# Des patrons morpho-syntaxiques pour le repérage automatique de l'antonomase du nom propre

Sarah Leroy

Université de Bretagne Occidentale et U.M.R. C.N.R.S. 5475 – Praxiling – France

## Abstract

We set out here one side of the advantage that may represent a corpus automatic processing for the linguist. From a chain of rather simple processings on a press corpus, we manage to automatically locate proper names antonomasias by using the morphosyntactical tagging of the corpus and the methods of information retrieval.

## Résumé

Nous présentons ici un aspect de l'intérêt que peut avoir un traitement automatique de corpus pour le linguiste. À partir d'une chaîne de traitements assez simple, on parvient à un repérage automatique de l'antonomase du nom propre dans un corpus de presse, en s'appuyant sur l'étiquetage morpho-syntaxique du corpus et sur les méthodes de l'extraction d'information.

**Mots-clés :** Antonomase. Nom propre. Patrons morpho-syntaxiques. Extraction d'information.

## 1. L'antonomase par l'exemple ; les matériaux de l'analyse

Ce travail s'inscrit dans le cadre des travaux linguistiques[1] sur l'antonomase du nom propre, phénomène qu'on peut succinctement décrire en ces termes : « un nom propre en antonomase est un nom propre employé comme un nom commun »[2]. Le corpus d'étude de l'antonomase du nom propre utilisé dans les approches linguistiques est avant tout convoqué pour exemplifier l'analyse linguistique, au point que parfois l'illustration se substitue presque entièrement à la description et à l'analyse. Le but de ce travail est de préparer une démarche inverse d'analyse, basée sur l'observation de productions attestées sur corpus. Cette démarche rejoint celle des linguistiques de corpus[3], pour lesquelles la donnée linguistique correspond au corpus, qui contient l'observable, les productions « authentiques » qui s'y trouvent.

### *1.1. Pourquoi un traitement automatique de l'antonomase ?*

On peut se demander ce qu'un traitement automatique peut apporter à un travail sur l'antonomase. Nous distinguerons trois niveaux d'utilité.

Le premier niveau est celui de la constitution d'un corpus, dans la perspective non seulement de l'utilisation de données attestées, mais aussi d'une certaine représentativité de la langue, ou au moins d'un aspect, d'un niveau de la langue. La constitution manuelle d'un corpus d'antonomases est particulièrement fastidieuse ; elle peut également donner lieu à une sélection subjective, si elle n'est pas précédée de l'établissement d'une définition et de critères de distinction du fait recherché. Un repérage automatique doit donc, à terme, permettre

[1] Sans proposer une bibliographie exhaustive, on renverra à Jonasson (1994) et Gary-Prieur (1994) pour les travaux en linguistique du nom propre.
[2] Exemples : Ce jeune homme est un véritable Don Juan ; L'auteur, sorte de Boris Vian japonais…
[3] Voir Habert *et al.* (1997).

d'économiser du temps et des efforts ; il oblige, auparavant, à réfléchir sur les critères opératoires de reconnaissance de l'antonomase. Enfin, il permet de traiter plus de données ; on peut ainsi augmenter la représentativité du corpus.

À un deuxième niveau, un traitement automatique de l'antonomase permet d'envisager une quantification des différents aspects de ce fait langagier, à partir de laquelle on peut mesurer le rôle de chacun des éléments constituant le GN antonomasique et les corrélations qui s'établissent entre eux. On peut ainsi espérer donner la prototypicité de telle ou telle structure antonomasique et aboutir à une grammaire de l'antonomase.

Enfin, une telle réalisation peut, à son modeste niveau, contribuer à l'amélioration de la prise en compte par le traitement automatique des langues des emplois non-prototypiques du nom propre, jusqu'à présent assez négligés, comme le constatent Daille et Morin (2000 : 618) : « les variations morphosyntaxiques ou métaphoriques [du nom propre] n'ont jamais été considérées par les systèmes existants alors qu'elles sont très productives dans la langue ».

### *1.2. Quelles méthodes, quels corpus ?*

Les méthodes d'analyse retenues s'inspirent pour une part de celles des études de l'antonomase effectuées dans les approches linguistiques centrées sur les emplois modifiés du nom propre, mais également des traitements automatiques de corpus. La démarche adoptée s'inspire de celles des « linguistiques de corpus »[4], qui conjuguent « tradition anglo-saxonne de linguistique descriptive s'appuyant sur les corpus électroniques » et « traitement automatique du langage naturel » (Habert *et al.* 1997 : 8). Notre travail s'inscrit donc dans le cadre d'« une linguistique faisant appel […] à des corpus électroniques pour développer, à partir de " faits " rassemblés, des dictionnaires et des grammaires descriptives, mais aussi pour tester des hypothèses, confronter un modèle postulé aux réalisations effectives » (Habert *et al.* 1997 : 8).

Dans la perspective de ces linguistiques de corpus ou linguistiques « sur corpus » (Habert *et al.* 1997 : 9), le terme corpus désigne « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (Sinclair 1996 : 4) ou, plus précisément, « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue » (Habert 2000 : 13). Il s'agit, le plus souvent, de corpus annotés, c'est-à-dire « regroup[ant] sous un même chef, un même type, des réalisations distinctes d'un même phénomène, ses occurrences » (Habert *et al.* 1997 : 11), quel que soit le type de représentation[5] des données textuelles.

