# Analyse informatique du roman proustien « *Du côté de chez Swann* »

*Katia ZELLAGUI*
Laboratoire le LASELDI – Université de Franche-Comté
30 Rue Mégevand - 25000 Besançon (France)
katia.zellagui@univ-fcomte.fr

## Mots-clefs – Keywords:
Automate fini, grammaire locale, dictionnaire électronique, étiquetage morpho-syntaxique, désambiguïsation, textes littéraires.
Finite state automata, local grammar, electronic dictionary, morpho-syntactic tagging, disambiguation, literary texts.

## Résumé - Abstract
Dans le cadre du développement des environnements d'analyse linguistique, d'étiquetage de corpus et d'analyse statistique afin de traiter des corpus de grande taille, nous proposons de mettre au point des procédures nouvelles d'étiquetage morpho-syntaxique et sémantique. Nous présentons un ensemble de ressources linguistiques - dictionnaires et grammaires - dans le but d'étiqueter entièrement le roman proustien : « *Du côté de chez Swann* ». Notre recherche avance deux atouts majeurs : la précision des étiquettes attribuées aux formes linguistiques du texte ; et le repérage et étiquetage exhaustifs des mots composés.
To deal with a great amount of corpus data within the framework of environmental development of linguistic analysis of corpus' tagging and statistic analysis, we propose to establish new procedures of syntactic and semantic tagging. We present some general linguistic resources, such as dictionary and grammar built-in the way to entirely tag the novel of Proust «*Du côté de chez Swann*». Our research leads to two main advantages: precise tagging assigned to linguistic forms of the text and identification and exhaustive tagging of compound nouns.

## Introduction
Face à l'expansion des ressources électroniques (e-book, sites Internet, CD-Rom), de plus en plus de données textuelles sont disponibles sur support électronique. Cette expansion implique inévitablement le développement des outils d'analyse de textes : que ce soit des analyseurs ou des logiciels d'étiquetage (e.g. Brill, Cordial, Lexico, Hyperbase, INTEX), et oblige les chercheurs en linguistique informatique à repenser la gestion et l'organisation des données textuelles (e.g. Text Encoding Initiative). Le problème des textes sur support électronique actuellement disponibles est qu'ils ne sont pas décrits suffisamment d'un point de vue lexical, syntaxique et sémantique. Dans le cadre de nos recherches, nous proposons de fournir aux chercheurs en littérature et en linguistique des procédés spécifiques d'étiquetage lexical et syntaxique de corpus littéraires, à partir desquels il sera possible de réaliser (à court terme) des analyses linguistiques et littéraires fines et de construire (à long terme) des hypertextes d'un nouveau type : les liens seront établis entre les unités linguistiques (plutôt qu'entre les formes superficielles).

## 1 Cadre théorique
Notre démarche s'inspire en grande partie des travaux du LADL et plus spécifiquement des travaux menés par Maurice Gross (1975) sur le lexique- grammaire.

L'équipe qui s'intéresse essentiellement au traitement automatique et à la description à large couverture des langues naturelles constitue depuis les années 60 une base de données de descriptions linguistiques très importante sous la forme principalement de dictionnaires électroniques - les DELA (Courtois, 1990) - et de grammaires locales (Gross, 1997).
Notre objectif est d'élaborer des ressources linguistiques afin de réaliser de façon automatique voire semi-automatique l'étiquetage du roman proustien « *Du côté de chez Swann* »[1]. L'étiquetage consiste à *« associer à des segments de texte, le plus souvent les 'mots', une ou plusieurs étiquettes, le plus souvent leur catégorie grammaticale voire leur lemme.* » (Habert & al., 1997). L'étiquetage que nous proposons de réaliser sera systématique (toutes les formes du texte doivent avoir au minimum une étiquette lexicale et syntaxique. Une telle entreprise rencontre deux obstacles majeurs : le choix des unités linguistiques à traiter et le choix des outils d'étiquetage. Or, une des conditions sine qua non d'un étiquetage visant un taux d'erreur proche de zéro réside dans le repérage exhaustif et l'étiquetage des mots composés dès les premières phases du traitement du corpus. Nous avons donc besoin d'un outil capable d'une part, de traiter les mots composés ; et d'autre part, d'un outil permettant de créer rapidement et facilement des règles de grammaires. Notre choix s'est donc porté sur le système INTEX (Silberztein, 1993). Cet outil utilise les vastes bases de données de descriptions linguistiques fines développées au LADL et permet de développer des ressources linguistiques (dictionnaires et grammaires locales), puis de les appliquer au corpus. Après avoir présenté les problématiques liées à une telle entreprise, nous tenterons de proposer des solutions en décrivant les ressources que nous avons créées afin d'étiqueter le corpus Swann.

