal, considéré comme la plus petite partie du discours porteuse de référence à la réalité extralinguistique, il est possible de développer une méthodologie de conception de base de données textuelles (Bouché, 1988: 1).

Les syntagmes nominaux

Un syntagme est un groupe de mots formant une unité à l'intérieur de la phrase. Plusieurs grammairiens, avec Charles Bally, estiment que tout syntagme peut être considéré comme binaire, c'est-à-dire formé de deux éléments: un déterminé et un déterminant. Pour la grammaire dite "nouvelle", toute phrase verbale est constituée par deux syntagmes: le syntagme nominal et le syntagme verbal, solidairement unis.

Pour Michel Le Guern «le syntagme nominal est la plus petite unité nominale de discours susceptible de servir de base à une relation référentielle autonome qui permet de désigner un objet» (Le Guern, 1991: 24).

En persan, les syntagmes nominaux peuvent être divisés en trois grandes catégories: simples, continus et liés.

Les syntagmes nominaux liés

Les syntagmes nominaux liés (SNL), dont les éléments constitutifs sont reliés par des signes de l'*ezāfe*, sont d'une importance considérable en persan.

Ils sont constitués d'un noyau et d'une ou plusieurs expansions. Comme tout SN en persan la présence des prédéterminants (têtes) n'est pas obligatoire pour la constitution d'un SNL. Les expansions sont des éléments étroitement associés au "noyau", au point de former avec lui un seul SN. Les SNL ont une structure telle que le rapport entre l'expansion et le noyau, grâce à l'*ezāfe*, est une solidarité. La présence physique des *ezāfe*-s facilitera sensiblement la reconnaissance automatique de tels SN.

Schématiquement, un syntagme nominal lié est ainsi constitué de trois couches différentes: Tête (T), Noyau (N) et Expansion (E).

towse'e-ye 'olūm-e ensānī
Le développement des sciences humaines.

Le processus de reconnaissance automatique des SNL

La reconnaissance des SNL impose a priori une segmentation correcte et cohérente du texte en phrases et en mots. Ainsi, lorsque les phrases sont segmentées en mots, chaque mot du texte doit subir une analyse morphologique pour désigner sa catégorie morphologique pertinente. Puis il faudra effectuer une analyse syntaxique qui est basée sur les règles syntaxiques dites les règles de réécriture (ou PS, par abréviation de l'anglais Phrase Structure).

Les règles de réécriture permettent le cheminement de l'analyseur pour la reconnaissance automatique des formes morpho-syntaxiques différentes. L'ensemble de ces règles, appelé "Grammaire de réécriture", constitue le schéma référentiel de l'analyseur. Ces règles peuvent être facilement traduites en langage Prolog.

Ainsi par exemple, le fait qu'une phrase (P) puisse être composée d'un syntagme nominal (SN) suivi d'un syntagme verbal (SV) sera représentée par une règle de la forme:

P $\longrightarrow$ SN + SV.

De même un SN pourrait être représenté, à son tour, par une autre règle de la forme:

SN $\longrightarrow$ nom + adj., ainsi de suite.

Le processus de reconnaissance des syntagmes nominaux en persan suppose donc la mise en place d'un système de reconnaissance qui contiendrait toutes les règles de réécriture nécessaires pour le repérage des SN.

La reconnaissance des SN est en principe une tâche très difficile et complexe. En ce qui concerne le persan, cette tâche est beaucoup plus compliquée. Comment peut-on réaliser une analyse automatique correcte et cohérente alors qu'on ne dispose pas de certains éléments informatifs qui sont indispensables pour la reconnaissance automatique des SN, comme par exemple les marques de l'*ezāfe* au sein des textes.

De nombreuses expériences ont mis en évidence que cette absence graphique de la particule de l'*ezāfe*, en surface des textes, est un véritable problème pour le traitement informa-

tique du persan. Notons que dans le langage parlé, il n'y a aucune ambiguïté, les voyelles brèves, les *eżāfe*-s et les pauses sont explicitement exprimés. Pour montrer l'importance de la particule de l'*eżāfe*, examinons un exemple:

La séquence *zakhm-[e] zabān* (écrite sans les marques de l'*eżāfe* en surface du texte), qui est constituée de deux mots, *zakhm* (blessure, plaie) et *zabān* (langue), peut recevoir trois interprétations différentes:

1. la séquence est un mot composé: *zakhm-e zabān* (sarcasme, raillerie).
Ex: *zakhm-e zabān-e mardom rā čegūneh taḥammol konam* (comment puis-je supporter les sarcasmes des gens?);

2. la séquence est un SNL.
Ex: *zakhm-e zabān-e u behtar šod* (la blessure de sa langue s'est améliorée);

3. la séquence est constituée de deux syntagmes nominaux (simple et lié).
Ex: *zakhm (SNS) zabān-e u (SNL) rā pūšāndeh* (la blessure a couvert sa langue.

Cet exemple montre que l'absence des marques de l'*eżāfe* au sein des textes persans peut sensiblement affecter le résultat d'une analyse automatique et entraîner une confusion importante entre une structure composée syntaxique et une structure lexicale. De ce fait, il faudra adopter, au préalable, un système adéquat qui permettrait de réduire l'ampleur de telles confusions.

