DETECTION AUTOMATIQUE DE FRONTIERES DES PHRASES - UN SYSTEME ADAPTATIF MULTI-LANGUES

Anna Pappa,
Doctorante - Informatique
ap@ai.univ-paris8.fr , + 33 1 49 40 64 12

Gilles Bernard,
Professeur en Sciences de Informatique
gb@ai.univ-paris8.fr + 33 1 49 40 64 12

Hind Oukerradi,
Doctorante - Informatique
hind@ai.univ-paris8.fr +33 1 49 40 64 12

Adresse professionnelle
Laboratoire Intelligence Artificielle, Groupe CSAR, Université Paris 8 ★ 2, rue de la liberté ★ 93526 St Denis Cedex

Résumé : Cet article présente un système robuste adaptatif multi-langues, qui reconnaît les phrases dans un texte, en balisant les phrases et les signes de ponctuation. Il a été testé avec succès sur différents corpus de langues naturelles différentes (comme le français, le grec et l'arabe) avec un taux de réussite qui dépasse 99%. Le système est basé sur un algorithme de segmentation sous forme de règles. Ces règles sont issues des statistiques, elles-mêmes basées sur l'analyse distributionnelle [Harris, 1954]. Le traitement ne nécessite aucune connaissance préalable ni aucun dictionnaire. Sa portabilité et son adaptabilité le rendent précieux pour toute application ou analyse sur le langage naturel telles que MT, IE, TA, DS[1].

Summary : This article presents a robust system of phrase punctuation recognition. It is an adaptive multilingual one. Its accuracy exceeds 99% in different linguistic corpora (French, greek and Arabic). The system is a rule-based segmentation algorithm. The rules are statistics issued themselves based on distributional analysis [Harris, 1954]. No previous knowledge or dictionary are previously. System's portability as well as adaptability are precious for any further application or analysis like machine translation, information extraction, text alignment and document summarization.

Mots clés : TALN[2], ponctuation, règles, corpus multi-langues.

[1] (MT) Machine Translation, (IE) Information Extraction, (TA) Text Alignment, (DS) Document Summarization

[2] Traitement Automatique du Langage Naturel

Détection automatique de frontières des phrases -

Un système adaptatif multi-langues

Cet article présente un système robuste adaptatif multi-langues, qui reconnaît les phrases dans un texte, en balisant les phrases et les signes de ponctuation. Il a été testé avec succès sur différents corpus de langues naturelles différentes (comme le français, le grec et l'arabe) avec un taux de réussite qui dépasse 99%. Le système est basé sur un algorithme de segmentation sous forme de règles. Ces règles sont issues des statistiques, elles-mêmes basées sur l'analyse distributionnelle [Harris, 1954]. Le traitement ne nécessite aucune connaissance préalable ni aucun dictionnaire. Sa portabilité et son adaptabilité le rendent précieux pour toute application ou analyse sur le langage naturel telles que MT, IE, TA, DS[3]..

1. INTRODUCTION

La grande majorité des applications de traitements du langage naturel présuppose le découpage de textes en phrases. Cette tâche est depuis très longtemps automatisée, on parle alors de reconnaissance de frontières des phrases[4]. La phrase est considérée comme l'unité centrale des processus du traitement du langage naturel, comme par exemple l'étiquetage. On reconnaît comme phrase la suite des mots qui se trouvent entre des signes de ponctuation dits majeurs tels que le point, le point d'exclamation, le point d'interrogation et d'autres qui précédent ou suivent ces signes.

Formellement la ponctuation désigne l'ensemble des signes qui permettent l'interprétation des textes écrits. Tradition-nellement la ponctuation est le module de la compétence langagière le plus difficile à maîtriser, d'une part à cause de son caractère écrit, d'autre part à cause de différents styles appliqués par les auteurs dans la littérature. Les marqueurs de ponctuation jouent un rôle important pour indiquer les relations structurelles dans le discours (écrit), comme par exemple les relations rhétoriques particulières qui résident dans un texte [Dale, 1991]. Avant de procéder à la construction de systèmes de génération ou d'analyse automatique des documents écrits, il faut s'assurer que ces systèmes incorporent un modèle adéquat pour ces aspects textuels.

