

MOURAD G. (1999). La segmentation de textes par l'étude de la ponctuation; Acte de colloque international, CIDE'99, Document Electronique Dynamique, pp. 155-171, Damas, Syrie.

La segmentation de textes par l'étude de la ponctuation

Ghassan Mourad

Equipe Langage, Logique, Informatique et Cognition (LaLIC)

Centre d'Analyse et de Mathématiques Sociales (CAMS)

UMR 8557 du CNRS, EHESS, Paris-Sorbonne

96, Boulevard Raspail 75006 PARIS – France

Tél. : (33) 01 44 39 89 63

Ghassan.Mourad@paris4.sorbonne.fr

Résumé : La segmentation de texte est une phase nécessaire pour un très grand nombre d'applications en traitement automatique du langage : par exemple pour l'alignement des phrases dans les systèmes de TAO, pour l'analyse syntaxique, pour le résumé automatique, pour le filtrage de textes, etc.

Nous essayerons, dans cet article, d'expliquer notre démarche pour développer un segmenteur de texte en segments textuels. Nous aborderons également les problèmes d'ambiguïté que suscitent les différents emplois des signes typographiques. Enfin nous signalerons brièvement ceux liés à la segmentation des textes d'autres langues et en particulier l'arabe.

Mots-clefs : exploration contextuelle, ponctuation, segmentation, segment textuel, TALN.

1. Introduction

La segmentation de textes en phrases (segments textuels) reste une phase préalable pour le traitement automatique des langues. Cette phase de traitement (à notre connaissance) n'est pas prise très au sérieux par la plupart des laboratoires qui traitent la langue automatiquement. Chaque équipe de recherches développe un outil provisoire pour des corpus bien définis (ou bien nettoyés de toute sorte de « scories »), ou n'a recours qu'à un traitement manuel. Or dans de nombreuses applications les textes ne sont pas préparés.

La segmentation de textes est basée sur l'étude linguistique d'une part, sur une modélisation informatique d'autre part. Ces deux études se complètent. La

segmentation a, comme d'autres types de traitement automatique de la langue, ses particularités, que ce soit au niveau linguistique, ou au niveau informatique. Et comme le signale C. Fuchs, la phrase a une place privilégiée dans le traitement linguistique pour des raisons diverses : d'une part on sait mal décrire des unités de taille supérieure ; d'autre part, en raison de la spécificité du fonctionnement qui ne se trouve pas dans les unités inférieures (mots ou syntagmes) [FUCH93]. C'est au niveau de la phrase que se construit la prédication. Cette définition répond à la question de la grammaire classique : qu'est ce qu'une phrase ? Mais les définitions traditionnelles ne sont pas toujours applicables en traitement automatique des langues, surtout si on prend en compte le traitement de la ponctuation, car les signes de ponctuation constituent une marque pivot pour identifier les phrases. Chaque signe de ponctuation porte en soi des informations qui peuvent remplacer des phrases ou des énoncés. Cette information, dans bien des cas, est aussi informative que la phrase dans sa définition classique, comme nous le montre l'exemple suivant :

(1) « - Alors je suis allé travailler.
- ?

Il sourit de nouveau pour dire que mon étonnement ne le surprend pas et répond :

- Oui, je suis parti en quelque sorte à la quête de la vérité concrète, ça m'a permis de réfléchir sur la réalité. » Exemple emprunté de L. Védénina [VÉDÉ89].

Le point d'interrogation évoque l'étonnement (état émotionnel) et le temps écoulé entre les répliques. Selon Védénina l'un des personnages ne trouve pas de paroles pour s'exprimer.

Notre approche est de définir un segmenteur de textes en segment textuel¹ à partir d'une étude systématique des marques de ponctuation. Dans bien des cas, nous identifions des unités textuelles qui coïncident avec des unités qui correspondent aux critères classiques de la langue. La segmentation est basée premièrement sur des marques de ponctuation « . », « ; », « : », « ! », « ? », « \r »² (qui sont considérées comme des marques pivot pour le déclenchement des règles de segmentation), et deuxièmement sur une étude des contextes gauches et droites de ces marqueurs.

L'implémentation informatique a été effectuée sous JAVA en collaboration avec G. Crispino³. Le projet a bénéficié du soutien du programme ECOS. Uruguay (action U9701).

2. Étude de l'existant

Les outils qui existent sur le marché en tant que segmenteurs-baliseurs de textes comme HTML utilisent, pour des textes bien structurés, des balises jusqu'au

¹ Pour les raisons liées à la définition de la phrase, dans cet article la notion de segment textuel et de phrase peuvent coïncider.

² Retour à la ligne

³ Enseignant chercheur à l'Université de Montevideo, Uruguay.

niveau paragraphe ; mais la segmentation des textes en unités plus petites («phrases ») reste une tâche qui n'est pas bien définie actuellement.

Pour la segmentation des textes français, Anne Dister, de l'université de Liège, a développé un segmenteur en utilisant le système INTEX [SILB93]. Pour procéder au découpage du texte en phrases, elle a appliqué un transducteur (automate qui lit une séquence dans un texte et par rapport à l'information associée à celui-ci, l'automate insère la marque de fin ou de non-fin de cette séquence), les marque utilisées pour cette phase de segmentation sont les « . », « ! », « ? ». La règle générale est appliquée (après la levée de certains cas d'ambiguïté) dans les cas où l'on rencontre la séquence « . » ou « ? » ou « ! » suivi d'une majuscule [DEFA98].