La représentation choisie ici est l'étiquetage morpho-syntaxique, traitement informatisé de corpus qui ajoute des informations aux données textuelles en leur associant un certain nombre d'« étiquettes ». Le programme spécifique, « étiqueteur » ou « tagger », qui produit l'étiquetage automatique effectue généralement une analyse morpho-syntaxique du texte qui lui est donné en entrée pour résoudre les ambiguïtés et affecter une étiquette à chaque élément du texte. Cette analyse morpho-syntaxique complète peut être affinée par une analyse partielle basée sur l'observation des contextes gauche et droit de l'élément traité.

[4] Voir Habert *et al.* (1997).

[5] « Étiquetage » morphosyntaxique (les constituants du corpus sont assortis d'étiquettes morphosyntaxiques), « arborescence » (les relations entre les constituants sont représentées par des arbres syntaxiques) ou « étiquetage » sémantique (le corpus est structuré sur la base de catégories lexicales ou conceptuelles). Chacun des ces types correspond à un niveau successif d'annotation.

Le corpus rassemblé pour observer en discours l'antonomase du nom propre est, en ce qui concerne le type de discours observé, relativement homogène, puisqu'il est exclusivement constitué d'articles de presse[6]. Il s'agit d'un corpus fermé[7] et spécialisé[8], qu'on divise en quatre sous-corpus respectivement nommés *Entraînement*, *Test*, *Portraits* et *Films*. Les deux premiers, *Entraînement* et *Test*, sont identiques du point de vue de leur contenu : ils rassemblent des articles[9] comportant tous au moins une occurrence d'antonomase du nom propre. Ils interviennent à différents moments de la partie « apprentissage » du traitement automatique de l'antonomase. Les deux autres sous-corpus, *Portraits* et *Films*, constituent un autre groupe de textes, qui eux sont utilisés dans la partie « application » du traitement automatique de l'antonomase. Chacun de ces sous-corpus est constitué d'une série d'articles[10] qui ne comportent pas forcément d'antonomase du nom propre et ont pour particularité de former des sous-types textuels[11] qui présentent des régularités formelles, structurelles et linguistiques, qui justifient leur exploitation dans un traitement automatique de l'antonomase.

## 2. Le traitement automatique de l'antonomase

Un traitement automatique s'incrit nécessairement dans un traitement automatique des noms propres. Cet aspect du traitement automatique des langues est désormais bien balisé ; il intervient par exemple dans « l'indexation, la recherche d'information ou la traduction » (Daille et Morin 2000 : 602), mais joue aussi un rôle capital pour la compréhension automatique, la communication homme-machine, et ce pour l'oral[12] comme pour l'écrit.

La définition du nom propre en traitement automatique des langues s'inspire de définitions linguistiques, mais aussi pragmatiques[13], avec en particulier l'introduction de la notion d'« entité nommée » qui « représente une catégorisation bien plus large que celle du nom propre [tel qu'il est abordé en linguistique], puisqu'elle inclut des expressions temporelles […] ou numériques […], des maladies et des drogues […] » (Daille et Morin 2000 : 606).

Le traitement automatique des noms propres s'articule autour de deux étapes principales : « l'identification des noms propres connus ou la découverte des nouveaux noms propres, et leur catégorisation » (Daille et Morin 2000 : 602). La première étape, l'étiquetage des noms propres, cruciale pour le traitement automatique de l'antonomase, peut être réalisée à l'aide d'« indices internes » au nom propre, tels que la majuscule, d'« indices externes [qui] proviennent du contexte dans lequel le nom propre apparaît » ou encore à l'aide de « liste[s] spécifiée[s] » (Bodenreider et Zweigenbaum 2000 : 730-731). La plupart des programmes d'étiquetage utilisent, outre le critère interne de la présence de la majuscule, la projection de dictionnaires de candidats-noms propres[14] et/ou l'analyse des contextes gauche et droit du

[6] 425 articles tirés de la presse quotidienne, nationale ou régionale, hebdomadaire ou mensuelle, d'où malgré tout une certaine hétérogénéité à l'intérieur même du type de discours.
[7] Selon la classification de Habert (2000 : 12-14) : « mis au point une fois pour toutes » pour les besoins d'une recherche ponctuelle, les corpus fermés s'opposent aux « réservoirs à corpus », tels que le BNC (*British National Corpus*) ou *Frantext*, qui consituent en fait des « *bases textuelles* », et aux « *corpus de suivi* » qui, « visant à capter en continu des données », « ne cesse[nt] de croître ».
[8] Les corpus spécialisés, « limités à une situation de communication ou à un domaine » (Habert *et al.* 1997 : 144), s'opposent aux « corpus de référence », conçus « pour fournir une information en profondeur sur une langue » (Sinclair 1996 : 10, cité par Habert *et al.* 1997 : 144).
[9] *Entraînement* compte 182 articles, *Test* en compte 48.
[10] 96 articles pour *Portraits*, 99 pour *Films*.
[11] *Portraits* est composé des portraits de dernière page de *Libération* ; *Films* rassemble des critiques de films du même journal.
[12] Voir Béchet et Yvon (2000).
[13] On retrouve là certains des « problèmes liés à la délimitation de la catégorie du Npr » évoqués par Jonasson (1994 : 13).
[14] La constitution de ces listes, ou acquisition lexicale, constitue une tâche spécifique.

candidat selon des patrons de fouille lexicaux, syntaxiques ou lexico-syntaxiques[15], utilisant des « mots-clés »[16] comme indices de la catégorie nom propre. La seconde étape, celle de la catégorisation, c'est-à-dire la classification des noms propres pour l'indexation, exploite ces différentes méthodes (analyse des contextes, utilisation de listes) pour mettre au point des taxinomies représentant la diversité des noms propres[17].

