**Reconnaissance automatique des parties du discours**

*Anna Pappa*

**Abstract**

Cet article présente un système informatique, combinant l'intelligence artificielle et la linguistique, capable de reconnaître les parties du discours et de faire émerger des structures du langage très variées sans aucune connaissance préalable. L'algorithme ne comporte pas de dictionnaire et il utilise un minimum de règles grammaticales et syntaxiques.

# 1 Introduction

Le système développé est basé sur des règles grammaticales [1] qui sont établies sans recours à l'étiquetage traditionnel [2]- même en utilisant des procédures automatiques [3]- ou à d'énormes bases de données d'entrées lexicales. Il procède à la détermination de l'étendue de différents groupes (ou syntagmes) syntaxiques qui composent une phrase : les syntagmes substantifs (SS), les syntagmes verbaux (SV), et les syntagmes prépositionnels (SP), qui se décomposent en une préposition suivie d'un SS ou SV. À la fin de l'analyse, le système crée un dictionnaire séparant les mots en deux catégories : noms et verbes. Le cadre de travail est énonciatif [4], structuraliste [5], fonctionnaliste [6] et prend en compte l'aspect dynamique du langage. Pour notre analyse, nous avons utilisé un corpus (environ 4.000.000 de mots) de textes français de différents auteurs sur des sujets et des styles différents.

# 2 Méthode suivie

La méthode que nous avons adoptée est basée sur l'étude des catégories grammaticales - syntaxiques sans avoir recours aux connaissances lexicales. Notre méthode fait partie des programmes qui utilisent :

- la division du système des règles en niveaux et,
- la méthode d'analyse morphologique (découpage et interprétation).

Les statistiques effectuées sur le corpus sont basées sur l'analyse distributionnelle [7]. Les régularités relevées forment les règles de notre système.

# 3 Système réalisé

Les textes sont découpés en phrases. Les mots sont classées dans les colonnes d'un tableau. Le **noyau** de ces phrases est la colonne qui comporte les mots dits grammaticaux (articles, pronoms, etc.), les colonnes qui précèdent le noyau contiennent les mots qui se trouvent juste avant les mots grammaticaux (MG)et constituent le contexte gauche (CG) et les colonnes qui suivent contiennent les mots qui se trouvent après les MG et forment le contexte droit (CD). Le tableau qui suit est un exemple, les phrases sont issues des textes différents :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *j'* | *ai* | un | *peu* | *de* | *fièvre* | *depuis* | *quelques* | *jours* |
| *et* | *voilà* | un | *jeune* | *homme* | *très* | *bien* | *fait* | *et* |
| *les* | *peuples* | des | *Nations* | *Unies* | *ont* | *proclamé* | *à* | *instaurer* |
| *et* | *qui* | la | *dévorait* | *des* | *yeux* | . | *CHAPITRE* | *HUITIEME* |

Le développement de cette démarche a été basé sur les marques grammaticales [8] ou marques de repérage. Les mots sont distinctement séparés en tenant compte des particularités de la langue française, par exemple les mots composés. Le système est un "parseur" basé sur des travaux statistiques [9] avec résolution des cas ambigüs[2], sans avoir étiqueté aucun mot auparavant, et il détermine les groupes syntaxiques sans recours aux grammaires des arbres [10] compactées ou non [11]. Nous citons quelques règles et nous décrivons leur fonctionnement:

| *N°* | *CG2* | *CG1* | Noyau | *CD1* | *CD2* | *CD3* | *CD4* | *CD5* | *CD6* | *Déc* | *Cas* | *Flag* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *1* | * | * | * | *=Verbal* | * | * | * | * | * | *SV* | *nég ou t.v* | *1* |
| *2* | * | * | * | * | *=De* | * | *=Relatif* | * | * | *CD4=m.a* | *Dét* | *4* |

La première ligne définit le rôle de chaque colonne. La première colonne indique le n° des règles[3]. Les deuxième et troisième colonnes contiennent le CG du noyau. La quatrième colonne contient les mots grammaticaux, ensuite viennent six colonnes[4] avec le CD du noyau. Les trois dernières colonnes indiquent la **décision**, où il est indiqué la marque d'arrêt (m.a) du syntagme, le **cas** où la décision se justifie, et le **flag** qui nous permettra de constituer notre dictionnaire à la fin du traitement. Nous présentons quelques exemples

2. Par exemple l'ambiguïté entre article défini et pronom personnel : le, la, les, l'.

3. L'étoile signifie n'importe quel mot, le "="renvoit automatiquement à une fonction traitée par le moteur(Parsing). Le moteur appelle la fonction la fonction (=Verbal) ou (=Relatif) et autres mentionnées dans le tableau de règles.