Des études linguistiques sérieuses sur le rôle de la ponctuation ont commencé récemment. [Nunberg, 1990] présente le début d'une théorie linguistique sur la ponctuation où les signes de ponctuation sont des indicateurs de surface pour plusieurs catégories syntaxiques.

Le problème de la détection automatique des phrases se pose à cause de l'ambiguïté de certains signes de ponctuation. Par exemple un point peut être utilisé pour déclarer la fin d'une phrase mais aussi pour exprimer une abréviation ou un acronyme (ex. : E.D.F. ou U.S.A.), ou même un nombre décimal, par exemple : 3.14 (écriture anglosaxone).

Les phrases[5] qui suivent montrent l'ambiguïté des signes de ponctuation comme le point :

a) Là où cela n'engage à rien, M.Filoche est pour les 37,5 annuités.

b) Les pays membres de l'U.E. seront confrontés à une reforme concernant leur régime de retraites.

c) Ce qui figure dans l'avant-projet de loi Fillon est considéré comme entièrement maintenu… et donc avalisé par la CFDT.

[3] (MT) Machine Translation, (IE) Information Extraction, (TA) Text Alignment, (DS) Document Summarization

[4] Dans la bibliographie anglosaxone la phrase est sentence et les propositions (principales et secondaires) sont des clauses.

[5] Ces exemples sont extraits du corpus traité composé des journaux quotidiens, hebdomadaires ou mensuels ainsi que des textes littéraires.

d) Alors T., mère de deux enfants, ouvrit sa porte.

e) « Est-ce qu'il sera là ? » demanda-t-il à son tour.

f) (…) appuyés par des hélicoptères, qui, - une fois n'est pas coutume ! – n'ont tiré ni missiles, ni roquettes, ni balles, (…)

Cette ambiguïté varie selon le type de texte ou de corpus. Des statistiques effectuées à partir d'un corpus composé d'articles du *Wall Street Journal* montrent que plus de 47% des points utilisés sont des points qui se trouvent dans les abréviations, contrairement à 10% pour le corpus *Brown* [Church, 1991]. Ceci démontre que si l'on coupe un texte en phrases sans tenir compte des particularités de l'ambiguïté on n'aurait que 53% des phrases correctement découpées pour le premier et 90% pour le second. Nous pouvons deviner que la plus fréquente source d'ambiguïté est le point d'une abréviation., ex. : Chap. 1.

2. TRAVAUX EFFECTUES

Les travaux concernant la détection automatique des phrases représentent un premier pas pour une analyse morphologique ou syntaxique et ne font pas l'objet d'une recherche. [Amman et al. 1997] ont regroupé les approches utilisées pour la ponctuation. Les techniques utilisées (souvent des listes grammaticales et de longues listes d'abréviations) visent à reconnaître les cas les plus courants. Dans leur majorité ces techniques sont orientées vers la détection des phrases dans de corpus ou dans une langue de type donné, ce qui rend difficile de les adapter à un autre type de textes ou à une autre langue sans avoir à modifier l'algorithme. De plus, puisque la détection de phrases est juste une première étape dans le traitement automatique du langage naturel, elle ne doit pas demander trop de ressources ni de calcul au niveau exécution. Donc, le développement d'un système basé sur de grandes listes d'abréviations, ou de lexiques spécialisés avec des informations qui ne s'utiliseront plus par la suite, n'est pas la meilleure solution.

Plusieurs travaux mentionnent l'utilisation d'un détecteur de phrases, mais aucune information sur sa conception ou sur sa performance n'est donnée [Voutilainen, 1993]. L'approche la plus simple consiste à l'usage des règles qui reconnaissent des suites de caractères (par exemple : point – espace – lettre majuscule), accompagnées de longues listes d'exceptions pour les abréviations. Une différente approche a été proposée par [Muller, 1980] : au lieu des listes d'abréviations, une analyse morphologique filtre les mots ayant les mêmes suffixes, pour lesquels la probabilité d'être une abréviation est minime ou nulle. De telles approches demandent plusieurs heures de travail pour construire et renouveler les listes de règles et d'abréviations. De plus, elles sont orientées à un type particulier de corpus.

L'approche avec le plus grand taux de détection [Riley, 1985] est basée sur un corpus annoté de 25 millions de mots. Sa performance pour le corpus Brown atteint 99,8%. Pour chaque mot du dictionnaire, ce modèle demande le calcul de probabilités qu'il soit le premier ou le dernier mot d'une phrase. Cette information est inutile aux analyses qui peuvent suivre morphologiques ou syntaxiques.