L'étiqueteur employé pour le traitement automatique de l'antonomase, *Cordial 6 "Universités"* pour *Windows*[18], utilise, pour le traitement et l'étiquetage des noms propres, un dictionnaire de noms propres de plus de 30 000 entrées ; de plus, un mot inconnu commençant par une majuscule est considéré comme un nom propre, ce qui permet en principe de traiter les noms propres inconnus. La combinaison de ces deux méthodes donne des résultats globalement satisfaisants : les erreurs sont peu fréquentes[19] et concernent principalement des noms propres ne figurant pas dans le dictionnaire des noms propres mais dont existe une forme homonyme[20]. Certaines erreurs classiques d'étiquetage du nom propre sont évitées : les gentilés ne sont généralement pas considérés comme des noms propres, pas plus que les titres en français. L'étiquetage des noms propres effectué par *Cordial* comporte également une caractérisation en genre et en nombre. Cette précision, si elle peut être utile par ailleurs, introduit une diversité qui n'est pas souhaitable pour un traitement automatique des noms propres et est éliminée lors de la préparation du corpus.

### *2.1. Enjeux et méthodes du traitement automatique de l'antonomase*

Le point de vue de certaines analyses linguistiques, qui considérent l'antonomase comme un emploi modifié du nom propre[21], met en lumière des caractéristiques syntaxiques qui permettent de considérer le nom propre en antonomase comme le nom-tête d'un GN antonomasique[22]. Du point de vue d'un traitement automatique, on gagne à considérer l'antonomase non pas du strict point de vue du nom propre, mais de celui du GN antonomasique. Ce choix[23] permet de délimiter de façon claire le phénomène. On définira donc ici l'antonomase du nom propre comme un GN ayant pour tête un nom propre, nom propre qui peut alors servir de repère pour la recherche de ces GN. La recherche d'une catégorie syntaxique et non d'une occurrence lexicale implique une vision plus générale des corpus, favorisée par l'étiquetage morpho-syntaxique.

Ces deux données (une donnée technique, l'existence d'un outil d'étiquetage globalement correct en ce qui concerne les noms propres et une donnée théorique, la conception de l'antonomase du nom propre comme un GN ayant pour tête un nom propre et présentant des caractéristiques syntaxiques précises), conduisent à s'appuyer sur des méthodes issues des techniques d'extraction d'information pour constituer un système de traitement automatique des GN antonomasiques.

---

[15] Voir, par exemple, Coates-Stephens (1993), Wolinski *et al.* (1995).
[16] Ou « mots déclencheurs (*trigger words*) » (Daille et Morin 2000 : 612).
[17] Daille et Morin (2000 : 603-606) présentent diverses classifications des noms propres et des entités nommées.
[18] Il s'agit d'un « correcteur global de la langue française » dont la version « universitaire », plus spécifiquement destinée à l'analyse de données textuelles, comporte une fonction d'étiquetage de texte.
[19] Les erreurs d'étiquetage de nom propre, c'est-à-dire l'affectation à un nom propre d'une autre étiquette que « nom propre », sont relativement rares ; l'erreur inverse, l'attribution de l'étiquette « nom propre » à un élément relevant d'une autre catégorie, semble plus fréquente.
[20] *Minou*, dans *Minou Drouet*, est étiqueté « nom commun » alors qu'il s'agit d'un prénom.
[21] Nous renvoyons ici à Jonasson (1994 : 171 : 238).
[22] Jonasson (1994 : 214) résume ces caractéristiques syntaxiques de ce GN : « Le Npr métaphorique est en général précédé d'un déterminant, et souvent accompagné de divers compléments ».
[23] Qui est déjà celui de Jonasson (1994 : 215), qui évoque « le Npr métaphorique [et] le SN dont il constitue la tête ».

L'extraction d'information[24], système d'analyse de données textuelles basé sur une analyse locale, n'effectue pas une analyse globale du texte, mais une recherche d'éléments textuels correspondant à un besoin donné. Le repérage des GN antonomasiques s'inscrit dans cette perspective : il s'agit de rechercher et d'extraire les segments de texte présentant des structures susceptibles de relever de l'antonomase du nom propre.

L'extraction d'information utilise, pour faire apparaître ces éléments textuels, des patrons d'extraction, syntaxiques, morpho-syntaxiques ou lexico-syntaxiques. Il s'agit, contrairement à la compréhension globale de textes ou à des approches lexicométriques, d'une « approche descendante » (Morin 1999 : 145) de l'analyse de corpus, processus d'analyse « guidé par la connaissance a priori des informations recherchées » (Poibeau et Nazarenko 1999 : 98), ce qui « suppose de savoir par avance ce qu'on cherche et de pouvoir le décrire par des indices de surface ». La recherche des GN antonomasiques doit donc être précédée de l'élaboration des patrons morpho-syntaxiques de l'antonomase.

*2.1.1. Description générale du système*

La démarche générale du traitement automatique de l'antonomase s'articule autour d'un corpus partitionné en deux, puis quatre sous-corpus. Ces différents groupes de textes sont soumis à une chaîne de traitements destinés à déterminer des patrons syntaxiques caractéristiques de l'antonomase (phase d'apprentissage), puis à repérer, sur des textes inconnus, des candidats-antonomases, séquences relevant des structures morpho-syntaxiques de l'antonomase (phase d'application).