Conclusion

Pour achever, il convient ici d'évoquer brièvement quelques propositions générales, dans la mesure où on veut concevoir un analyseur morpho-syntaxique du persan pour la reconnaissance automatique des SN. Voici trois solutions qui nous sembleraient actuellement envisageables:

La transcription du persan

La conception d'un système de transcription intermédiaire qui permettrait le passage d'un système phonétique vers un système graphique, c'est-à-dire la possibilité d'une reproduction ou une représentation des voyelles brèves au sein des textes persans, peut pratiquement éliminer la plupart des confusions qui sont dues à l'écriture du persan. Dans cette optique, pour que le système fonctionne bien, il faudra que les séquences graphiques transcrites soient lisibles au sens où elles doivent restituer toutes les caractéristiques phonétiques et morphologiques du texte initial (discours oral).

Lorsqu'il s'agit du persan, tout système de transcription, du fait que certains claviers persans ne disposent pas de la totalité des caractères nécessaires, se heurte à une difficulté assez sérieuse: la reproduction des voyelles brèves et d'accessoires de l'alphabet qui ne sont pas exprimés dans l'écriture. Dans cette perspective, nous pouvons, au moins, trouver un symbole provisoire pour représenter les marques de l'*eżāfe*. Ce symbole pourrait être par exemple une étoile * qui se trouve en principe sur n'importe quel clavier du monde.

La constitution du lexique

Le lexique, qui est conçu à l'origine pour servir au repérage des SN, contient en principe un ensemble organisé de mots et leurs traits grammaticaux (catégories, sous catégories, valeurs de variables). Ce lexique fournit à l'analyseur tous les éléments informatifs nécessaires, dans l'ordre de la morphologie et de la syntaxe, pour la reconnaissance des unités du discours. Chaque élément du lexique considéré comme une entrée lexicale, sera organisé dans le lexique comme une forme ou une unité de traitement qui est la suite de caractères comprise entre deux blancs et ne comportant aucun blanc. Un texte est défini comme une suite de formes dont chaque forme ne peut s'approprier qu'une seule catégorie grammaticale ou morphologique.

Dans le cas du persan, en l'absence d'un système de transcription, la construction d'un lexique spécifique qui contienne aussi des mots composés formés avec ou sans l'*eżāfe*, peut sensiblement réduire une grande partie des confusions qui existent entre les composés lexicaux et syntaxiques. Cependant, du fait que les éléments constitutifs des mots composés sont en principe séparés par un espace, pour éviter tout découpage abusif et des démarches embarrassantes, on peut les ranger dans le lexique comme des unités indépendantes; dans ce cas les espaces internes ne seront pas pris en compte.

La génération automatique des *eżāfe*-s

En dehors de toute solution éventuelle, la génération automatique des marques de l'*eżāfe* au sein des textes est aussi un moyen très important pour l'élimination de nombreuses ambiguïtés morpho-syntaxiques et phonologiques. Cette solution, qui est la plus délicate à effectuer, nécessite au préalable une réflexion profonde basée sur une analyse morpho-syntaxique du texte. Les différentes étapes d'une génération automatique de l'*eżāfe* peuvent se résumer ainsi:

Après avoir effectué la segmentation et la catégorisation des formes lexico-syntaxiques du texte, il faut repérer toutes les séquences de mots qui sont susceptibles de se ranger parmi les SNL. Cette étape s'effectue, conformément aux règles de reconnaissance syntaxique, à partir d'une analyse syntaxique complète. Ainsi, lorsque les différentes séquences des SNL sont identifiées, l'emplacement des marques de l'*eżāfe* serait identifié et les marques de l'*eżāfe* respectives seront intercalées et représentées (Mahmoudi, 90, 94).

Bibliographie

Berrendonner, Alain, 1978, *Les référents nominaux du fran- çais et la structure de l'énoncé,* thèse d'Etat, Université de Lyon II.

Dupont, Pierre, 1983, *Eléments logico-sémantiques pour une analyse du français,* thèse d'Etat, Université Lumière-Lyon II.

Fouchécour, Charles-Henri De, 1981, *Eléments du persan,* Paris, PUF.

Lazard, Gilbert, 1957, *Grammaire du persan contemporain,* Paris, Klinksieck.

Le Guern, Michel, 1991, «Un analyseur morpho-syntaxique pour l'indexation automatique», *Le français moderne,* n° 1, juin.

Mahmoudi, Seyed Mohammad, 1990, *Analyse automatique de la langue persane: la génération automatique des eżāfe,* mémoire de DEA, présenté devant l'Université Lumière-Lyon II.

— 1994, *Contribution au traitement automatique de la langue persane: l'analyse et reconnaissance automatique des SN,* thèse de doctorat présenté devant l'Université Lumière-Lyon II.

— 1996, «Traitement des langues naturelles: évolution et perspectives», *Cahiers de recherches du Laboratoire d'Informatique Cognitive* (LIC), Université Lumière-Lyon II, 1er trimestre.

Moïnfar, Mohammad Djafar, 1978, *Grammaire comparée de l'arabe et du persan,* 2° fascicule: *Grammaire du persan,* Documents de linguistique quantitative, Jean Favard.

— 1991, «L'écriture, piège du sens»; *Contrastes,* n° 20-21, avril.

Mo'in, Mohammad, 1984, *Eżāfe* (en persane), Amir-Kabir, Téhéran.

Périer, J. B., 1901, *Nouvelle grammaire arabe,* Paris.

Roman, André, 1993, «La voie des hypertextes ?» in *Travaux du C.R.T.T.: aspects du vocabulaire,* sous la direction de Pierre J., L. Arnaud et Philippe Thoiron, PUL.