# 2 Unités linguistiques et ambiguïtés

## 2.1 - Les unités linguistiques

Nous choisissons de traiter deux types d'unités linguistiques : les mots simples (ou formes simples) et les mots composés (ou formes composées).
En ce qui concerne la définition d'une forme simple, nous partons d'une définition formelle que nous empruntons à Courtois (1990). Les formes simples sont : « […] *des unités de texte définies sur l'alphabet des codes ASCII ou EBCDIC à 256 caractères, et ne comportant aucun séparateur*. » Les lettres (éléments de base) appartiennent à un alphabet déterminé (e.g. l'alphabet français comprend 41 lettres).
Il n'existe pas de définition précise et admise de la composition. Gross (1988) annonce clairement qu'il est inutile car impossible de proposer une définition unique et stable du mot composé. Il propose d'appréhender cette notion par le biais de contraintes linguistiques. Sa démarche consiste « *à montrer que le figement n'est pas une valeur absolue mais relève d'une gradation correspondante à des propriétés transformationnelles potentielles réalisées à des degrés différents* ». Il propose donc de calculer le degré de figement en terme de propriétés non observées (Gaston, 1986). Les listes des mots composés élaborées en grande partie à partir de ces contraintes par les chercheurs du LADL et du LLI[2] ont servi à l'élaboration du dictionnaires électroniques des mots composés : le DELAC.
La reconnaissance des mots composés (au même titre que les mots simples) est nécessaire pour deux raisons principales :

1. ils forment une unité linguistique à part entière, e.g. *parce que, pomme de terre* ;
2. ils peuvent générer des erreurs lors de l'étiquetage s'ils ne sont pas reconnus.

[1] Le texte sur support électronique (issu de la reconnaissance optique) a été fourni par les éditions Champions (Paris). Nous le nommons corpus Swann.
[2] LLI : Laboratoire de Linguistique Informatique, Paris 13.

## 2.2 Les ambiguïtés

La notion d'ambiguïté est fondamentale car elle pose des problèmes majeurs dans l'analyse de texte. Certaines unités de la langue (mots simples ou mots composés) sont effectivement ambiguës et possèdent plusieurs étiquettes morpho-lexicales et syntaxiques dans les ressources d'INTEX, ce qui nécessite un travail de désambiguïsation. Par exemple, la forme simple **la** possède trois entrées dans le DELAF : *la, la.N+z1:ms:mp : **la** est un nom ; la, le.DET+z1:fs : **la** est un déterminant; la, le.PRO+z1:3fs : **la** est un pronom*.
L'ambiguïté que nous venons de décrire relève de la syntaxe. Il existe effectivement six grands types d'ambiguïtés : les ambiguïtés orthographiques, morpho-flexionnelles, morpho-dérivationnelles, syntaxiques, sémantiques, et les ambiguïtés pragmatiques (Fuchs, 1996).
Dans le cadre de notre recherche, nous traiterons deux types d'ambiguïtés :

1. les ambiguïtés syntaxiques qui se situent au niveau de la structure des énoncés;
2. les ambiguïtés morpho-flexionnelles qui se situent au niveau de la flexion des formes verbales, e.g. *j'aime / il aime* (le verbe aimer est conjugué à la 1ère et 3ème personne du singulier du présent de l'indicatif ou du présent du subjonctif) ;

Une simple exploration des contextes gauche-droite de la forme suffit souvent à lever l'ambiguïté (les grammaires locales sont alors très efficaces). Mais certaines formes ne peuvent être désambiguïsées qu'en tenant compte du contexte syntaxique ou sémantique global, au niveau de la phrase ou même du discours (e.g. *table ronde*).