*2.1.1.1. Le corpus, les sous-corpus*

Les sous-corpus *Entraînement* et *Test* sont destinés à la partie « apprentissage » du traitement automatique de l'antonomase, tandis que les sous-corpus *Portraits* et *Films* sont exploités dans la partie « application ». On a donc un sous-corpus d'apprentissage, sur lequel s'effectue l'acquisition de patrons syntaxiques de l'antonomase, et un sous-corpus d'application, qui permet de valider les méthodes utilisées et de rassembler un certain nombre d'occurrences antonomasiques. L'ensemble du corpus est soumis à un prétraitement destiné à homogénéiser le matériau de l'analyse[25].

Les fichiers de chacun des sous-corpus sont étiquetés à l'aide de la fonction d'étiquetage de texte de *Cordial*. Le paramétrage de cette fonction est effectué dans le but de 1) conserver le maximum d'informations tout en 2) normalisant les lignes du fichier. Ainsi, le traitement des erreurs d'orthographe, qui introduit d'importantes variations dans la structure de la ligne d'étiquetage, est supprimé, ainsi que, pour la même raison, le relevé des ambiguïtés.

Le fichier *Nom.cnr* fourni à la suite de l'étiquetage présente une segmentation du texte en mots[26] et en phrases, marquées par un délimiteur de début ou de fin de phrase. Entre ces délimiteurs, chaque mot de la phrase, suivi d'une série d'étiquettes, occupe une ligne.

[24] Présentée dans Poibeau et Nazarenko (1999).

[25] Chacun des articles constitue un fichier distinct. Après une première normalisation effectuée manuellement (la ponctuation est mise en forme ; le gras, mais non les italiques, est supprimé, le soulignement est réservé à la mise en valeur de l'antonomase, de façon à retrouver facilement les occurrences (pour les corpus d'apprentissage) ; la police, le corps et l'interligne sont uniformisés, et les lignes vides supprimées), chaque fichier est conservé en version document Word (*Nom.doc*) et en version texte seul (*Nom.txt*) ; cette dernière version est destinée à l'étiquetage, tandis que la première reste une version de vérification et de retour au texte originel.

[26] Certains problèmes classiques du découpage en mots, tels que l'apostrophe (*aujourd'hui*) sont bien traités par *Cordial*. D'autres, comme le trait d'union (*peut-être*, *avant-hier*) donnent lieu à des incohérences de traitement.

Chacune de ces étiquettes attache une information au mot concerné[27] : place du mot dans la phrase, mot, lemme, catégorie grammaticale du mot, numéro du mot-pivot du syntagme ou du sous-syntagme auquel appartient le mot, fonction grammaticale du mot, rang, dans la phrase, de la proposition à laquelle appartient le mot, type de cette proposition[28].

*2.1.1.2. Les étapes du système*

Le traitement automatique de l'antonomase peut être ramené à une suite de tâches distinctes[29]. Ces tâches sont à leur tour atomisées en traitements élémentaires effectués par une série de scripts *Perl* successifs.

La première tâche concerne la préparation des données textuelles. Deux groupes de données sont tout d'abord préparés pour l'analyse : les textes étiquetés sont nettoyés et formatés, pour aboutir à la forme sous laquelle ils seront exploités.

La seconde tâche est l'acquisition des patrons morpho-syntaxiques du GN antonomasique à partir de critères linguistiques externes et d'acquisition interne aux corpus. Elle se subdivise en plusieurs étapes qui sont décrites plus précisément ci-dessous.

La troisième est la phase de repérage des GN, suivie d'une évaluation et d'un affinement des patrons.

Dans une dernière étape, ces patrons sont projetés sur les corpus d'application.

Le système de traitement automatique de l'antonomase est donc modifié et enrichi au fur et à mesure, pour arriver à un outil capable de tester, sur un corpus d'application, une liste finie[30] de patrons morpho-syntaxiques de l'antonomase du nom propre.

*2.1.2. Les patrons*

Adopter une approche descendante, qui « part de connaissances plus ou moins riches sur le fonctionnement de la langue pour modéliser les informations à extraire » (Morin 1999 : 145), suppose de passer d'une définition naturelle de l'antonomase du nom propre à une définition opératoire, transcriptible en un (des) algorithme(s). Cela implique l'établissement, par l'appel à des connaissances externes au corpus, d'« indices de surface » (Poibeau et Nazarenko 1999 : 98) propres à décrire le phénomène recherché. Cette étape est souvent difficile et coûteuse : « Globalement, dans ce type d'approche, la phase de description des patrons, souvent manuelle, reste une forte contrainte » (Morin 1999 : 146). Le passage de la notion d'antonomase du nom propre à celle de GN antonomasique constitue un premier pas en ce sens.