Pour les textes anglais, Jeffrey C. Reynar et Adwait Ratnaparkhi de « *Departement of Computer and Information Science* de l'université de Pennsylvania) ont développé un outil de segmentation de texte (« *A Maximum Entropy Approach to Identifying Sentence Boundaries* ») en utilisant les « . », « ! », « ? » pour segmenter un texte. Leur travail consiste à définir le suffixe ou le préfixe d'un signe candidat et des caractères qui y sont liés, ainsi qu'une liste des abréviations. Dans d'autres cas, ils étudient la morphologie d'un mot avant ou après ce candidat [JEFF97].

L'outil SATZ (phrase en allemand) est un autre système développé par David D. Palmer de l'université de California (SATZ – *AN Adaptative Sentence Segmentation System*). Palmer utilise un réseau neuronal, en étudiant par des critères lexicaux le contexte gauche et le contexte droit de chaque candidat (dans son cas les « . », « ! », « ? » définissent la fin d'une phrase) [PALM94].

La plupart des segmenteurs existants sont limités à la simple utilisation des marques de ponctuation « . », « ! », « ? », avec une étude des quelques cas d'ambiguïté sur des corpus bien déterminés, et une utilisation de dictionnaires de sigles. Le procédé de siglaison peut être ainsi introduit sans aucune convention, donc le dictionnaire doit être enrichi de chaque nouveau sigle (pour plus de détail sur la siglaison cf. [CALV80]).

Dans les trois outils mentionnés ci-dessus, des problèmes d'ambiguïté apparaissent : premièrement ils n'ont pas pris en compte l'espace entre les candidats « . », « ! », « ? » et le segment suivant ; deuxièmement ils n'ont pas étudié les cas des points entre parenthèses ; troisièmement ils n'ont pas pris en compte les points à l'intérieur des guillemets ; quatrièmement les cas des segments qui commencent par des chiffres arabes (ex. 1) , par des guillemets, par des parenthèses etc. Que dire enfin des segments qui se terminent par des points-virgules ou par deux-points (nécessaires pour une étude syntaxique) ? Par ailleurs, un problème n'est pas mentionné : le cas des énumérations hiérarchiques. Selon nous, une segmentation fiable devrait s'intéresser à ces problèmes.

(1) «*Du rêve à la réalité, la population de Mexico a donc baissé de 58 % en quinze ans, passant de 31 millions à 18 millions d'habitants. 13 millions d'hommes et de femmes se sont ainsi évanouis d'un cauchemar purement statistique en moins de vingt ans.* »

3. Étude linguistique associée à la conception du système

Précisions:

– L'étude linguistique a été basée sur une étude de corpus de textes usuels et spécialisés.

– Nos corpus de textes sont les suivants : 60 articles de taille longue du «Le Monde Diplomatique » CD-ROM 1989-1996 ; un fichier d'aide de WINDOWS 97 ; et un corpus de textes qui contient des articles scientifiques, linguistiques, psycholinguistiques (corpus *Spirale*⁴).

– Une faute de ponctuation est considérée comme une faute d'orthographe de l'auteur. Exemples : laisser deux espaces après un point à l'intérieur d'une phrase, ou ne pas laisser un espace après un point pour le début d'une phrase. Ce problème peut être réglé en supprimant les espaces excédentaires.

– Ce travail a été effectué dans le cadre de la plate-forme CONTEXT⁵. La segmentation de textes est adaptée à chaque tâche, mais les traitements ne sont pas limités par des types d'applications envisagées dans CONTEXT.

– Des exemples sont pris dans des livres qui traitent de la ponctuation afin que ce segmenteur puisse être utilisé pour tout type de traitement.

Dans le projet CONTEXT, nous réalisons d'abord l'extraction des phrases pertinentes selon un certain point de vue, puis, par une étude de la structure discursive et enfin l'étude de cohérence textuelle, nous pouvons de plus nous intéresser aux phrases suivantes ou précédentes, aux paragraphes suivants ou précédents.

La ponctuation en tant que système graphique est pauvre car le nombre de signes est réduit, mais très riche par rapport à l'utilisation de ces derniers. Les signes graphiques qui jouent un rôle dans la phase de segmentation sont tous ambigus. Cette ambiguïté dans laquelle le point joue un grand rôle ne sera pas levée si on ne prend pas en compte les contextes gauches et droites. Le schéma (fig. 1) [MOUR99] du caractère « point » (qui a été le premier signe utilisé, avant même l'espace entre les mots du système de ponctuation [CATA77][CATA94]) en tant que marque typographique nous montre différentes utilisations de ce signe de ponctuation :

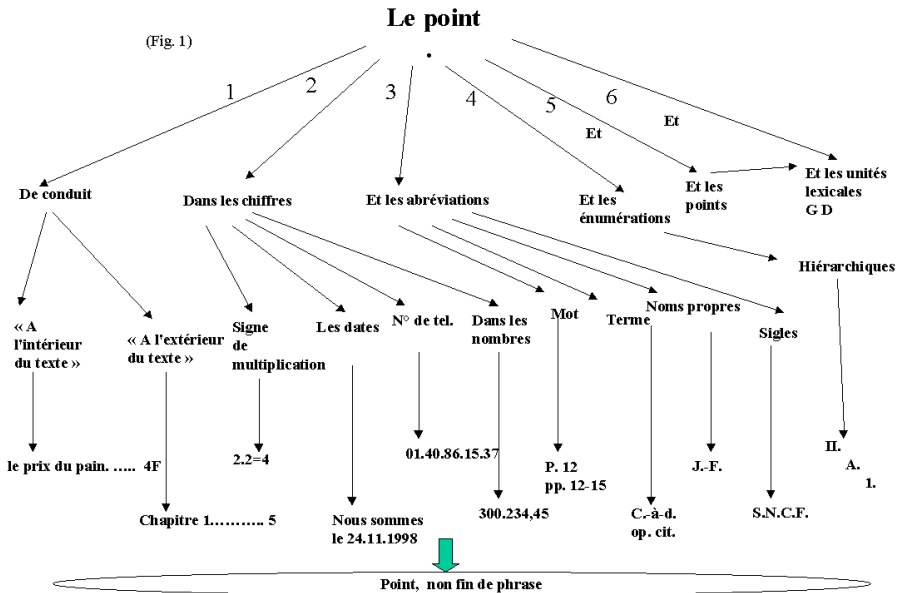
Une lecture descendante de ce schéma nous donne un premier aperçu du point en tant que marque de non fin de segment textuel.