# 3 Les outils informatiques pour l'étiquetage

Nous allons à présent décrire les outils utilisés pour le traitement automatique et linguistique du corpus : les dictionnaires électroniques et les grammaires locales.
INTEX dispose de ressources lexicales. Pour le français, il s'agit des dictionnaires électroniques du LADL. Les deux dictionnaires principaux du système INTEX sont : le DELAF [dictionnaire des mots simples généré automatiquement à partir du DELAS (Courtois, 1990)] et le DELACF [dictionnaire des mots composés généré semi-automatiquement à partir du DELACF (Silberztein, 1989)]. Ces ressources (une fois appliquées au corpus) permettent d'obtenir le vocabulaire complet du texte. Les ambiguïtés sont alors repérables (i.e. une forme ambiguë possède plusieurs entrées dans le dictionnaire).
Les grammaires locales s'avèrent être des outils très efficaces afin de gérer les ambiguïtés non résolues par consultation des dictionnaires. Elles se présentent sous forme d'automates. Dans le processus d'étiquetage, nous utiliserons des grammaires locales de désambiguïsation sous forme de transducteurs. Les transducteurs diffèrent des automates dans la mesure où ils produisent en plus de l'information « *les transitions sont étiquetées par des couples de symbole (Sr/Sp), où Sr est un symbole reconnu, et Sp un symbole produit*. » (Silberztein, 1993).

# 4 L'étiquetage du texte

La phase étiquetage se déroule en trois étapes successives : le pré-traitement du texte brut, l'étiquetage des formes composées et l'étiquetage des formes simples.

## 4.1 Pré-traitement

Avant la phase d'étiquetage, il faut réaliser l'étape dite de pré-traitement. Sous INTEX, cette étape consiste d'une part à segmenter le texte en phrases, ce qui passe par l'application d'une grammaire locale qui traite les ambiguïtés liées au point (signe de ponctuation forte ou signe d'abréviation) et qui insère le symbole {S} après chaque fin de phrase ; et d'autre part, à reconnaître et étiqueter les mots composés non ambigus : e.g. *parce que, aujourd'hui* (par le biais du dictionnaire Ucompound.dic intégré au système INTEX).

Il s'agit ensuite d'appliquer les dictionnaires du système : le DELAF qui contient 746 214 entrées, et le DELACF qui compte 248 885 entrées. Une fois cette opération réalisée, INTEX fournit deux listes qui correspondent au vocabulaire du texte : la liste des mots simples (47 592 entrées), et la liste des mots composés (3 408 entrées). Il est alors possible de procéder à la phase d'étiquetage, ce qui revient à gérer les ambiguïtés du texte. L'étape de désambiguïsation se déroule en deux temps : le traitement des formes composées puis celui des formes simples.

## 4.2 Traitement des formes composées

Le DLC contient donc 3 408 entrées, ce qui représente environ 7 240 mots composés dans le texte. Le problème majeur du traitement des mots composés réside dans le fait que certaines séquences reconnues par une consultation de dictionnaire sont en fait des séquences libres de mots simples (ex. *bien que*). C'est pourquoi INTEX traite tous les mots composés comme a priori ambigus. La désambiguïsation des mots composés ne peut être réalisée que par une analyse manuelle du contexte de leurs occurrences. En effet, une séquence peut être figée dans un certain contexte, et ne pas l'être dans un autre (e.g. *Je me rappelle **bien que** je n'ai pas dormi ; cela lui ferait plus de **bien que** son lit ; […] il le disait **bien que** cela ne se fasse pas !*).
Nous utilisons un programme interactif de convivialité d'étiquetage : DIATAG[3] afin de procéder à la levée d'ambiguïté des mots composés en contexte. DIATAG (intégré au système INTEX) présente la forme avec son contexte (gauche et droite) et la ou les étiquettes candidates. À partir de l'observation directe, l'utilisateur peut alors valider la bonne étiquette qui sera intégrée directement dans le texte.
Les séquences non validées seront ensuite traitées comme des séquences de mots simples.
À l'issue de cette opération, le DLC ne contient plus que **3 382** entrées et le texte possède **6 580** mots composés.

## 4.3 Traitement des formes simples

Une fois les formes composées étiquetées, il s'agit alors de traiter les formes simples. Ce traitement est entrepris en deux phases successives :

1. le traitement du plus grand nombre d'ambiguïtés est réalisé avec INTEX. La bibliothèque de grammaires locales de levée d'ambiguïté contient deux types de grammaires : les grammaires locales dites générales applicables à n'importe quel texte de la langue française, et les grammaires locales dites ad hoc (i.e. grammaires spécifiques qui ne fonctionnent que sur le corpus *Swann)*. La bibliothèque (dans son état actuel) contient une cinquantaine de grammaires locales. Ces grammaires locales présentées sous forme de transducteurs fonctionnent par reconnaissance d'information et production d'information. Nous présentons deux grammaires de désambiguïsation. La grammaire locale *forme s.grf* (cf. Figure 1) fonctionne sur tous les textes de la langue française.