*2.1.2.1. Des critères opératoires*

Il s'agit de dégager, de la manière la plus objective possible, des critères de repérage de l'antonomase. On distingue deux types de critères, des critères syntaxiques[31] et des critères

[27] Lorsque cette étiquette n'a pas lieu d'être, elle est remplacée par un tiret (sauf l'étiquette *Type Prop.*).
[28] Lorsque cette étiquette n'est pas présente, la ligne se termine par une tabulation.
[29] Une représentation de la chaîne de traitements est donnée en fin d'article.
[30] Ce point peut laisser croire que seront établis tous les patrons syntaxiques possibles de l'antonomase du nom propre. Ce n'est évidemment pas le cas, et un des prolongement possibles de ce travail serait l'amélioration de système de traitement dans la direction d'un apprentissage automatique permanent de nouveaux patrons. Ainsi, d'éventuelles antonomases présentes dans un corpus de référence et relevant d'un patron non prévu par le système ne seront pas reconnues.
[31] Le nom propre en antonomase est obligatoirement précédé d'une **détermination** (déterminant défini, indéfini, possessif, démonstratif, numéral...), **OU** dans une position syntaxique qui exclut la détermination : en apposition, objet d'un verbe comme *traiter de...*, *qualifier de...*, *traiter en...* et optionnellement précédé (entre le déterminant et le nom propre : *ce* ***petit***

sémantico-référentiels[32]. Seuls les critères du premier type sont facilement formalisables et sont donc à la base du traitement automatique de l'antonomase. Les critères du second type, qui correspondent à des critères non-opératoires, interviennent au moment de l'extraction manuelle des patrons, ainsi que pour évaluer la pertinence des résultats.

Les connaissances linguistiques externes ne permettent pas d'établir, à elles seules, des règles strictes sur l'entourage syntaxique du nom propre en antonomase. La difficulté d'établissement d'une norme externe pour décrire les patrons syntaxiques du groupe antonomasique conduit au choix d'une norme interne : les patrons sont élaborés à partir de groupes antonomasiques existants et reconnus comme tels, à l'intérieur du sous-corpus *Entraînement*. Cette démarche repose sur un postulat de représentativité du corpus, censé contenir toutes les possibilités syntaxiques du groupe antonomasique, bien que cette représentativité ne puisse pas être totale, en particulier en ce qui concerne le contexte droit.

*2.1.2.2. Acquisition des patrons*

L'acquisition des patrons morpho-syntaxiques de l'antonomase se fait en trois étapes principales : une élaboration manuelle, suivie d'une généralisation, puis d'un affinement des patrons.

On procède tout d'abord à une extraction manuelle des antonomases du sous-corpus *Entraînement* : l'ensemble des GN antonomasiques d'*Entraînement*[33] est relevé et soumis à un traitement qui sélectionne les étiquettes situées à gauche et à droite du nom propre. La position du nom propre dans le GN antonomasique pose parfois problème : s'il est entouré d'une détermination à gauche et d'une complémentation à droite, ces deux éléments forment les contextes gauches et droits du nom propre et sont stockés séparement dans des fichiers distincts. Lorsqu'en revanche le nom propre en antonomase constitue la limite gauche[34] ou droite[35] du GN, il est nécessaire de retenir des éléments de contexte situés hors du GN. Ces « faux » contextes gauches et droits sont également stockés dans des fichiers distincts.

L'étape suivante est celle de la généralisation : on effectue une « factorisation » des contextes gauches et droits : chacun des contextes gauches est couplé avec chacun des contextes droits et des faux contextes droits ; chacun des faux contextes gauches est couplé avec chacun des contextes droits. Cette opération a pour effet de multiplier les patrons possibles, tout en évitant de produire un patron formé, à gauche comme à droite, de faux contextes.

L'ensemble de ces patrons est alors projeté sur les corpus d'apprentissage *Entraînement* et *Test* et propose un nombre important de groupes de mots, correspondant aux structures morpho-syntaxiques formées par les couples de contextes. Ces groupes de mots sont des candidats au statut d'antonomase. Une intervention manuelle est nécessaire pour évaluer ces résultats.

---

*Mussolini*) et/ou suivi (*un Godard* ***de sous-préfecture***) d'un ou plusieurs (***une sorte de jeune*** *Boris Vian* ***du polar***) **compléments** (adjectif, complément du nom, proposition relative).

[32] Le nom propre en antonomase est associé à un référent (personne, lieu...) qui bénéficie d'une certaine **notoriété** (notoriété « historique », très large (*Homère*, *Aristote*, *Néron*, *Mozart*...) ; notoriété « médiatique », plus restreinte dans l'espace et dans le temps (*Gabin, Mandela, Bill Gates*, *Bernard Tapie*...) ; notoriété « discursive », locale, (le référent du nom propre a été ou sera présenté au cours du discours précédant ou suivant l'antonomase)) mais il ne désigne **pas**, en l'occurrence, le référent qui y est habituellement associé.

[33] 235 occurrences.

[34] Comme par exemple dans le cas de l'apposition sans détermination : Mel Brooks, *Don Quichotte du rire*, … (Exemple de Jonasson (1994)).

[35] Comme par exemple en l'absence de complémentation : Hauteclaire Stassin était sérieuse comme *une Clorinde*. (Exemple de Fromilhague (1995)).

On procède alors à l'amélioration et à l'affinement des patrons. Les fichiers de contextes correspondants sont complétés par certains contextes gauches ou droits, non présents au sein du sous-corpus *Entraînement*, dont l'absence a empêché le repérage d'occurrences du sous-corpus *Test*. Chacun des patrons composé d'un couple (faux) contexte gauche / (faux) contexte droit est évalué en fonction de ses résultats et de la proportion d'antonomases correspondant à ces résultats, pour obtenir un classement des patrons selon leur pertinence[36]. Lorsqu'un même patron a obtenu des résultats différents sur *Entraînement* et sur *Test*, le taux de pertinence est la moyenne des deux taux.