Les branches 1 à 4 de la figure 1 détaillent les quatre situations dans lesquelles le point est utilisé ailleurs qu'en fin de segment textuel. Le point peut être utilisé à l'intérieur d'un texte pour signifier un prix (branche 1). Ici, les « points de conduite » ont une signification sémantique : « le prix du pain (est de) 4 F. »

⁴ *Spirale* est une revue semestrielle de recherche en éducation qui paraît depuis 1988).

⁵ Logiciel développé au CAMS par l'équipe LaLIC (groupe CONTEXT) sous la direction de J.-P. Desclés sur l'extraction des connaissances : résumé automatique, extraction de citations, extraction de concepts et de relations statiques, extraction de relations causales, relations sémantiques entre texte : image, tableau et encadré.

On remarque d'ailleurs, dans le cas des sigles ou des chiffres (ex. S.N.C.F. ou 24.11.98) la non-présence d'un espace après le point. L'absence d'espace après le point permet dans ces cas, de lever l'ambiguïté par rapport à la fin ou à la non-fin du segment textuel ⁶.



La branche 5 est explicitée dans la figure 2, qui montre (par une lecture descendante) comment, historiquement, le point a été associé à d'autres signes graphiques pour donner nos signes actuels de ponctuation.

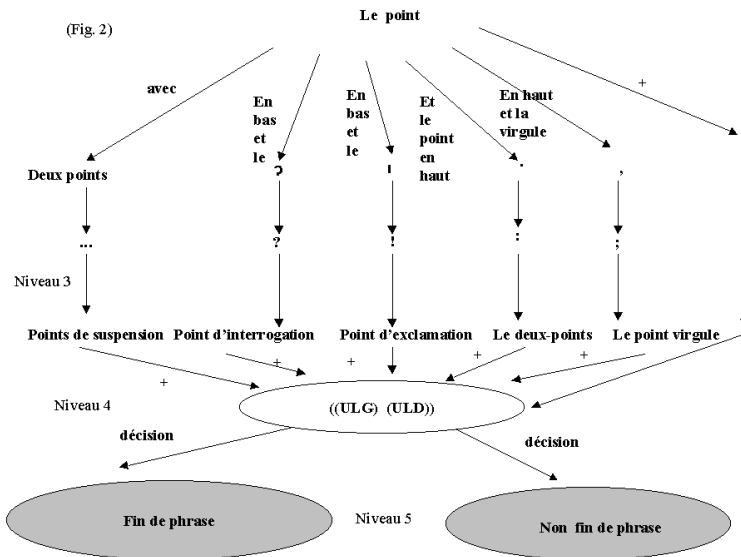
Les nouveaux 3, 4, et 5 de celle-ci montrent notre démarche pour effectuer la prise de décision de fin ou de non fin d'un segment textuel. ((ULG) (ULD)) constituent les unités lexicales (token) gauches et droites de chaque signe.

4. Adaptation de la méthode d'Exploration Contextuelle

La méthode d'exploration contextuelle (EC) est une procédure, fondée sur des règles heuristiques qui permettent de prendre la décision relative à la tâche que l'on veut résoudre. Cette question a été initialement soulevée par J.-P. DESCLES. La conception de cette méthode sur le langage a donné lieu à diverses publications [DESC93, 95], [MAIR90], ainsi qu'à plusieurs applications informatiques : identification des valeurs aspectuelles d'une proposition (SECAT) [DESC93], filtrage automatique de phrases importantes dans un texte en vue d'un résumé automatique

⁶ La distinction établie entre le point abrégatif et le point final dépend de l'application informatique, et non du niveau de l'interprétation linguistique.

(système SAPHIR) [BERR96], modélisation des connaissances par analyse des marqueurs linguistiques de relation entre concepts (SEEK) [JOU193], acquisition de connaissances causales à partir de textes (COATIS) [GARC98].



La méthode d'EC consiste à identifier en premier lieu dans un texte les unités linguistiques appelées indicateurs qui sont nécessaires au déclenchement des règles d'exploration contextuelle. Ces règles doivent identifier dans le contexte gauche et/ou droite des indices linguistiques qui orientent vers une prise de la décision adéquate.

Dans le système de segmentation de textes, l'application de la méthode d'EC a été réalisée à partir des marqueurs typographiques qui jouent un rôle important dans le déclenchement des règles de segmentation ; ces règles font appel à un examen du contexte gauche et/ou droite, ce qui permet de lever les ambiguïtés.

Une règle d'EC pour la segmentation de textes a la forme générale suivante :

Soit un marqueur pivot X

SI le contexte gauche de X est G

ET SI le contexte droite de X est D

ALORS prendre la décision Y (fin ou non fin d'un segment)

Exemple d'une règle :

SI l'on rencontre dans un texte un point *PT*

ET SI *PT* est suivi d'un espace *BL*

ET SI *BL* est suivi d'un guillemet *GI*

ET SI *GI est suivie d'un BL*
 ET SI *BL est suivie d'une majuscule*
 ALORS *PT (le dernier de (trois points) est la fin d'un segment textuel*
 (insérer la marque d'une fin d'un segment après le guillemet).

