I. ANALYSE LEXICALE DES TEXTES GRECS

L'analyse lexicale est la première étape de tout traitement linguistique. Toutes les applications de reconnaissance de textes écrits, d'indexation, de correction orthographique ou grammaticale, de traduction automatique, etc., intègrent un analyseur lexical dans le but de reconnaître et de déterminer les informations qui leur sont nécessaires. Il paraît donc naturel de débuter par cette étape quand on s'intéresse au traitement automatique des langues naturelles. Dans cette étude, notre domaine de recherche est l'analyse du grec moderne. Puisqu'il s'agit d'une étape nécessaire à toute application relative aux langues naturelles, l'analyse lexicale doit être fiable si on vise des résultats satisfaisants et précis.

1 Processus de l'analyse

Pour analyser des textes en langue naturelle, il faut modéliser des phénomènes plus ou moins bien délimités, par exemple typographiques, morphologiques et syntaxiques. Lors de l'analyse typographique, nous devons identifier la fonction de chaque caractère dans la phrase (par exemple, le point peut être utilisé comme signe de ponctuation, de séparateur numérique, d'abréviation etc.). Les règles typographiques étant propres à chaque langue, l'analyse typographique est différente pour chaque langue traitée.

Ensuite, il faut identifier les unités minimales de traitement, c'est-à-dire les mots (simples ou composés) du texte, et les ramener à leur forme canonique. C'est le rôle de l'analyse morphologique. Ces traitements lexicaux peuvent être obtenus par une simple consultation de dictionnaires. L'analyse morphologique conduit à représenter un texte avec un grand nombre d'ambiguïtés qui sont introduites par la consultation des dictionnaires électroniques (M. Silberztein, 1993).

Au cours de l'analyse syntaxique, la forme canonique de chaque mot (simple ou composé) renvoie aux différentes unités syntaxiques répertoriées dans un dictionnaire syntaxique. Par exemple, la forme canonique *βάφω* (peindre, maquiller) renvoie aux tables du lexique-grammaire (c.f. Analyse syntaxique : la description des verbes) dans lesquelles sont décrits les différents emplois du verbe. De façon générale, l'analyse syntaxique identifie les propriétés distributionnelles et transformationnelles des phrases simples, qui constituent les unités élémentaires de sens. Elle permet ainsi de :

– lever les ambiguïtés lexicales en tenant compte des propriétés syntaxiques des mots dans les phrases,
– rendre équivalentes (du point de vue de l'interprétation) des formes différentes d'un même mot ou d'une phrase.

Cette brève présentation du processus de l'analyse montre que les problèmes majeurs qui doivent être traités lors de l'analyse automatique d'un texte sont, d'une part, les descriptions linguistiques qui doivent être complètes, précises et exhaustives et, d'autre part, la levée des nombreuses ambiguïtés de la langue naturelle. Le grec, comme les autres langues européennes, présente des ambiguïtés à tous les niveaux, à savoir :

– typographique (à cause, par exemple, des emplois multiples du point),
– morphologique (par exemple la forme *καθιστά*/assis/rendre, peut être analysée comme la troisième personne du singulier du présent de l'indicatif du verbe *καθιστώ* ou comme le pluriel de l'adjectif *καθιστός*/assis au neutre),
– syntaxico-sémantique (par exemple le verbe *αλλάζω*/changer a au moins deux emplois différents : *Ο καιρός άλλαξε*/Le temps a changé - *Ο Γιάννης άλλαξε το μωρό*/Jean a changé le bébé),
– contextuelle (il s'agit par exemple des ambiguïtés entre mot composé et séquence de mots simples comme *μαύρη ζώνη*/ceinture noire).

Par conséquent, les ambiguïtés ne concernent pas uniquement le niveau sémantique mais tous les niveaux de la description d'une langue donnée. Les ambiguïtés compliquent évidemment le processus de l'analyse automatique dans le sens où à un moment de l'analyse, plus d'une solution est possible pour la machine. Il est donc nécessaire de procéder à la levée des ambiguïtés le plus tôt possible. M. Silberztein (1993) a démontré qu'il est envisageable de lever un certain nombre d'ambiguïtés, par exemple avant l'analyse syntaxique, grâce à des grammaires locales qui explorent sommairement les contextes de certains mots. Cependant, il faut noter que, malgré les efforts qui lui sont

consacrés depuis plusieurs dizaines d'années, l'analyse automatique des textes ne donne aujourd'hui que des résultats trop approximatifs pour être véritablement utilisables.