De plus la détection n'est qu'un moyen pour permettre à divers outils d'effectuer leur analyse des phrases découpées, donc le coût de calcul doit être minime. C'est pour cette raison que [Reynar et al. 1997] proposent un modèle basé sur la plus grande entropie qui ne nécessite aucun calcul ni le coût d'une information compliquée. L'information utilisée par ce modèle est fondée sur le segment (*token*) qui contient le signe de la ponctuation candidat pour la fin d'une phrase, ainsi que les autres *tokens,* juste avant et juste après celui-là. Pour la construction d'une liste d'abréviations extraite du corpus annoté, un algorithme simple est utilisé. L'approche de la plus grande entropie atteint 97,7% de solution pour le corpus *Brown*, avec utilisation d'une liste manuellement construite d'abréviations, d'appellations et d'acronymes. Si on enlève cette information supplémentaire, le taux est de 97,5%.

Une autre approche est le système *SATZ* [Palmer et al, 1997], qui utilise un réseau neuronal pour la désambiguïsation des frontières d'une phrase. Il est basé sur des probabilités d'appartenance d'un mot à une partie du discours principale (*prior part-of-speech*). Par exemple, le mot *portes* peut être un verbe mais il est plus probable qu'il soit un nom selon les textes utilisés. Ce système utilise un dictionnaire de 30000 entrées lexicales et

une information sur 6 différents *tokens* concernant leur contexte: 3 *tokens* avant et 3 après la fin candidate, et sa performance est de 98,5%, sur un corpus composé d'articles du *Wall Street Journal*. Le système SATZ peut être utilisé pour d'autres types de corpus ou d'autres langues naturelles après un apprentissage. Le problème demeure du fait qu'il existe des types de corpus, ou même des langues, pour lesquelles des dictionnaires spécifiques contenant des informations sur les principales parties du discours ou sur la façon de les détecter automatiquement n'existent pas. Les auteurs insistent sur le fait que la performance du système n'est pas influencée par la diminution de la taille du dictionnaire. Même si cette information est vraie, et peut être utile par la suite à un autre traitement du langage, il y a de traitements pour lesquels elle ne s'y applique pas, comme l'alignement des phrases [Santos, 2001] (*sentence alignement*).

Il faut préciser que le système SATZ ainsi que l'approche de l'entropie maximale, ne distinguent pas les différents usages des signes de ponctuation qui montrent la fin d'une phrase, qu'et ils utilisent les mêmes règles partout. Or, il y a des ambiguïtés qui ne sont pas traitées. Par exemple, un point peut être utilisé pour une abréviation mais ce n'est pas le cas d'un point d'interrogation ou d'un point d'exclamation. Pour la langue grecque nous pouvons citer le travail de [Stamatatos, 2000], qui décrit un système de découpage en phrases basé sur la technique de *transformation-based learning* introduite par [Brill, 1993]. Cette théorie, dont les applications comme par exemple le *text chuncking* [Ramshaw, 1995] sont largement répandues dans le domaine du TALN, consiste à extraire automatiquement de la connaissance linguistique sous forme de règles à partir des corpus annotés.

3. METHODOLOGIE

Notre méthode est fondée sur :

- des paramètres simples comme la lettre finale d'un mot ou sa longueur dont le coût de calcul est minime,

- des règles de désambiguïsation pour les signes de ponctuation. Ces règles sont extraites des informations tirées des statistiques sur des corpus non annotés et sont applicables pour les trois langues testées (français, grec, arabe).

Avant de procéder à la détection des frontières des phrases d'un texte (pour toutes les langues testées), nous définissons un ensemble de segments (*tokens*), ex. : caractères, nombres, signes de ponctuation. Pour chaque langue, ces signes sont différents, à part quelques signes comme le point final. Le point d'exclamation, qui sont communs aux trois langues.

La procédure de la segmentation suit la règle suivante : deux segments sont séparés par des espaces. Un segment qui est suivi par un des signes de ponctuation décrits ci-dessous, est considéré comme fin probable de la phrase (en respectant la forme particulière de chaque langue).