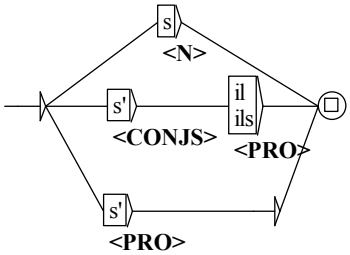


*Figure1 : Grammaire locale de désambiguïsation : forme s.grf.*

[3] http://www.nyu.edu/pages/linguistics/intex/#diatag

Le premier chemin analyse la forme *s* comme étant un nom (et impose la contrainte <N> ; la forme *s'* suivie des pronoms *il* et *ils*, est systématiquement une conjonction ; dans tous les autres cas, le *s'* est un pronom. 944 ambiguïtés sont résolues en appliquant cette grammaire.
La grammaire locale *je plus V.grf* (cf. Figure 2) appartient à l'ensemble des grammaires locales ad hoc. Elle permet de désambiguïser un grand nombre de pronoms (e.g. *le, la, nous, lui*) et de préciser la flexion du verbe.

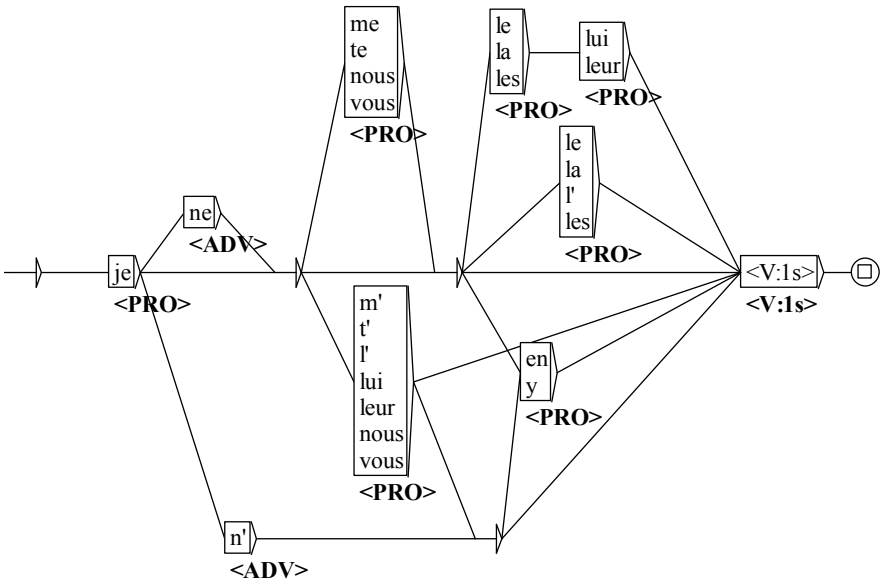


*Figure 2 : Grammaire locale de désambiguïsation : je plus verbe.grf.*

Le graphe *je plus verbe.grf* reconnaît 1 565 occurrences dans le corpus et permet de lever 1 378 ambiguïtés, par exemple : *{je,.PRO+PpvIL+z1} {me,me.PRO+z1} {suis,être.V+aux+z1:P1s} {couché,coucher.V+se+p+i+E+z1:Kms}*.
Les grammaires locales que nous avons construites ont permis de lever plus de 80% des ambiguïtés de façon automatique.

2. Traitement des ambiguïtés résiduelles avec DIATAG

Face à l'impossibilité de lever certaines ambiguïtés de façon automatique, nous entreprenons une partie de l'étiquetage des formes simples de façon manuelle à l'aide du programme DIATAG. La démarche est alors similaire à celle que nous avons suivie pour le traitement des formes composées. Prenons l'exemple de la forme **que** (qui possède cinq étiquettes syntaxiques : pronom relatif ou interrogatif, conjonction de subordination, introducteur ou adverbe). Compte tenu de la complexité de certaines phrases proustiennes, l'étiquetage de cette forme ne peut être réalisé qu'en se référant au contexte. Ce type d'ambiguïté est donc résolu au cas par cas avec DIATAG. Nous avons eu recours à ce procédé essentiellement afin de traiter les mots grammaticaux (Dister, 2001). À l'issue de ce traitement, le DLS ne contient plus que **16 372** entrées.