2 Analyse typographique

Les textes que nous analysons sont issus de systèmes informatiques. Ils sont destinés à être lus ou échangés via les réseaux. Par conséquent, les fichiers correspondants contiennent des codes d'enrichissement typographique, comme par exemple « gras », « Times New Roman 12 points », etc. À noter que cet enrichissement typographique est intégré dans les textes sous forme de balises (SGML, XML) ou séparé des données du texte. Il suffit donc de les repérer dans les textes et de les extraire. Par ailleurs, les traitements de texte permettent la sauvegarde des textes sans enrichissement typographique. Quel que soit le codage adopté, il est toujours possible de convertir ces fichiers en fichiers normés ne contenant que la séquence des lettres, chiffres, séparateurs et signes de ponctuation. En plus des lettres de l'alphabet de base d'une langue, les textes sont susceptibles de contenir des lettres capitales, des lettres étrangères, des apostrophes, des traits d'union etc. Par la suite, nous présenterons certains signes typographiques qui présentent des particularités en grec, et la façon dont ils ont été traités dans le système d'analyse que nous implémentons. L'analyse typographique ne fait pas appel à des dictionnaires mais à des grammaires qui décrivent les contextes caractéristiques des formes traitées. Pour cette partie, nous nous sommes largement inspirée de l'étude réalisée par M. Silberztein pour le français (1993).

2.1 L'apostrophe

En grec, l'apostrophe peut apparaître entre deux mots (élision). Les grammaires traditionnelles (cf. M. Triandaphyllidis, 2000) distinguent trois cas. Le premier est le phénomène de « έκθλιψη »/ekthlipsi[3], qui concerne deux mots dont le premier se termine par une voyelle et le deuxième commence par une voyelle et il y a effacement de la dernière voyelle du premier mot :

το αβγό - τ' αβγό (l'œuf),
το αγόρασα - τ' αγόρασα (je l'ai acheté),
θα αγοράσω - θ' αγοράσω (j'achèterai).

Le deuxième est celui de « αφαίρεση »/aphérèse, qui concerne deux mots

3. Translittération.

dont le premier se termine par une voyelle et le deuxième commence par une voyelle et il y a effacement de la première voyelle du deuxième mot :

θα έρθω - θα 'ρθω (je viendrai)

Enfin, le dernier est le phénomène de « αποκοπή »/apocope, qui affecte deux mots dont le premier se termine par une voyelle et le deuxième commence par une consonne :

πάρ' το (prends-le)

L'apostrophe devrait être traitée lors de l'analyse typographique, mais cela n'est pas possible sans dictionnaires (cf. section 3.1).

2.2 Le trait d'union

En grec, le trait d'union est utilisé dans différents cas[4] (cf. aussi D. Holton, P. Mackridge, I. Philipaki-Warburton, 2000 et M. Triandaphyllidis, 1975 et 2000) :

- comme signe de ponctuation lors d'une énumération ou d'un dialogue (il s'agit des tirets),
- comme division dans un mot coupé par la fin d'une ligne (situation parfois improprement dénotée par le terme césure),
- dans un certain nombre de mots composés : *παιδί-θαύμα* (enfant miracle) et les noms de famille : *Αναστασιάδη-Συμεωνίδη* (Anastassiadis-Symeonidis),
- dans des symboles mathématiques pour désigner le négatif ou la soustraction,
- dans les dates (10-2-2002) ou pour indiquer une période chronologique (1949-1980).