- Point .
- Point d'exclamation !
- Point d'interrogation ?
- Points de suspension …

Une autre frontière probable est quand les signes mentionnés ci-dessus se trouvent juste avant ou après des suites de caractères doubles comme les parenthèses (), les crochets {}, les guillemets, comme dans l'exemple : « super! ». Le tableau qui suit montre le découpage effectué dans une partie du corpus.

Les tableaux contiennent des extraits des corpus (français, arabe, grec), les phrases sont découpées et balisées ; les balises <s> <\s> désignent la phrase (sentence), et les balises <po> <\po> la ponctuation de la fin de phrase.

1	<s>Ainsi avons-nous analysé la Théorie physique <po>.</po> </s>
2	<s>La première tache question que nous rencontrions est celle-ci<po>:</po> </s>
3	<s>quel est l'objet d'une théorie physique <po> ?</po> </s>

Tableau 1 : Extrait du corpus français:

Les corpus sont composés de textes des styles divers: journaux, magazines, littérature, etc. et ils sont non annotés. Aucune connaissance préalable ou dictionnaire ne sont utilisés pour notre algorithme.

Il faut préciser que le corpus français atteint les 100 millions des mots, les corpus grec et arabe, qui sont moins volumineux néanmoins pas moins significatifs, atteignent progressivement le million.

3.1. Particularités de langues naturelles

En effet, il n'y a pas de différences quant à l'utilisation de la ponctuation dans les langues étudiées au sein de notre recherche. Les différences se résument à la graphie et aux particularités morphologiques que nous décrivons brièvement. Les particularités morphologiques de la langue arabe ont été prises en compte - en traitant les textes codés en UTF-8, afin de finaliser l'algorithme de la détection de phrases et désambiguïser l'utilisation de signes de ponctuation.

Le mot arabe graphique (il existe aussi la notion du mot arabe phonique mais ce n'est pas le cas de notre recherche), est facile à identifier: c'est ce qui s'écrit en un seul bloc entre deux blancs [Kouloughli, 1994]. En arabe, on ne distingue pas minuscules ou majuscules, mais il existe néanmoins trois formes graphiques: initiale, médiale et finale [Siamak, 2001]. Ceci est intéressant pour la reconnaissance des acronymes car nous pouvons les identifier suivant les mêmes règles que pour les langues occidentales: "Majuscule - Point - Majuscule - Point" etc. Mais il y a encore des problèmes d'ambiguïté qui persistent, car l'unicode ne permet pas de distinction (au niveau programmation) des trois formes mentionnées plus haut. Mais dans le corpus traité nous n'avons pas trouvé d'acronymes non identifiables.

1	<s> فالذين يتهموننا بالامبريالية هم جاهلون او سيئو النية [1] <po>.</po> </s>
2	<s> بالطبع، فقد جعل النبيذ اوضاء مدينة الوانال البقارة يسني ما اقطع من المكسيك نالكينيمودلاو يتياه على هيمنة والهم ابوك ديقيتو مضو اوغاراكين حايتجاو ايبمولوك من امنب عازتناو الفيليبين <po>...</po> </s>
3	<s> عام 1823، وفي رسالة موجهة الى الكونغرس اطلق الرئيس الاميركي جيمس مونرو العقيدة التي ال تزال تحمل اسمه <po>.</po> </s>

Tableau 2 : Extrait du corpus arabe.

Il faut juste mentionner que pour le grec moderne le signe de ponctuation point virgule (;) correspond au point d'interrogation et signifie donc la fin d'une phrase (voir exemple n° 2 dans le tableau 3 ci-dessous). Il faut aussi mentionner que le double point « : », étant donné qu'il peut signifier la fin d'une proposition, est marqueur de fin de phrase. Ceci est le cas pour plusieurs langues comme le français, l'italien, l'espagnol.

1	<s>Η φύση του νου σας, είναι στην πραγματικότητα καθαρή<po>.</po> </s>
2	<s>Η αντανάκλαση εμφανίζεται ακόμα κι αν ο ήλιος ή το φεγγάρι δεν έχουν αυτή την πρόθεση <po>:</po> </s>
3	<s>Πώς μπορούν τ' αποτυπώματα, τα οποία δεν έχουν μορφή να καθορίσουν κάτι φυσικό όπως τα σώματά μας<po>;</po> </s>

Tableau 3 : Extrait du corpus grec.