Exemple :

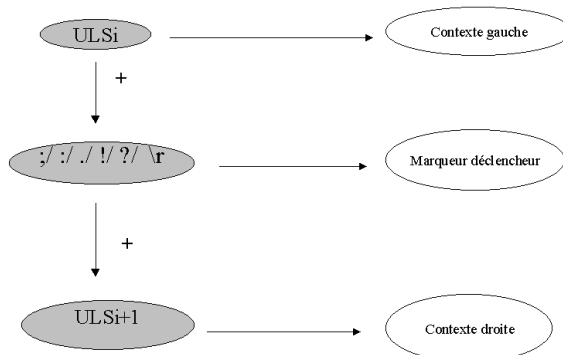
<a> « (...) Ce n'est donc pas l'idée d'une action commune de la France et de l'Allemagne qu'il faut remettre en cause, mais son mode opératoire (à partir de la monnaie) et son contenu social et politique... » <a> La pensée de Jean-Pierre Chevènement, loin d'être seulement critique, est foncièrement positive et débouche sur un projet pour l'Europe, reposant sur une coopération entre les peuples.
 (CD ROM - Le Monde Diplomatique Avril 1996, page 18 ; 19).

<a> désignent respectivement le début et la fin d'un segment textuel.

L'étude des corpus cités ci-dessus nous a permis d'identifier toutes les unités lexicales permettant de segmenter les textes. Ces unités sont :

<L'espace BL> <Le point PT> <Le deux-points DPT> <Le point virgule PTV> <Le point d'interrogation PINT> <Le point d'exclamation PTEX>
 <Parenthèse ouvrante PO> <Parenthèse fermante PF> <Crochet ouvrant CO> <Crochet fermant CF> <Tiret TIRET> <Guillemets GI> <Retour à la ligne \r> <Début de ligne \n> <Tabulation \t> <Chiffres Arabes CA>
 <Chiffres Romains CR> <Lettres majuscules LM> <Lettres minuscules LM>
 <Un mot qui commence par une majuscule MMaj> <Un mot qui commence par une minuscule MMin> <Liste des particules d'interjection (Ah, oh, Ouf...) LPI> <Liste des abréviations toujours suivies d'un chiffre ou d'une majuscule (p., pp., MM., ...) LA> <Liste des indices pour la segmentation dans le cas de deux-points (http, « , ...) LIDPT> <Liste des indices qui sont suivis par une lettre majuscule (vitamine C, Hépatite A,...) LILM> <Texte TXT>.

(Fig. 3)

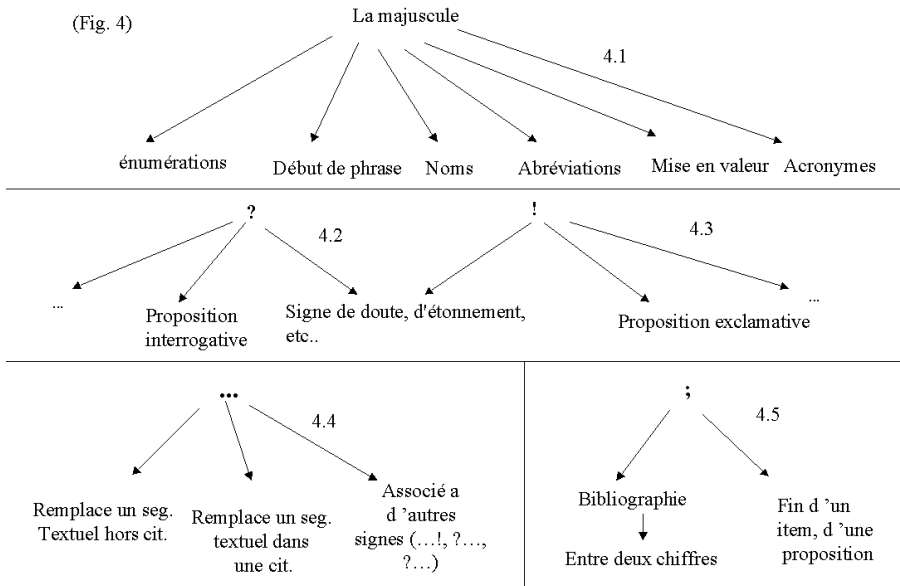


L'application des règles d'EC à celles-ci permet de repérer non seulement les fins de segments textuels, mais également leur début.

On considère que les déclencheurs qui déterminent le début et la fin d'un segment sont : ;/ :/ ./ !/ ?/ //r

Nous nous intéresserons donc aux contextes gauches et aux contextes droits de chacun de ces déclencheurs. La figure 3 schématise cette démarche.

Les figures 4.2, 4.3, 4.4 et 4.5 montrent que ces déclencheurs ne sont pas seulement utilisés en fin de segment textuel.



Il existe d'ailleurs des ambiguïtés au sujet des points qui délimitent une unité sémantique.

En effet, prenons les exemples :

(2) «*Eh bien ! Bravo ! Tu fais là un bel exploit !* »

(3) «*Selon les organisateurs du défilé, il y avait au moins 50 000 manifestants (?) au Champs-de-Mars.* »

(4) «*Voulez-vous ouvrir votre coffre ?, dit le douanier.*»

(5) «*Veux-tu savoir si je suis un humaniste ? oui, je le crois.*»

Ainsi nous remarquons que dans l'exemple (2) (corpus *Spirale*) le point d'exclamation exprime l'ironie ; dans l'exemple (3) (cité par J.-P. Collignon) le point d'interrogation exprime le doute ; dans l'exemple (4) le point d'interrogation est suivi d'une incise qui détermine l'énonciateur ; enfin dans l'exemple (5) (cité par J. Drillon [DRIL91]) le point d'interrogation précède la question de l'énonciateur.