Les tirets sont identifiés lors de l'analyse typographique et ils sont considérés comme des signes de ponctuation. Les divisions en fin de ligne ne sont pas traitées car les codes correspondants peuvent être supprimés au niveau du traitement de texte. Les traits d'union dans les mots composés, les noms propres et les dates sont identifiés respectivement lors de la reconnaissance des mots composés (cf. chapitre 8) et des dates (Voyatzi, 2002), c'est-à-dire au niveau lexical. Les autres doivent être examinés dans des applications particulières.

2.3 Les signes de ponctuation

Les signes de ponctuation grecs diffèrent des signes des autres langues

4. Nous ne présentons ici que les cas qui intéressent le traitement automatique.

européennes et sont exposés ci-dessous :

> la virgule (,), le point (.), les deux points (:), le point-virgule marqué par le signe (·), le point d'exclamation (!), le point d'interrogation marqué par le signe (;), les points de suspension (...), les parenthèses ouvrante et fermante(), le tiret (-), ainsi que les guillemets (« ») et (").

En grec, la virgule, le point, les deux points, le point-virgule, le point d'exclamation, le point d'interrogation, les points de suspension, les parenthèses fermantes et les guillemets fermants sont collés au mot qui les précède. Les parenthèses ouvrantes et les guillemets ouvrants sont collés avec le mot qui les suit.

Enfin, rappelons que les autres séparateurs utilisés pour la reconnaissance des mots simples sont les blancs et les retours-chariot.

2.4 Les séparateurs de phrases

En grec, le point (.), le signe (·) utilisé comme un point-virgule, le point d'exclamation (!), le point d'interrogation (;) et les points de suspension (...) peuvent être des séparateurs de phrase, mais pas uniquement (cf. D. Holton, P. Mackridge, I. Philipaki-Warburton, 2000). Les différents emplois de ces signes seront explicités ci-dessous.

Le point est utilisé :
- à la fin de la phrase ;
- dans des abréviations *(π.Χ.*[5]*, π.χ.*[6]*)* mais pas toujours : *ΗΠΑ*[7] ;
- dans des prénoms abrégés suivis des noms (*Α. Παπαδόπουλος*/A. Papadopoulos) ;
- dans les chiffres (1.234.567) et la notation de l'heure (10.45).

Le signe (·) a les mêmes emplois que le point-virgule des langues européennes, à savoir :
- séparation entre des phrases qui sont fortement liées entre elles ;
- avant une phrase qui est une explication de la précédente.

Le point d'exclamation (!), hors parenthèses, marque la fin d'une phrase qui exprime l'admiration ou se place à l'intérieur d'une phrase après une interjection.

Les points de suspension (...) sont utilisés à la fin d'une phrase pour

5. *προ Χριστού*/avant Jésus-Christ
6. *παραδείγματος χάρη*/par exemple
7. *Ηνωμένες Πολιτείες Αμερικής*/Etats-Unis

montrer qu'elle est incomplète ou encore à l'intérieur d'une phrase après un mot qui va étonner le lecteur ou même à la place d'un mot.

Enfin, le point d'interrogation (;), hors parenthèses, est utilisé à la fin d'une phrase interrogative.

2.5 Problèmes liés à la délimitation d'une phrase

Pour le traitement automatique des textes, la délimitation des phrases constitue la première étape délicate. En effet, si on ne reconnaît pas les phrases d'un texte, il est difficile de faire une analyse linguistique acceptable.

Nous venons de voir (à la section précédente) que les signes de ponctuation ne marquent pas forcément la fin d'une phrase. De plus, la règle élémentaire qui impose une majuscule en début de phrase n'est pas toujours une indication suffisante. En effet, une majuscule peut se trouver :

– au début d'une phrase :
  Παρομοίως, μια επιρρηματική φράση μπορεί να χωρίζεται με κόμματα.
  De même, une phrase adverbiale peut être séparée par des virgules.
– dans les abréviations (*Φ.Π.Α*/T.V.A) ou les noms propres (*Μαρκόπουλος*/Markopoulos, *γλώσσα προγραμματισμού C*/langage de programmation C).