# 5 Résultats

Nous avons développé un environnement qui a permis de lever plus de 80% des ambiguïtés du texte de façon automatique. Ces ressources comprennent des dictionnaires électroniques et des grammaires locales de levées d'ambiguïté[4]. Le vocabulaire final du texte contient 164 060 mots

[4] Les grammaires locales de désambiguïsation sont disponibles sur le site INTEX.

simples et 6 580 mots composés. Afin de réduire les erreurs d'étiquetage, nous concevons une chaîne de traitements incluant des procédures de contrôle qualité (permettant de détecter le maximum d'erreurs). La qualité de l'étiquetage est ensuite évaluée en comparant le corpus étiqueté avec la version étiquetée proposée par FRANTEXT[5]. Notre objectif étant d'évaluer la qualité de l'étiquetage, nous précisons que nous comparons dans ce cas précis les résultats en ayant conscience du fait que les méthodes d'étiquetage sont différentes. Nous sélectionnons de façon aléatoire 10 paragraphes d'environ 200 mots chacun. Nous relevons dans la version FRANTEXT plusieurs erreurs dues essentiellement au fait que les mots composés n'aient pas été reconnus (*« cette/DTN:sg femme/SUB:sg que/SUB$ j'/PRV:sg avais/ACJ:sg quittée/VPAR:sg '/' il/PRV:sg y/PRV:++ avait/ACJ:sg quelques/DTN:pl moments/SUB :pl* ***à/PREP peine/SUB:sg*** *;/; »*[6]*)*
Un objectif futur serait d'appliquer ces ressources et notre méthodologie à un corpus plus conséquent ; nous pensons bien entendu à la totalité de *La Recherche*. En s'appuyant sur l'expérience que nous avons acquise au cours de ce projet, nous pensons qu'un travail d'étiquetage de l'ensemble de *La Recherche* pourrait être effectué par un chercheur en quelques mois.

# Références

COURTOIS B. (1990), « Un système de dictionnaires électroniques pour les mots simples du français », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n°87, Paris : Larousse, pp. 11-22.
DISTER A. (2001), « Levée d'ambiguïté sur les mots lexicaux et grammaticaux », fascicule spécial, In *Description et levée des ambiguïtés*, Éditions Éric Laporte, Linguisticae Investigationes, Amsterdam/Philadelphia, John Benjamins, pp. 105-126.
FUCHS C. (1996), *Les ambiguïtés du français*, Collection l'Essentiel Français.
GROSS G. (1988), « Degré de figement dans les noms composés », In *Les locutions figées*, Éditions Laurence Danlos, Langages, n° 90, Paris : Larousse, pp. 57-72.
GROSS G. (1990), « Définition des noms composés dans un lexique-grammaire », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n° 87, Paris: Larousse, pp. 84-90.
GROSS Gaston (1996), *Les expressions figées en français* : *noms composés et autres locutions*, Paris : Ophrys.
GROSS M. (1975), *Méthodes en syntaxe*. Paris : Hermann.
GROSS M. (1997), «The construction of local grammars », In *Finite-State Language Processing*, E. Roche and Y. Schabes (eds.), Cambridge, Mass/London, England:MIT Press, pp. 329-354.
GROSS M. (2001), « Les ambiguïtés », In *Description et levée des ambiguïtés*, Éditions Éric Laporte, Linguisticae Investigationes, vol. 24, fascicule 1, John Benjamins Publishing Company, pp. 3-41.
HABERT B., NAZARENKO A., SALEM A. (1997), *Les linguistiques de corpus*. Collection U Linguistique. Paris : Armand Colin/Masson.
SILBERZTEIN M. (1990), « Le dictionnaire électronique des noms composés », In *Dictionnaires électroniques du français*, Éditions Blandine Courtois et Max Silberztein, Langue Française, n° 87, Paris: Larousse, pp. 71-84.
SILBERZTEIN M. (1993), *Dictionnaires électroniques et analyses automatiques de textes* : *le système INTEX*, Paris : Masson.

[5] Version étiquetée par Josette Lecomte à l'INaLF – mars 2002.
[6] Nous avons souligné les erreurs d'étiquetage.