Dans tous ces cas on considère que ces points ne sont pas des séparateurs des segments textuels. De même, les ambiguïtés concernant les points de suspension sont très nombreuses. Considérons les exemples suivants :

(6) «BRUNER J. (1990), ...car la culture donne forme à l'esprit, Paris, ESHEL (trad. 1991). »

(7) «Washington s'inquiète et réclame l'éviction de trois officiers – proches du président - impliqués dans des atteintes aux droits de l'homme en... 1991. »

(8) «Certains préféraient se laisser mourir en arrivant... » «Crimes contre l'humanité restés impunis... »

(9) «La lecture d'articles des années 80 montre à quel point ils ont pu rivaliser dans le choix et le nombre des qualifiants : fonctions "énonciative", "distanciante", "expressive", "rhétorique" etc... (nous en avons relevé une trentaine !...). »

Dans l'exemple (6) (corpus *Spirale*) les points de suspension se trouvent en début de segment textuel ; dans l'exemple (7) au milieu d'un segment sans qu'ils ne déterminent pour autant une fin de phrase ; dans l'exemple (8) on les retrouve à la fin d'un segment dans lequel le dernier point des trois points de suspension remplace le point final (on remarque que le deuxième segment après celle-ci commence par un guillemet) ; et dans l'exemple (9) ils sont précédés d'un "etc" qui n'est pas la fin d'un segment premièrement et deuxièmement associés à d'autres points (ex. : (7), (8) et (9) corpus *Le Monde*).

Selon les livres de grammaire et les traités de ponctuation, le point-virgule termine une proposition grammaticalement complète mais sémantiquement liée à la phrase principale (or on considère que cette proposition est un segment textuel important, mais mal considéré en linguistique). Ainsi le point-virgule peut également, dans les textes journalistiques surtout, se trouver dans les bibliographies pour séparer les noms des auteurs ou les numéros des pages, comme le montre l'exemple (10).

(10) «*Le Monde diplomatique*, juillet 1996, page 6 ; 7 ; 8. »

Il nous faut également aborder le cas des points suivis de deux parenthèses, dans lequel le contenu des parenthèses est sémantiquement lié au segment précédent.

(11) « *Le modèle est généralement illustré par un texte adéquat alors que parfois, quelques pages plus loin, on peut trouver des textes qui attestent d'autres systèmes !!! (19). »*

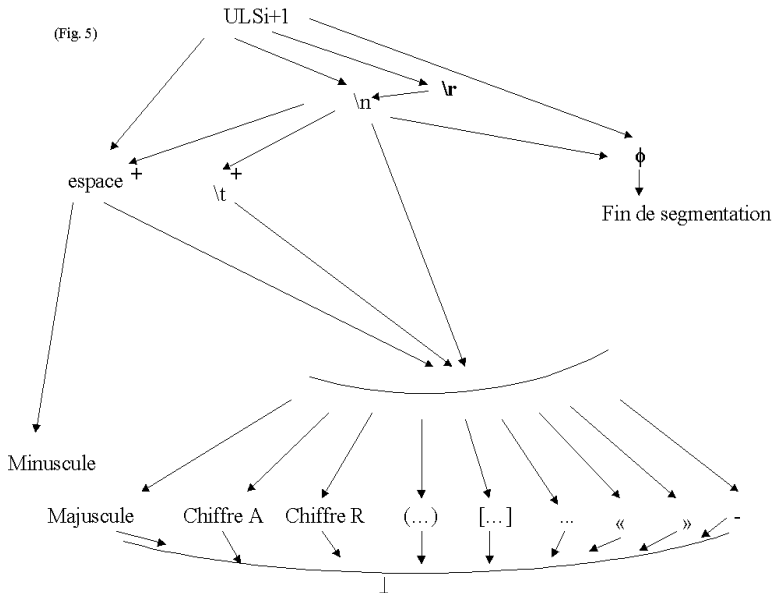
Le chiffre 19 entre parenthèses est ici lié à la phrase qu'il précède et renvoie à une note dont le contenu est sémantiquement lié à cette phrase. Dans les textes actuels, ces types de renvoi de notes sont liés par des liens hypertextes. Dans [MOU99], on considère que la couleur bleue qui est utilisée par défaut dans le système de navigation sous Internet (*Internet Explorer*, *Netscape*) est un nouveau signe typographique.

La figure 4.1 schématise l'utilisation d'une majuscule dans des textes. Il montre que la règle selon laquelle la majuscule permet de repérer le début d'une phrase est insuffisante. En effet, le début de segment textuel ne se limite pas à la présence d'une majuscule, mais peut être commencé par n'importe quel signe, comme le

montre la figure 5. Cette figure désigne les cas dans lesquels on trouve les unités lexicales droites (contexte droit) :

ULSi+1 (unités lexicales droites) désigne les cas « possibles » qui peuvent suivre un marqueur (la flèche marque la relation PEUT-ÊTRE- SUIVI-PAR).

Le "+" signifie UNE-OU-PLUSIEURS-FOIS



Le contexte gauche ULSi est schématisé par la figure 6.