Les patrons ainsi classés par ordre de pertinence sont alors projetés sur les corpus d'application. À l'issue de ce passage, le système propose des candidats-antonomases dont l'évaluation est facilitée par l'indication d'un niveau de confiance tiré du taux de pertinence du patron

### *2.2. Les résultats*

On utilise pour l'évaluation des résultats du traitement automatique des noms propres les notions de précision, de rappel, de bruit et de silence[37], mesurées et analysées pour la partie « apprentissage » puis pour la partie « application » du traitement.

#### *2.2.1. Les corpus d'apprentissage*

L'évaluation des résultats des deux sous-corpus d'apprentissage valide une première étape du traitement, celle de la constitution et du test des patrons morpho-syntaxiques. Les résultats reflètent deux niveaux de difficulté, dans la mesure où le repérage des antonomases, appliqué à *Entraînement*, a pour fonction de retrouver, sous forme de patrons syntaxiques, les GN antonomasiques qui en ont été extraits, tandis qu'il doit, sur *Test*, repérer des GN antonomasiques inconnus. On peut donc s'attendre à un taux d'erreurs plus important sur le second sous-corpus que sur le premier.

##### *2.2.1.1. Précision et bruit*

Le taux de précision du sous-corpus *Entraînement* est d'environ 10%[38], celui du sous-corpus *Test* est donc d'environ 7%[39]. Les problèmes révélés par ces taux de précision assez faibles sont éclairés par l'examen du bruit, dont le taux du sous-corpus *Entraînement* est d'environ 90%[40], tandis que celui du sous-corpus *Test* est d'environ 94%[41].

---

[36] La pertinence d'un patron est notée par un chiffre entre 0 et 1, selon le nombre de segments ramenés correspondant effectivement à un GN antonomasique (0 : aucun des segments ramenés n'est un GN antonomasique ; 1 : tous les segments ramenés sont des GN antonomasiques, 0,5 : la moitié des segments ramenés sont des GN antonomasiques). Certains patrons ne donnent aucun résultat ; leur pertinence est alors notée 00.

[37] « En recherche documentaire, la *précision* représente la proportion de réponses pertinentes données par rapport au total des réponses extraites. Le *rappel* est la proportion des réponses pertinentes extraites par rapport au total des réponses pertinentes possibles. Le *silence* correspond alors aux réponses pertinentes non extraites. Le *bruit* renvoie aux informations non pertinentes produites » (Habert *et al.* 1997 : 11). Pour le calcul des taux de précision, de rappel, de silence et de bruit, nous distinguerons le nombre de bonnes réponses cherchées, qui correspond au nombre réel d'occurrences d'antonomases dans le sous-corpus, du nombre de bonnes réponses trouvées, qui peut être supérieur au nombre réel d'occurrences d'antonomases dans le sous-corpus car certains GN antonomasiques peuvent être repérés plusieurs fois, plus ou moins tronqués. Ainsi, le nombre de bonnes réponses cherchées du sous-corpus *Entraînement* est de 235, tandis que celui des bonnes réponses trouvées est de 340 ; le nombre de bonnes réponses cherchées du sous-corpus *Test* est de 64, tandis que celui des bonnes réponses trouvées est de 67.

[38] Précision *Entraînement* = 340/3660 = 0,0928961.

[39] Précision *Test* = 67/1061 = 0,0631479.

[40] Bruit *Entraînement* = 3314/3660 = 0,9054644.

[41] Bruit *Test* = 994/1061 = 0,9368520.

Le taux de bruit, sur l'un et l'autre sous-corpus, est très important et doit être mis en rapport avec la pertinence des patrons. Une énorme majorité de patrons a une pertinence 00[42]. Ces patrons ne produisent pas directement de bruit, mais sont susceptibles de le faire sur un corpus plus volumineux[43]. D'autres patrons, qui reçoivent un taux de pertinence de 0, ne ramènent que des segments non pertinents, constituant ainsi le bruit réel. Ils ne doivent cependant pas être éliminés : le patron *Da-fs-d (Np)+ Afpfs*, qui ne ramène pas de GN antonomasique (mais des noms propres caractérisés tels que *la France métropolitaine* ou *la Telecaster noire*) doit être conservé, car il représente une des structures de GN antonomasique possibles, même s'il n'en ramène aucun sur les corpus soumis au repérage[44].

Le bruit provient également, partiellement, des patrons à taux de repérage compris entre 0 et 1, ce qui relève de plusieurs types de problèmes. Il se peut tout d'abord qu'à ce patron corresponde un GN antonomasique qui relève d'une structure exceptionnelle, mais attestée : le patron *Da-ms-d (Np)+ Yps* ramène 113 segments, dont un seul (*le César.*) est un GN antonomasique, ce qui lui donne un taux très faiblement supérieur à 0.

Le taux de pertinence d'un patron peut être abaissé par des erreurs d'étiquetage : 2 des 3 segments ramenés le patron *Da-ms-i (Np)+ Sp Da-fs-d Ncfs* sont des antonomases, ce qui lui confère un taux de pertinence de 0,66. Le segment « fautif » est *un ruban sur la joue*, où *ruban* est par erreur étiqueté « nom propre » par *Cordial*. Cette erreur d'étiquetage ne dépendant pas du système de repérage, on affecte à ce patron un taux de pertinence 1.

