

Une application de la grammaire structurale: L'analyseur syntaxique du français SYGFRAN

Jacques Chauché¹

(1) LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
jacques.chauche@lirmm.fr

Résumé. La démonstration présentée produit une analyse syntaxique du français. Elle est écrite en SYGMART, fournie avec les actes, exécutable à l'adresse : <http://www.lirmm.fr/chauche/ExempleAnl.html> et téléchargeable à l'adresse : <http://www.sygtext.fr>.

Abstract. The software produces a syntactic analysis of french. It is written in SYGMART, including acts, runnable at <http://www.lirmm.fr/chauche/ExempleAnl.html> and downloadable at : <http://www.sygtext.fr>.

Mots-clés : Analyse syntaxique.

Keywords: syntactic analysis.

1 Modèle de traitement

Les algorithmes de Markov ont la puissance d'une machine de Turing. Ils peuvent donc représenter n'importe quel algorithme. Ces traitements se font à partir d'un traitement de chaîne par la substitution d'infices. La grammaire structurale transforme des structures arborescentes. Elle correspond donc à une extension des algorithmes de Markov appliquée aux structures arborescentes : au lieu de remplacer un infixe d'une chaîne on remplace une sous-structure d'une arborescence. La grammaire structurale permet de définir des récurrences ou des compositions d'applications. Le système SYGMART est un outil qui permet l'écriture et l'exécution de grammaires structurales. Ce système a comme entrée soit une structure déjà produite par ailleurs, soit un texte quelconque. Il produit soit un texte soit une structure. Par exemple une application récente concernait la manipulation de fichiers xml afin de modifier et/ou de composer leurs structures. D'autres applications textuelles permettent d'écrire un traducteur, un tagueur, un compresseur de texte, un classifieur,

2 Caractéristiques de SYGFRAN

L'analyse s'effectue en deux temps :

- une analyse morphologique produisant l'ensemble des solutions possibles pour chaque mot dans une arborescence simple. Cette arborescence comporte deux niveaux. Dans le premier niveau chaque point correspond à un mot du texte. Chaque point de ce niveau est la racine des différentes solutions pour le mot associé. Cette étape est basée sur un automate d'états finis.
- Une analyse syntaxique basée sur la grammaire structurale. La grammaire comprend environ 250 parties et 20000 règles.

Sur un PC ou un Mac comprenant au moins 2 GO de mémoire l'analyse supporte le traitement d'un texte représentant une arborescence syntaxique de 200000 points. L'analyse du livre "Le petit prince" de Saint Exupéry traite et produit une structure d'environ 35000 points et mets 3mn sur un ordinateur MacBookPro Intel core 2 Duo, 4GO de mémoire.