3.2. Les paramètres et les règles

Nous allons décrire le choix des paramètres pour la reconnaissance des fins des phrases en prenant comme modèle la langue grecque, sachant que les informations équivalentes ont été extraites des autres langues testées telles le français et l'arabe (les particularités de chaque langue faisant partie d'un module spécifique où ,selon le codage, le module s'active ou non). La procédure est empirique et exploite les caractéristiques du grec moderne [Chatzivasiliou, 1995]. Les informations extraites des statistiques qui nous ont permis de fabriquer les règles de reconnaissance sont résumées ci-dessous :

a. La longueur (en caractères) du dernier mot avant la fin de la phrase a peu de probabilité d'être petite (1 ou 2 caractères).

b. La grande majorité des mots en grec moderne ont comme caractère final le ς ou le ν, ou les voyelles.

c. Les signes de ponctuation qui se trouvent à coté d'une possible fin de phrase sont significatifs quant à leur utilisation.

	Type de signe de ponctuation	Exemples
	Initial	(, [, {, «
	Final	),], }, »
	Aucun	

	Type de caractère	Exemples
français	fréquent minuscule	e, s, t, n, a, etc.
	très peu fréquent minuscule	w, k, j, etc.
	fréquent majuscule	I, II, V, etc.
	très peu fréquent majuscule	W, K, J, etc.
grec	fréquent minuscule	α, ε, η, etc.
	très peu fréquent minuscule	μ, κ, λ, etc.

fréquent majuscule	A, E, H, etc.
très peu fréquent majuscule	M, K, Λ, etc.
nombre	0, 1, 2, etc.
spécial	%, #, $, etc.

Tableau 5 : Informations statistiques sur les caractères finaux en français et en grec ;

Les informations extraites des statistiques effectuées sur les corpus à partir des segments (comme l'exemple donné en figure 1) composent les paramètres suivants :

SG : la longueur (en caractères), le type du caractère, contient-il un point.

SD : la longueur (en caractères), le type du caractère, contient-il un point.

SPF : type de signe de ponctuation qui se trouve en position finale (à la fin d'un mot).

SPI : type de signe de ponctuation qui se trouve en position initiale (au début d'un mot).

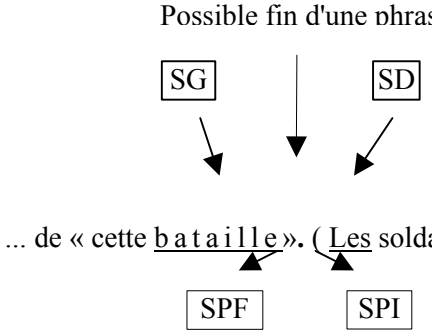

Figure 1 : Exemple des segments utilisés

Le tableau qui suit donne les informations extraites de l'exemple donné en figure 1.

SG	longueur : 8 type du premier caractère : minuscule, peu probable pour se trouver comme caractère final d'un mot type du dernier caractère : minuscule, final très probable contient-il un point : non
SD	longueur : 3 type du premier caractère : majuscule, probable final type du dernier caractère : minuscule probable final contient-il un point : non
SPF	type de signe de ponctuation : final - couple peut encadrer un paragraphe.
SPI	type de signe de ponctuation : initial - couple.

Tableau 6 : Explicatif des segments utilisés.

Le processus de la désambiguïsation des signes de ponctuation, ainsi que l'extraction automatique de règles que nous décrivons, est une variante de la théorie des règles de transformation. La différence consiste dans le fait que nous utilisons de corpus non annoté et que nous avons adopté le même algorithme à trois langues différentes.

Selon notre méthode, la reconnaissance des frontières d'une phrase peut se diviser en trois parties : dans la première nous effectuons une série de statistiques pour évaluer le contexte des signes de ponctuation. L'analyse distributionnelle nous permet de dégager des règles qui évaluent ensuite le contexte d'un signe particulier. Dans un premier temps nous considérons que tous les signes de ponctuation qui peuvent être signes de la fin d'une phrase indiquent une fin de phrase. La précision atteinte est la plus basse de notre système.