Notons aussi que pour des raisons stylistiques ou d'emphase certains auteurs écrivent une partie de leur texte tout en majuscules :

> *Επιλέγει επίσης τις ταινίες του ΔΙΕΘΝΟΥΣ ΦΕΣΤΙΒΑΛ ΚΙΝΗΜΑΤΟΓΡΑΦΟΥ και καθορίζει την ημέρα προβολής τους.*
> Il sélectionne aussi les films du FESTIVAL INTERNATIONAL DU FILM et détermine le jour de leur projection.

Le fait qu'un signe de ponctuation soit suivi d'une majuscule n'est pas suffisant pour marquer la fin d'une phrase et le début d'une autre. Pour illustrer cette difficulté, examinons les exemples ci-dessous qui incluent le point, le signe de ponctuation le plus ambigu.

(1) Οι καρέκλες είναι μαύρες. Το τραπέζι επίσης.
  Les chaises sont noires. La table aussi.
(2) Το όνομα του γράφεται με Ι. Μου το είπε η Μαρία.
  Son nom s'écrit avec un I. Marie me l'a dit.
(3) Το ανέφερε ο Ι. Μελισσανίδης στην συνέντευξή του.
  Cela a été cité par I. Melissanidis lors de son interview.
(4) 1.2.3. Επίλογος
  1.2.3. Conclusion

(5) Εμείς στο ΔΗ.Κ.ΚΙ αγαπητοί συνάδελφοι είμαστε χαρούμενοι άνθρωποι.
Nous au DI.K.KI[8] chers collègues nous sommes des gens heureux.
(6) Ο κ. Σημίτης θα βρίσκεται στις Βρυξέλλες για τη σύνοδο κορυφής.
M. Simitis se trouvera à Bruxelles lors de la conférence au sommet.
(7) Ο Κ. Σημίτης θα συναντηθεί με πολλούς συναρμόδιους υπουργούς για να αποφασίσουν.
K. Simitis rencontrera plusieurs ministres compétents pour prendre une décision.

Les phrases (1) et (2) sont des cas simples où le point suivi d'une majuscule marque la fin d'une phrase. Dans les phrases (3)-(7), le point ne marque pas la fin de la phrase. Dans la phrase (3), le « I. » est la lettre initiale du nom propre *Ιωάννης* (Jean). Le « I » de la phrase (2) n'est pas une abréviation et le point est la fin de la phrase. L'exemple (4) est un titre de section. On peut trouver le même type de numérotation dans des énumérations : 1., 2., ...

Signalons aussi que les numéros des titres et des énumérations peuvent apparaître entre parenthèses. Dans l'exemple (5), « ΔΗ.Κ.ΚΙ. » est l'abréviation du nom d'un parti politique. Enfin, dans les deux derniers exemples les « κ. », « Κ. » sont des abréviations mais elles ne désignent pas la même chose. En effet, « κ. » est l'abréviation courante pour le mot *κύριος*/monsieur, et « Κ. » en majuscule est la lettre initiale du nom propre *Κωνσταντίνος*/Konstantinos ou *Κώστας*/Kostas.