Il arrive également qu'on soit confronté aux limites des critères morpho-syntaxiques tenus pour opératoires. C'est le cas lorsque le nom propre du segment ramené est utilisé dans un emploi modifié non métaphorique ; seule une évaluation manuelle tenant compte du second groupe de critères, sémantico-référentiels, peut l'écarter. C'est par exemple le cas pour le patron *Ds1.s. Np Afpfs*, qui ramène le segment *notre Marthe Richard nationale*. Il ne s'agit pas d'un nom propre en antonomase, mais d'une caractérisation au sein d'une locution figée. Or, on trouve ailleurs la même structure[45] pour un GN antonomasique : *notre Bambi monarchiste*. On trouve également d'autres emplois modifiés, qui ne peuvent être distingués de l'emploi antonomasique que de façon manuelle. Ce type d'incident survient aussi bien sur les patrons à taux de confiance 0 que sur ceux qui atteignent un taux de confiance compris entre 0 et 1.

Certains éléments des patrons morpho-syntaxiques sont à l'origine de bruit, en particulier des éléments de contexte gauche ou droit qui ne sont pertinents que pour une réalisation lexicale précise, tous les autres produisant du bruit. On s'aperçoit par exemple que le patron *Ncfs Sp Np Sp* ramène, sur le sous-corpus *Entraînement*, 26 segments, dont seulement un GN antonomasique. Ce patron n'est globalement pas pertinent pour l'antonomase du nom propre (comme le montrent les autres segments ramenés), il ne l'est que si l'élément « mot » correspondant à l'étiquette Ncfs est *sorte*[46].

*2.2.1.2. Rappel et silence*

[42] A l'issue de la phase d'apprentissage, sur les 8541 patrons formés par le couplage des contextes gauches et droits, 7923 a un taux de pertinence 00 ; 309 ont un taux 0, 204 un taux 1 et 105 un taux compris entre 0 et 1.

[43] Un certain nombre d'entre eux n'extrairont cependant jamais rien, car ils forment des segments incorrects ; c'est par exemple le cas des patrons qui associent des contextes gauche et droit de genre et/ou de nombre différents.

[44] Ce problème est à relier à celui des marques de genre, nombre et personne car ce patron, dans sa version « masculine » a un taux de confiance supérieur à 0.

[45] La seule différence est que le masculin y remplace le féminin.

[46] Ce problème concerne d'autres catégories que celle du nom commun : le patron *Afpms Np Yps* n'est pertinent que lorsque à l'élément morpho-syntaxique *Afpms* correspond un élément lexical *nouveau*. De même, à l'élément morpho-*syntaxique Da-.p-i*, dans le patron *Da-.p-i Np Afpmp*, doit correspondre à l'élément lexical *du*.

Le taux de rappel du sous-corpus *Entraînement* est d'environ 95%[47], celui du sous-corpus *Test* d'environ 74%[48]. Comme il était prévisible, il est meilleur pour *Entraînement* que pour *Test*. Les problèmes révélés par le taux de rappel sont éclairés par l'examen du silence, dont le taux est, sur le sous-corpus *Entraînement*, d'environ 5%[49] et, sur le sous-corpus *Test*, d'environ 26%[50]. L'examen de ces silences guide l'affinement des patrons et conduit à des modifications du système de repérage. On peut distinguer plusieurs causes de silences.

Il arrive tout d'abord qu'une erreur d'étiquetage du nom propre par *Cordial* soit à l'origine d'un certain nombre de silences[51]. Ce type de silence peut être évité au niveau de la préparation des sous-corpus d'apprentissage, mais ne pourra pas être prévu pour les corpus d'application.

Certains silences sont spécifiques du sous-corpus *Test*. Pour une moitié, il s'agit de « faux » silences : le GN antonomasique n'est pas reconnu en entier, mais tronqué sur sa droite. Il ne s'agit pas d'un échec total du repérage, mais d'un problème de délimitation à droite[52].

Les « véritables » silences sont dus à des insuffisances des patrons syntaxiques. Il s'agit, dans environ 2/3 des cas, d'un problème lié à l'absence totale ou, plus souvent, partielle du contexte gauche. Il arrive également que l'élément de contexte gauche soit trop spécifique, en particulier au niveau des genre et nombre. Dans la moitié des cas[53], c'est le contexte droit qui est en cause, qu'il soit inexistant ou incomplet dans la liste des contextes droits. Les contextes gauches ou droits non prévus sont ajoutés aux listes[54].

### *2.2.2. Les corpus d'application*

Les résultats de la projection des patrons morpho-syntaxiques de l'antonomase sur les corpus d'application sont comparables pour l'un et l'autre sous-corpus : *Portraits* comporte 29 antonomases, dont 28 sont reconnues par le système, *Films* en comporte 27, dont 22 sont ramenées automatiquement. Si le taux de rappel reste satisfaisant, le taux de bruit est, là encore, trop élevé. Cependant, le marquage des réponses par un taux de pertinence lié au patron permet d'effectuer un premier tri parmi les réponses.

Les segments ramenés par des patrons de taux 1 sont relativement rares. Il s'agit généralement de bonnes réponses, les erreurs étant dues aux causes évoquées ci-dessus[55]. Les segments rapportés par des patrons de taux 0 constituent quasiment tous de mauvaises réponses. Ce sont les patrons de taux compris entre 0 et 1 qui rapportent la plupart des GN antonomasiques, tout en nécessitant l'intervention manuelle la plus importante. Enfin, les patrons de taux 00 ramènent quelques segments, dont certains constituent des antonomases, ce qui justifie *a posteriori* la factorisation des contextes gauches et droits.