4. Les statistiques ont montré qu'il est vraiment rare de rencontrer un SS -beaucoup moins un SV- qui s'étend au-délà de 7 mots à son contexte droit après le déterminant.

extraits des textes traités :

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *385* | *ruisseaux* | *dont* | aucun | *ne* | *va* | *de* | *droit* | *fil* | , | *SV* | *sub* | *1* |
| *134* | *donc* | *comme* | une | *galerie* | *de* | *peintures* | *dont* | *les* | *traits* | *CD4=m.a* | *Relatif* | *4* |
| *171* | *roule* | *sur* | des | *choses* | *indifférentes* | . | *Après* | *dîner* | , | *CD3=m.a* | *Prép+SN* | *3* |

Notre analyse se termine avec *la création d'un dictionnaire* qui contient les mots qui se trouvent immédiatement à droite des mots grammaticaux : il y a des déterminants qui ne peuvent être suivis que par un nom (soit substantif soit épithète) dans la plupart des cas, par exemple les possessifs (sa maison). Pour ces cas où l'ambiguïté est peu probable (flag 1), nous avons développé une procédure qui enregistre les mots[5] de la colonne **CD1** dans un dictionnaire. Les mots sont étiquetés en deux catégories verbale ou nominale:

| **N°** | **CD1** | **Type** | **Texte d'origine** |
|---|---|---|---|
| 282 | abstinence | nominal | Paresse |
| 62 | armée | nominal | Luc |
| 68 | coin | nominal | Pin-Up |
| 254 | diable | nominal | Luc |
| 4627 | accablait | verbal | Les-Cenc |
| 4523 | accompagnai | verbal | domi |
| 11 | aller | verbal | Rêveries |
| 161 | apprirent | verbal | Luc |

# 4 Conclusion

Le système développé procède à une analyse des marques (mots grammaticaux) sans connaissance préalable sur leur nature et leur fonctionnement. Les taux de réussite sont répartis ainsi : 93% des cas résolus (reconnaissance des SS et SV précédés ou non d'une préposition), dont le taux d'erreur varie entre 0 et 1%, et 7% les cas non résolus. L'algorithme est utilisé également pour résourdre les cas ambigüs (entre article défini et pronom personnel). Cette analyse peut s'effectuer à d'autres langues qui ont la même structure syntaxique que le français (en cours d'étude la langue grecque). La création du dictionnaire à partir du contexte droit des mots grammaticaux peut s'avérer précieuse pour un traitement où le système chercherait des informations supplémentaires dans les entrées lexicales dont le type grammatical est déjà connu.

5. En leur forme fléchie

# Références


[1] *G. Sabah*, L'intelligence artificielle et le langage, processus de compréhension, éd. Hermès (1989) vol.2

[2] *J. Vergne, E. Giguet*, Regards théoriques sur le tagging, In proceedings of the fifth annual conference Le Traitement Automatique des Langues Naturelles (TALN 1998), Paris, France, June 10-12.

[3] *E. Brill*, Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging, To appear in Natural Language Processing Using Very Large Corpora. Kluwer Academic Press (1997)

[4] *A. Culioli*, Pour une linguistique de l'énonciation, Opérations et représentations, Orphys (1990) vol. 1

[5] *E. Benveniste*, Saussure après un demi-siècle, in Problèmes de Linguistique Générale, Paris, (1966) chap. III

[6] *J. Vachek*, Dictionnaire de linguistique de l'école de Prague, Anvers, Utrecht (1966)

[7] *Z.S. Harris*, Distributional Structure, Word (1954), p 146-162.

[8] *G. Bernard*, Typologie neuromimétique des substantifs, rapport de recherche 94-06-17-1, laboratoire d'Intelligence Artificielle de l'Université Paris 8 (1994)

[9] *E. Charniak*, Statistical parsing with a context-free grammar and word statistics, in Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press / MIT Press (1997)

[10] *M. Marcus, et al.*, The Penn Tree-Bank: Annotating predicate argument structure, in Proceedings of ARPA Speech and Natural Language Workshop (1994)

[11] *Y. Wilks, et al.*, Compacting the Penn Treebank Grammar, in Proceedings of the COLING-ACL (17th Int. Conf on Computational Linguistics), Montreal (1998)



*Anna Pappa*
Groupe CSAR - Laboratoire d'Intelligence Artificielle
Université Paris 8
2, rue de la liberté
93526 Saint Denis Cedex - France
ap@ai.univ-paris8.fr
http://www.ai.univ-paris8.fr/ ap