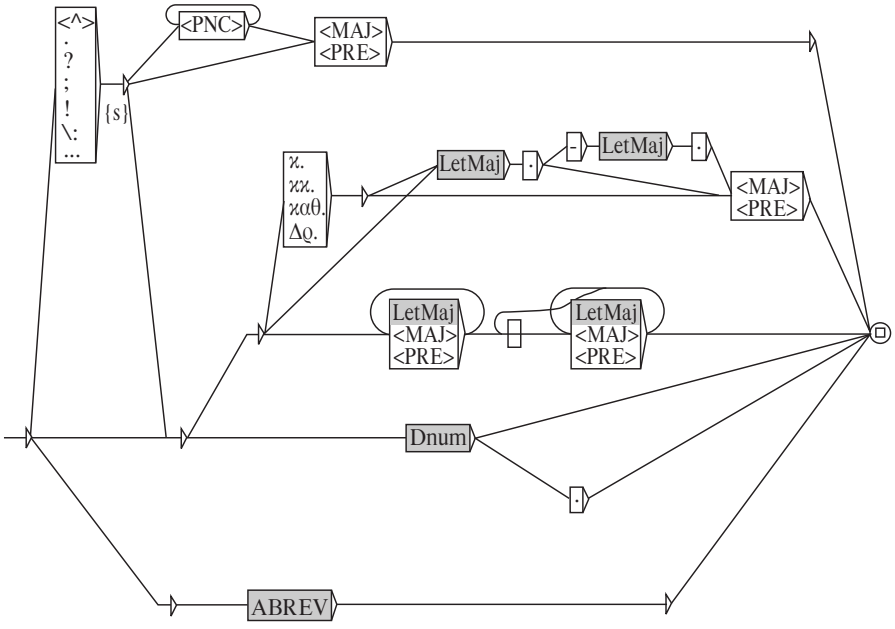
Ces exemples montrent que les ambiguïtés causées par les abréviations, les chiffres incluant des points, et les combinaisons : Abréviation-Nom Propre, Nom Propre-Nom Commun, rendent l'identification des phrases très difficile, voire impossible dans certains cas. La seule solution serait que l'analyseur accède à des listes complètes d'abréviations, de noms propres et de chiffres. Pour les chiffres, la grammaire a été faite et elle a été incorporée dans le graphe de reconnaissance des phrases (cf. section 2.7). La liste des abréviations ne comprend aujourd'hui que 350 abréviations courantes (cf. aussi section 3.4) comme *κ.ά.* (*και άλλα*/et d'autres), *κ.λ.π.* (*και τα λοιπά*/etc.) ainsi que les sigles d'organismes nationaux Δ.Ε.Η/E.D.F. Cette liste doit être complétée. De façon générale, nous traitons les abréviations « standard » mais il peut en exister d'autres plus dépendantes des auteurs ou des textes. Le traitement des abréviations des noms propres suppose tout d'abord d'avoir une liste complète des noms propres. Une telle opération est difficilement

8. Il s'agit d'un parti politique grec.

envisageable à l'heure actuelle, elle demanderait des années de travail (cf. aussi 3.5). En attendant d'avoir des listes exhaustives, le graphe de séparation de phrase prévoit que, quand on a une ou plusieurs majuscules suivies par des points, alors il ne s'agit pas d'une phrase. Cette règle donne des résultats satisfaisants mais certains cas perturbateurs comme *Γιάν. Κεφαλογιάννης* (Gian. Kefalogiannis) ne sont pas reconnus.

2.6 Graphe de délimitation d'une phrase

Nous présentons ci-après le graphe de délimitation des phrases que nous avons élaboré (cf. aussi section 3.4). Les résultats obtenus après application de ce graphe sont présentés en annexe.



Les parties grisées renvoient à des sous-graphes (M. Silberztein, 1993). Pour une présentation plus détaillée concernant les grammaires locales et les graphes voir aussi E. Roche, Y. Schabès, 1997, M. Gross, 1993, 1995, 1997.

Soulignons aussi que les proverbes complexes composés de deux phrases devront être rajoutés ultérieurement au graphe de reconnaissance des phrases. En effet, nous avons des proverbes du type :

Άγιε Γιώργη βόηθα με. Κούνα και συ τα χέρια σου.
Aide-toi, le Ciel t'aidera.

Pour cet exemple, le graphe actuel reconnaîtra deux phrases séparées, ce qui posera des problèmes au niveau de la reconnaissance des proverbes. Il faudra donc rajouter ces proverbes dans le graphe pour pouvoir reconnaître ces séquences complexes et leur attribuer une étiquette par exemple du type S1.

2.7 Alphabet et caractères latins

Les textes grecs que nous traitons sont écrits avec l'alphabet grec, qui comporte 24 lettres (minuscules et majuscules) décrites dans un fichier du système d'analyse. Cependant, il faut noter que nous rencontrons aussi, dans les textes, des mots écrits en caractères latins ou même des mots composés avec une partie écrite en caractère latin et une autre en caractère grec : *software-μωρό*/software-bébé (Tr. Littérale). Ces mots, qui appartiennent le plus souvent au domaine des Nouvelles Technologies, sont codés tels quels dans nos dictionnaires. Dans les textes traités, nous avons aussi trouvé des cas où le début du mot est écrit en caractères latins et la fin du mot (c'est-à-dire la terminaison[9]) en caractères grecs : *cliκάρεις*/tu cliques. Nous n'avons pas pu vérifier l'exactitude de ces formes pour trois raisons essentielles : elles sont absentes des dictionnaires, les locuteurs natifs n'ont pas pu nous renseigner et, statistiquement, le nombre de cas rencontrés est insuffisant pour pouvoir décider s'il s'agit de formes correctes ou d'erreurs. Par conséquent, ces formes n'apparaissent pas dans nos dictionnaires et ne sont pas reconnues actuellement par le système d'analyse.