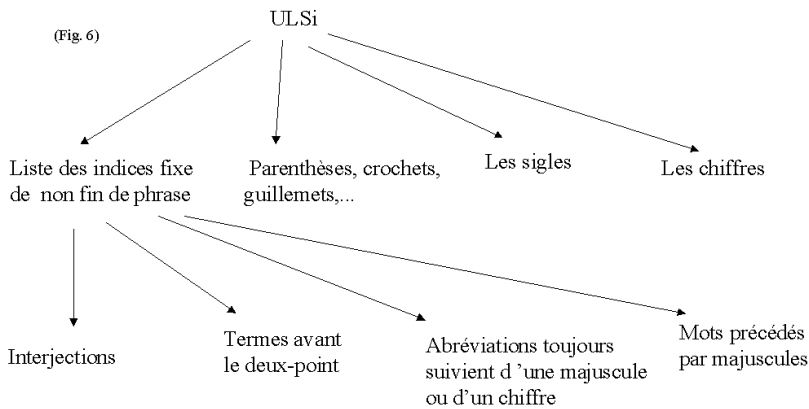
5. Conception de règles

Dans l'étude de notre corpus nous avons isolé 50 cas pour développer notre outil. Nous avons conçu des règles permettant au segmenteur de repérer d'abord les non-fin de segments textuels, ensuite les fins de segments textuels. Nous allons en présenter quelques-unes ci-dessous de façon non formelle.

On considère qu'un point suivi d'un espace suivi d'une parenthèse suivie d'un mot en minuscule n'est pas la fin d'un segment textuel. Car ce qui est contenu dans la parenthèse est lié sémantiquement à ce segment. Il peut contenir une précision, une référence bibliographique ou un commentaire (cf. ex. 9).

On considère aussi qu'un point suivi d'un espace suivi d'une parenthèse ouvrante suivie d'un chiffre suivi d'une parenthèse fermante suivie d'un point n'est pas la fin du segment textuel, car le chiffre dans ce cas renvoie à une précision sémantique par rapport à ce qui précède (cf. ex. 11).

Lorsqu'on rencontre trois points suivis d'un chiffre arabe suivi d'un point, le dernier des trois points n'est pas la fin d'un segment.



Le cas où un point d'exclamation est suivi d'un espace puis d'une majuscule ou d'un retour à la ligne est une fin de segment, sauf lorsque le point d'exclamation est précédé de particules interjectives (exemple : hélas, ah, oh, hé bien,...). Car l'interjection porte aussi en réalité sur le segment suivant. Un exemple spécifique d'exploration contextuelle de ce cas est étudié ci-après.

Exemple sur les particules d'interjections

(12) « *Ouf, j'ai pris mon cachet* »

« pour déterminer les valeurs sémantiques des temps (SECAT). Le passé composé **J'ai pris** est en soi indéterminé, il ne permet de caractériser complètement la valeur aspectuelle de la proposition. **Ouf** est un indice contextuel qui contribue à lever l'indétermination sémantique et qui oriente vers la valeur "d'état résultant." » [DESC96].

Ce exemple pouvait être écrit de plusieurs manières :

Ouf ! **J**'ai pris mon cachet

ou

Ouf ! **j**'ai pris mon cachet

ou

Ouf, j'ai pris mon cachet !

Si on segmente le « texte » qui contient des propositions similaires sans prendre en compte les interjections en début de segment, on se trouve dans le cas suivant :

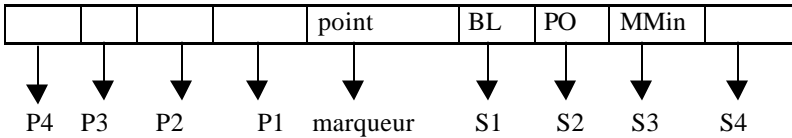
<a> Ouf ! <a> J'ai pris mon cachet

Il serait plus judicieux de considérer cet énoncé comme un seul segment textuel.

REGLES DE NON FIN DE SEGMENTS :

Les abréviations utilisées dans les exemples suivants sont mentionnées ci-dessus :

Prenons l'exemple suivant : *La lecture d'articles des années 80 montre à quel point ils ont pu rivaliser dans le choix et le nombre des qualificatifs : fonctions "énonciative", "distantiante", "expressive", "rhétorique" etc... (nous en avons relevé une trentaine !...).*

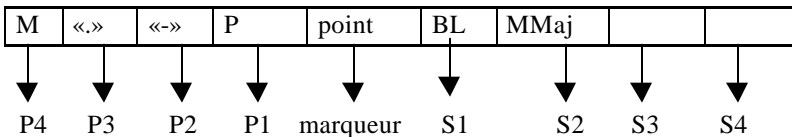


Le troisième point après «etc» n'est pas la fin d'un phrase.

Un point suivi d'un espace et d'une parenthèse ouvrante et d'une lettre minuscule n'est pas la fin fin d'une phrase

Règles pour les noms propres composés

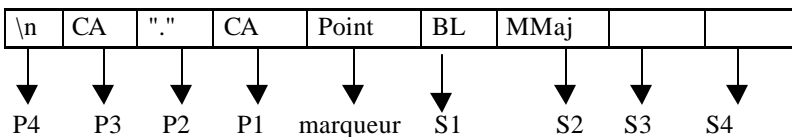
Si on rencontre un point précédé d'une lettre majuscule et suivi d'un Blanc (espace entre les mots) et d'une majuscule : le point n'est pas une fin de phrase, ex.: M.-P. Dupont.