En deuxième partie nous appliquons des règles de type 1 :

SI contexte

ALORS éloigner signe de fin de phrase

Le contexte décrit la condition qui active la règle qui se compose soit d'une combinaison entre SG et SD soit entre SPF et SPI. L'exemple de la figure 1 devient :

SG	longueur : *8* type du premier caractère : minuscule, peu fréquent comme caractère finale d'un mot type du dernier caractère : minuscule, final très probable contient-il un point : non
	ET

SD	longueur : *3*
	type du premier caractère : majuscule, probable final
	type du dernier caractère : minuscule probable final
	contient-il un point : non
	OU
SPF	type de signe de ponctuation : final
	ET
SPI	type de signe de ponctuation : initial

Tableau 7 : Exemple du contexte

Une fois les règles de type 1 appliquées, le système procède à l'application des règles de type 2 (troisième partie) :

SI contexte

ALORS introduire signe de fin de phrase

Le contexte décrit la condition qui active la règle comme elle est définie plus haut. La procédure de règles peut se résumer comme le montre la figure 2 ci-dessous : tout d'abord nous appliquons les règles qui transforment un signe de fin de phrase en un simple signe de ponctuation, puis toutes les règles qui transforment un simple signe de ponctuation en un signe de fin de phrase.

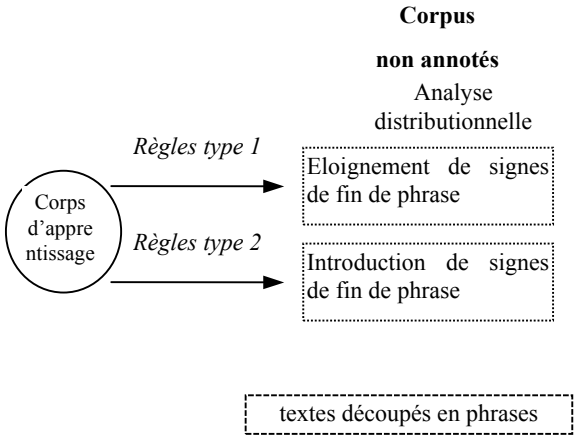

Figure 2 : Procédure de désambiguïsation

Pour chaque combinaison entre contexte et signe possible comme fin de phrase, les informations suivantes (clauses Prolog) sont extraites du corps d'apprentissage

Le SIGNE DE PONCTUATION est un des quatre possible comme fin de phase (point, point d'interrogation, point d'exclamation et points de suspension).

Le CONTEXTE est soit une combinaison entre SG - SD soit entre SPF - SPI.

Le C1 est un entier qui indique le nombre de fois qu'un signe de ponctuation particulier *n'est pas fin de phrase*.

Le C2 est un entier qui indique le nombre de fois qu'un signe de ponctuation particulier *est fin de phrase*.

Ensuite si l'on veut inclure une de ces clauses dans les règles de type 1 il faut respecter les contraintes suivantes :

$$C1 > C2$$

ET

$$C2 < SCFP * VERITE$$

Le *SCFP* représente la totalité de Signes Candidats pour Fin de Phrase, et *vérité* est un nombre entre 0 et 1 qui indique le degré de crédibilité des règles générées. Plus la vérité est grande, plus il y a des règles générées (moins de précision). Et plus la vérité est petite, plus les règles sont précises. Les tests effectués montrent une vérité de l'ordre de 0,01 (ou 99%) de crédibilité. Les contraintes pour les règles de type 2 sont respectivement :

$$C1 = 0$$

ET

$$C2 > 0$$

La définition des contraintes a été basée sur les statistiques. Dans un premier temps les contraintes pour les deux types de règles étaient symétriques mais l'analyse a montré que les contraintes de type 2 devaient être augmentées afin d'atteindre plus de précision.

Le système de découpage en phrases produit deux types d'erreur :

erreur positive : quand un signe de ponctuation est considéré comme fin de phrase, mais c'est faux, et

erreur négative : quand un signe de ponctuation qui est fin de phrase n'est pas reconnu par le système.

Le tableau qui suit regroupe quelques résultats pour les trois langues. La première colonne indique le choix des langues, la deuxième les 4 signes qui indiquent la fin d'une phrase (P pour le point, PI pour le point d'interrogation, PE pour point d'exclamation et PS pour points de suspension). La troisième colonne EP indique l'erreur positive, la suivante l'erreur négative (EN) et la dernière donne le taux de précision.