À noter aussi que la majorité des ordinateurs en Grèce incluent l'alphabet grec et un alphabet latin (anglais ou français en général). Cette spécificité entraîne beaucoup d'erreurs de la part des utilisateurs. Ces erreurs concernent essentiellement les caractères et les signes de ponctuation semblables des deux systèmes d'alphabet qui ne se distinguent pas à l'œil nu, comme par exemple A et A, B et B, N et N, ...et ..., etc. Ces erreurs compliquent tant l'analyse typographique que l'analyse morphologique du grec puisque, très souvent, les mauvais résultats d'analyse ne sont pas dus à des descriptions incomplètes et imprécises, mais à des erreurs liées aux deux alphabets. Ainsi, dans les textes journalistiques, nous avons repéré plusieurs fautes du type *Ν. Γκεσούλης*/N.

9. Cf. aussi section 7.4.2.

Gessoulis où le « . » était écrit en caractère latin ou encore *Αμερικανών* (des Américains) avec un A qui n'était pas de l'alphabet grec. De même, nous avons constaté quelques erreurs du même type dans nos dictionnaires où par exemple la dernière lettre « o » de l'adjectif *όμορφος*/beau était écrite en caractère latin. Nous citons ces problèmes pour montrer la difficulté du processus de l'analyse et la complexité de la vérification des données.

2.8 Corpus

Les textes que nous traitons sont issus du réseau Internet. Ils nous ont été fournis par le laboratoire CIS (dirigé par le professeur Franz Guenthner) de l'Université de Munich. Ce laboratoire a développé un programme qui intercepte les textes disponibles sur le Web et les trie par langue. Les textes grecs ainsi réunis représentaient 416MB. Mais ces textes, récupérés automatiquement, comportaient différents symboles informatiques qui gênaient le système d'analyse. Il a donc fallu « nettoyer » ces textes des caractères tels que #, {, }. Cette opération de « nettoyage » a été entreprise par une vingtaine d'étudiants de l'Université de Thessalonique.

Cependant, certains symboles doivent être identifiés dans les textes, puisqu'ils ont un sens linguistique précis et sont parfois même utilisés pour représenter des mots abrégés : *% τοις εκατό*/pour cent, $δολλάριο/dollar. Ces symboles n'ont pas été enlevés des textes et ils seront discutés dans la section 3.6.

Après le « nettoyage », la taille de notre corpus est de 280MB, ce qui représente environ 40 000 000 mots. On peut prétendre qu'il s'agit du plus grand corpus disponible aujourd'hui pour le grec. Pour son traitement automatique, il a été divisé en 3 fichiers.

3 Reconnaissance automatique des mots

Une fois la séparation des phrases terminée, l'analyseur lexical doit reconnaître les mots du texte. Ces mots peuvent être écrits en caractères grecs ou latins, en minuscules, en majuscules, ou les deux (*Παπαδόπουλος*/ Papadopoulos, *ΕΤΕπ (Ευρωπαϊκή Τράπεζα Επενδύσεων*/*Banque Européenne d'Investissements*, *FoxPro*, *kHz*). Le plus souvent les mots qui composent les textes apparaissent sous une forme fléchie : les verbes sont par exemple conjugués, les noms et les adjectifs déclinés. L'identification des mots consiste par conséquent à effectuer la lemmatisation des mots du texte, c'est-à-dire retrouver la forme canonique des verbes (1ère personne du présent de l'indicatif), des noms (le nominatif singulier) et des adjectifs (le nominatif masculin singulier), puis associer à