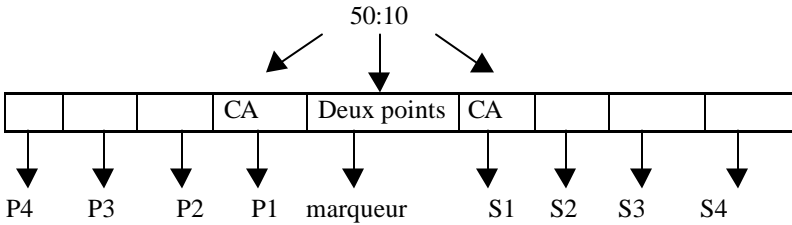
Règles pour les chiffres arabes en début de ligne

ex. : « 1.2. C'est ainsi que, pour couper à toute critique, M. Alain Minc déclare: "Ce n'est pas la pensée, c'est la réalité qui est unique." Il n'y a donc plus même à penser: le réel suffit. Le fait et la valeur ne font plus qu'un. »

Début de ligne suivi d'un chiffre arabe, suivi du marqueur, suivi d'un blanc suivi d'un mot en majuscule n'est pas la fin d'une phrase (on considère que ces chiffres portent des informations utiles pour un traitement automatique est surtout pour le balisage des textes en paragraphes).



Ex.: deux point en tant que signe de division



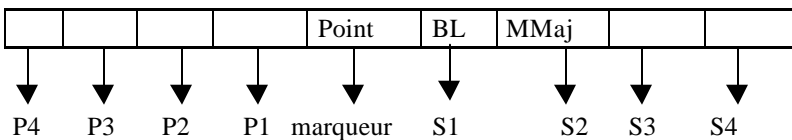
REGLES DE FIN DE SEGMENTS :

Le texte (en général) commence par une majuscule qui détermine le début d'un segment i.

Pour trouver la fin d'un segment i (sauf le dernier) il faut connaître le début du segment i+1

SI l'on rencontre dans un TXT un PT
 ET SI PT est suivi d'un espace BL
 ET SI BL est suivi d'une majuscule MMaj
 ALORS PT est la fin du segment

La première règle la plus simple après le lever d'ambiguïté :



6. Évaluation

Nous avons réalisé plusieurs segmenteurs adaptés à plusieurs tâches par la même méthode. Pour le résumé automatique et le filtrage des textes, les marqueurs-pivot sont « . » / « ; » / « : » / « ! » / « ? » / « \r ». Pour le système SEEK et l'étude de la causalité, les marqueurs sont « . » / « ; » / « ! » / « ? » / « \r ». Le deux-points est ici un indice contextuel complémentaire pour définir une relation de dépendance de causalité.

Deux types de segmenteurs sont réalisés pour la segmentation à l'intérieur des guillemets : l'un segmente, l'autre pas. Concernant celui qui segmente à l'intérieur des guillemets, il subsiste néanmoins une contrainte : chaque guillemet ouvrant doit être refermé. Car le code informatique (format texte .txt) de ces deux guillemets est le même. Ce problème est posé dans le cas des imbrications de citations, où la fin de la deuxième citation constitue la fin des citations. En d'autres termes le nombre des

guillemets doit être un nombre paire, où chaque guillemet ouvrant doit être suivi d'un guillemet fermant.

Le corpus du Monde diplomatique⁷ est segmenté sans échec (pour n'est pas dire à 10%) par notre système. Pour le corpus *Spirale*, aucun problème ne se pose, sauf quand le segmenteur ne doit pas segmenter à l'intérieur des guillemets (citations longues).

Certains problèmes n'ont pas été rencontrés dans le corpus, mais existent néanmoins. Ce sont les cas des segments terminés par un sigle et qui commencent par un nom propre et le cas d'un segment terminé par une lettre majuscule. Ces problèmes ne sont pas fréquents car la tendance actuelle est à l'utilisation des acronymes à la place des sigles. Un raffinement de ces deux problèmes a été effectué par la constitution : premièrement d'une liste de mots qui peuvent être suivis d'une seule lettre, deuxièmement, d'un fichier comprenant des débuts de phrases fréquents en français : la majuscule des articles, démonstratifs, pronoms personnels, quantificateurs, etc. La première liste a été étudiée par M. Silberztein et reprise par A. Dister. Ces lettres sont toujours précédées d'un autre mot, car les sigles en une lettre sont très vite épuisés. Calvet dans son étude de corpus n'en a trouvé que 0,9% [CALV80].

Le fichier d'entrée du segmenteur est un fichier en format texte, les fichiers des sortie sont en format texte (.txt) et HTML.

7. Perspective

Le segmenteur est aisément adaptable à plusieurs langues. Il est actuellement adapté à l'espagnol, et en phase d'adaptation à la langue russe, à l'anglais et au bulgare. La méthode utilisée peut en effet être transposée à toutes ces langues européennes.

Mais, pour appréhender un traitement automatique de langues de famille différente, telle que l'arabe et surtout une segmentation de textes, il faudrait se baser sur l'étude de la virgule associé à la particule « wa » (le « wa » en arabe est une unité de langue qui prend différentes acceptions notamment la coordination, la concomitance, le serment, etc. ; pour plus de détails, voir le traité de grammaire de Ibn Hishâm, Moughni Al Labîb), ainsi que sur une étude syntaxique qui doit définir la fin au la non fin d'une phrase.

Déjà dans le texte coranique il y avait des marques pour indiquer une pause pour la récitation du Coran. Dans les versions du Coran du moyen âge la fin d'un verset était marquée par les signes \circ , \odot , Φ . De même, on trouvait ces signes dans différents manuscrits désignant la fin d'une phrase comme dans le manuscrit de Al Thaalibi, 10^{ème} s., *Fiqh al Lougha (Al Hayat*, 4 février 1999 page 16).