Langue	Signe	Phrases	EP	EN	Précision en %
	P	1516868	3189	858	99,7
français	PI	80847	1	642	99,2
	PE	41812	3	353	99,1
	PS	50009	1631	128	96,4
total		1689536	4824	1981	*99,5*
	P	56672	168	99	99,5
grec	PI	3028	0	29	99,1
	PE	1566	0	17	98,9
	PS	1873	63	6	96,3
total		63139	231	151	*99,4*
	P	32721	96	59	99,5
arabe	PI	1118	0	12	98,9
	PE	856	0	7	99,1
	PS	1055	32	2	96,7
total		35750	128	80	*99,4*
TOTAL		**1788425**	**5183**	**2212**	**99,5**

Tableau 8 : Résultats avec l'erreur.

4. CONCLUSION

La procédure de découpage en phrases constitue un pré-traitement pour toute analyse du langage naturel. Nous avons développé un système robuste qui reconnaît les frontières d'une phrase et qui s'adapte à différentes langues. Il a été testé avec succès sur des corpus des langues naturelles comme le français le grec et l'arabe. Nous n'utilisons aucune connaissance préalable (dictionnaire, liste d'abréviations, etc.), et les corpus ne sont pas annotés. Le taux de découpage sans faute atteint en moyenne pour les trois corpus 99,5%. La désambiguïsation des signes de ponctuation tels que le point, s'effectue grâce à un moteur de règles simples extraites des statistiques basées sur l'analyse distributionnelle des segments différents. Les balises produites par le système au cours du traitement offrent un outil précieux pour toute analyse syntaxique postérieure.

BIBLIOGRAPHIE

Akman V., Say B. (1997). *Current Approaches to Punctuation in Computational Linguistics.* In Computers and the Humanities. 30(6): 457-469.

Brill E. (1993). *Automatic grammar induction and parsing free text : A transformation-based approach.* In Proceedings of the DARPA Speech and Natural Language Workshop, pages 237-242.

Church K., Liberman M. (1991). *Rapport sur ACL/DCI*. In Proc. of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora, pages: 80 – 91.

Dale R. (1991). *The role of Punctuation in Discourse Structure.* In Working Notes of the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation, Asilomar, pages : 13-14.

Harris Z. (1954). *Distributional Structure*, in Word, pages. 146-162.

Kouloughli D.E. (1994). *Grammaire de l'arabe d'aujourd'hui.* Edition Pocket Collection Langues pour tous.

Muller H., Amerl V., Natalis G. (1980). Worterkennungsverfahren als Grundlage einer Universalmethode zur automatischen Segmentierung von Texten in Satze. In Sprache und Datenvarbeitung, 1. Ein Verfahren zur maschinellen Satzgrenzen - bestimmung im Englischen.

Nunberg G (1990). *The linguistics of Punctuation*. CSLI Lecture Notes 18, Stanford, CA.

Palmer D., Hearst M (1997). *Adaptive Multilingual Sentence Boundary Disambiguation*.

Computational Linguistics, pages: 241-267(2).

Ramshaw L.A. and Marcus M.P. (1995). *Text Chunking Using Transformation-Based Learning.* In Proceedings of the ACL third Workshop on Very Large Corpora. Cambridge, pages 82-94.

Reynar J. Ratnaparkhi A (1997). *A maximum Entropy Apppoch to Identifying Sentence Boundaries*. In Proc. of the fifth Applied Natural Language Processing Conference.

Riley M. (1989). *Some Applications of Tree-based Modeling, to Speech and Language Indexing*. In Proc. of the DARPA Speech and Natural Language Workshop, pages: 339-352.

Santos D. (2001). *Punctuation and Multilinguality : some reflections from a language engineering perspective*. www.oslo.sintef.no/portug/Diana/download/ponctuacao.ps

Siamak R (2001). Tokenizing an Arabic Script Language. In Arabic NLP Workshop at ACL/EACL, France.

Stamatatos E. (2000). Statistical Identification of Genre and Author in Urestricted Modern Greek Text. Ph.D Thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece.

Voutilainen A. (1993). *NPTool, a Detector of English Noun Phrases*. In proc.of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio State University, pages: 48-57.

Chatzivasiliou G. (1995). *Synoptiki Neoelliniki Grammatiki kai Syntaxi*[6]. éd. GRIGORIS, Athènes, Grèce.

[6] Traduit par : Grammaire et Syntaxe Concises du Grec Moderne.