Les marques de ponctuation utilisées aujourd'hui sont le résultat de la Renaissance Arabe au 19^{ème} s. grâce à l'imprimerie, à la confrontation avec l'Europe et au mouvement de traduction d'œuvres étrangères. Les marques de ponctuation sont

⁷ Le segmenteur à été appliqué sur l'ensemble des textes sans aucun «nettoyage», comprenant les titres, les références, les adresses des auteurs, les dates, etc.

celles du système d'écriture européen, mais n'ont pas pour autant les mêmes valeurs, en particulier la virgule qui n'a pas toujours la fonction de coordination. D'autre part, le point en arabe n'est souvent utilisé que pour marquer la fin d'un paragraphe, alors que la virgule est utilisée pour déterminer la fin d'une phrase. En revanche les autres signes de ponctuation tels que les guillemets, les parenthèses, les points d'exclamation et d'interrogation, les trois points, etc. ont la même valeur que ceux des langues européennes.

A notre connaissance, il n'y a que trois études traitant de la ponctuation en arabe ; un livre de Ahmad Zaky (1912), un autre de Abed Al Rauf Albasri (1932) et un livre récent de ISAAK F. [ISAA96] qui selon nous n'est pas approprié à la langue arabe par le fait qu'il y applique le système de ponctuation européenne.

8. Conclusion

En traitement automatique des langues, les segments textuels ne correspondent pas toujours à la définition classique de phrase, ni à la définition des signes de ponctuation telle qu'on la trouve dans les traités. Une segmentation fiable (balisage ou niveau de phrase ou de paragraphe) doit prendre en compte tous les marqueurs typographiques, et que le point suivi d'une majuscule ne suffit pas pour détecter la fin ou le début d'un segment.

Le travail de segmentation, comme beaucoup d'autres problèmes de traitement automatique, ne peut atteindre 100% même si, dans notre corpus, on est parvenu à ce taux. Par exemple, les corpus traitant de sujets mathématiques où l'utilisation des inconnues X, Y, etc. est très courante, entraînent des taux d'échec plus considérables.

La segmentation de texte doit être adaptée au traitement qu'on doit effectuer. Elle dépend premièrement du corpus ; deuxièmement de la tâche qu'on applique à ce corpus.

9. Références

[BERR96] RERRI J., Contribution à la méthode d'exploration contextuelle, Applications au résumé automatique et aux représentation temporelles, réalisation informatique du système SERAPHIN, Thèse de doctorat, Paris-Sorbonne, mai 1996.

[CALV80] CALVET J.-L., Que sais-je ? n° 1811, Les sigles, 1980.

[CATA77] CATACH N., (sous la direction de), Recherches historiques et actuelles sur la ponctuation, Actes de la table ronde internationale CNRS, Paris, Publications CNRS-HESO, 1977 ; 1979.

[CATA94] CATACH N., La ponctuation, Presse Universitaire de France, Paris, 1994.

[DEFA97] DEFAYS J.-M et al. « À qui appartient la ponctuation ? », Actes du colloque international et interdisciplinaire de Liège (13 - 15 mars 1997), in Champs linguistique, Paris, Bruxelles, pp. 437-447, 1998.

[DESC93] DESCLÉS J.-P., C. JOUIS., L'exploration contextuelle : une méthode linguistique et informatique pour l'analyse informatique de textes. In *ILN' 93*, 339-335, 1993.

- [DESC95] DESCLÉS J.-P., BERRI J., Le ROUX D. MALRIEU D., MINEL J.-L., «Le résumé automatique par exploration contextuelle », Recueil des communication effectuées aux rencontres Cognisciences Est le 25 novembre 1994, Rapport interne 95/1 du CAMS-LaLIC 1995.
- [DESC96] DESCLÉS J.-P., table ronde sur le contexte, Avril 1996, Caen.
- [DRIL91] DRILLON J., Traité de la ponctuation française, Paris, Gallimard, 1991.
- [FUCH93] FUCHS C., Linguistique et Traitement automatique des langues. Hachette, Paris, 1993.
- [GARC98] GARCIA D., Analyse automatique des textes pour l'organisation causal des actions, système COATIS, thèse de doctorat, Paris-Sorbonne, mai 1998.
- [IMPR90] Imprimerie Nationale. Lexique des règles typographiques en usage à l'imprimerie Nationale. Imprimerie Nationale, Paris, 1990.
- [ISAA96] ISAAK F., ISAAK S., «*La grammaire de Ibn ISAAK*», Dar Al Saki, Beirut, Liban, 1996.
- [JEFF97] JEFFREY C. and al., «*A Maximum Entorpy Approach to Identifying Sentence Boundaries* ». In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, D.C., March 31 - April 3, 1997.
- [JOU193] JOUIS C., Contributions à la conceptualiation et à la modélisation des connaissances à partir d'une analyse linguistique de textes : réalisation d'un prototype, Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales, Paris. 1993.
- [MAIR91] MAIRE-REPPERT D., Les temps de l'indicatif du français en vue d'un traitement informatique: Imparfait, Thèse de doctorat, paris Sorbonne novembre 1991.
- [MOUR98] MOURAD G., Exposé au sèminaire LaLIC, Université de Paris-Sorbonne, ISHA, Novembre 1998.
- [MOUR99] MOURAD G., Journée d'études franco-quebecoise (LaLIC, LANCI, IDIST) autour des plats formes CONTEXT et CONTERM, ISHA 1999.
- [PALM94] PALMER D. and HEARST A., «*Adaptative sentence boundary disambiguation.*» In proceeding of the 1994 Conference on Applied Natural Language Processing, Stuttgart, Germany, October 1994.
- [SILB93] SILBERZTEIN M, Dictionnaires électroniques et analyse automatique de textes, Le système INTEX, Paris, Masson, 1993.
- [VÉDÉ89] VÉDÉNINA L, Pertinence linguistique de la présentation typographique, Paris 1989.