Les causes de silence restent les mêmes que lors de la phase d'apprentissage : erreurs d'étiquetage produites par *Cordial*, contexte ou gauche non prévu. Malgré tout, le taux de

[47] Rappel *Entraînement* = 222/235 = 0,9446808
[48] Rappel *Test* = 47/64 = 0,734375.
[49] Silence *Entraînement* = 11/235 = 0,0468085.
[50] Silence *Test* = 17/64 = 0,265625.
[51] Sur *Entraînement*, les noms propres de *sinistre Bérézina (.)*, *nos trois Gnafron (:)* et les *Césars (.)*, sont étiquetés « nom commun », tout comme, sur *Test*, ceux de *les Cassandres (qui)* et *du de Gaulle d' arrondissement*.
[52] Ces silences étant partiels, ils ne sont pas pris en compte dans le calcul du taux de rappel. Si l'on tient compte de ces GN incomplets, le taux de rappel du sous-corpus *Test* culmine à environ 50%, ce qui est nettement moins satisfaisant.
[53] Certains de ces GN antonomasiques cumulent les problèmes, en contexte gauche comme en contexte droit.
[54] Lorsqu'il s'agit d'un problème lié aux genre et nombre, on peut être tenté de supprimer ces marques, de faire des patrons « neutres ». Cela risque cependant d'entraîner une augmentation du bruit, et d'introduire des variations indésirables.
[55] Erreurs d'étiquetage, noms propres en emploi modifié non antonomasique.

rappel de près de 90 % peut être considéré comme satisfaisant.

Certains GN antonomasiques sont, là encore, cependant ramenés incomplets sur leur droite. Se pose donc à nouveau la question de la délimitation à droite de l'antonomase. L'évaluation du découpage de l'antonomase relève elle aussi, dans une certaine mesure, de critères non opératoires et il semble qu'on a là un problème de délimitation qui dépasse le cadre du traitement automatique de l'antonomase, mais que la démarche de traitement automatique contribue à mettre à jour.

*2.2.3. Perspectives*

Le système de traitement automatique de l'antonomase, si incomplet et rudimentaire soit-il, permet d'ouvrir deux types de perspectives : il paraît, d'une part, nécessaire d'améliorer l'outil de repérage, afin d'augmenter les taux de rappel et, surtout, de précision ; on peut, d'autre part, utiliser les résultats pour décrire les fonctionnements discursifs de l'antonomase, afin d'en déterminer les structures préférentielles, tant au niveau microstructural que macrostructural.

L'amélioration du système de traitement automatique de l'antonomase suppose un travail sur les patrons, ce qui implique des modifications du programme de repérage visant à tenir compte des étiquettes fonctionnelles dans la délimitation des GN antonomasiques. Or la représentation par étiquetage morpho-syntaxique effectuée par *Cordial* est sur ce point insuffisante. L'obtention d'un repérage exact à droite comme à gauche du nom propre en antonomase semble devoir s'articuler à une annotation fonctionnelle et syntaxique du corpus, dans une représentation arborescente et non plus linéaire. Dans cette même direction, on peut également envisager un marquage de l'antonomase sous forme de balises, visant à intégrer cet emploi du nom propre dans des structurations textuelles telles que celles recommandées par la *TEI* (*Text Encoding Initiative*)[56].

L'utilisation des résultats fournis par les patrons morpho-syntaxiques, même grossiers, est cependant possible. L'observation des fréquences des patrons morpho-syntaxiques permet de mener une analyse linguistique tenant compte des structures préférentielles de la figure, distinguant structure prototypique et structure exceptionnelle. On peut ainsi compléter des analyses sémantiques et syntaxiques par l'observation sur corpus du comportement de l'antonomase en discours.

## Références

Béchet F. et Yvon F. (2000). Les noms propres en traitement automatique de la parole. *Traitement automatique des langues* 41-3 : 671-707.

Bodenreider O. et Zweigenbaum P. (2000). Stratégies d'identification des noms propres à partir de nomenclatures médicales parallèles. *Traitement automatique des langues* 41-3 : 725-757.

Bruneseaux F. (1998). Noms propres, syntagmes nominaux, expressions référentielles : repérage et codage. *Langues* 1-1 : 46-59.

Daille B. et Morin E. (2000). Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *Traitement automatique des langues* 41-3 : 601-621.

Coates-Stephens S. (1993). The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *Computers and the Humanities* 26 : 441-456.

[56] Voir Bruneseaux (1998) et Flanders *et al.* (1998) pour l'intégration des noms propres (en emploi standard et dans une perspective classificatoire) au sein de la *TEI.*

Cucciarelli A., Luzi D et Velardi P. (1999). Semantic Tagging of Unknown Proper Names. *Natural Language Engineering* 5-2 : 171-185.

Flanders J, Bauman S., Caton P. et Cournane M. (1998). Names Proper and Improper : Applying the TEI to the Classification of Proper Nouns. *Computers and the Humanities* 31 : 285-300.

Fromilhague C. (1995). *Les Figures de style*. Nathan.

Gary-Prieur M.N. (1994). *Grammaire du nom propre*. Presses Universitaires de France.

Habert B., Nazarenko A. et Salem A. (1997). *Les Linguistiques de corpus*. Armand Colin / Masson.

Habert B., Fabre C. et Issac F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. Masson.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *Cahiers de l'Université de Perpignan* 31 : 11-58.

Jonasson K. (1994). *Le Nom propre. Constructions et interprétations*. Duculot.

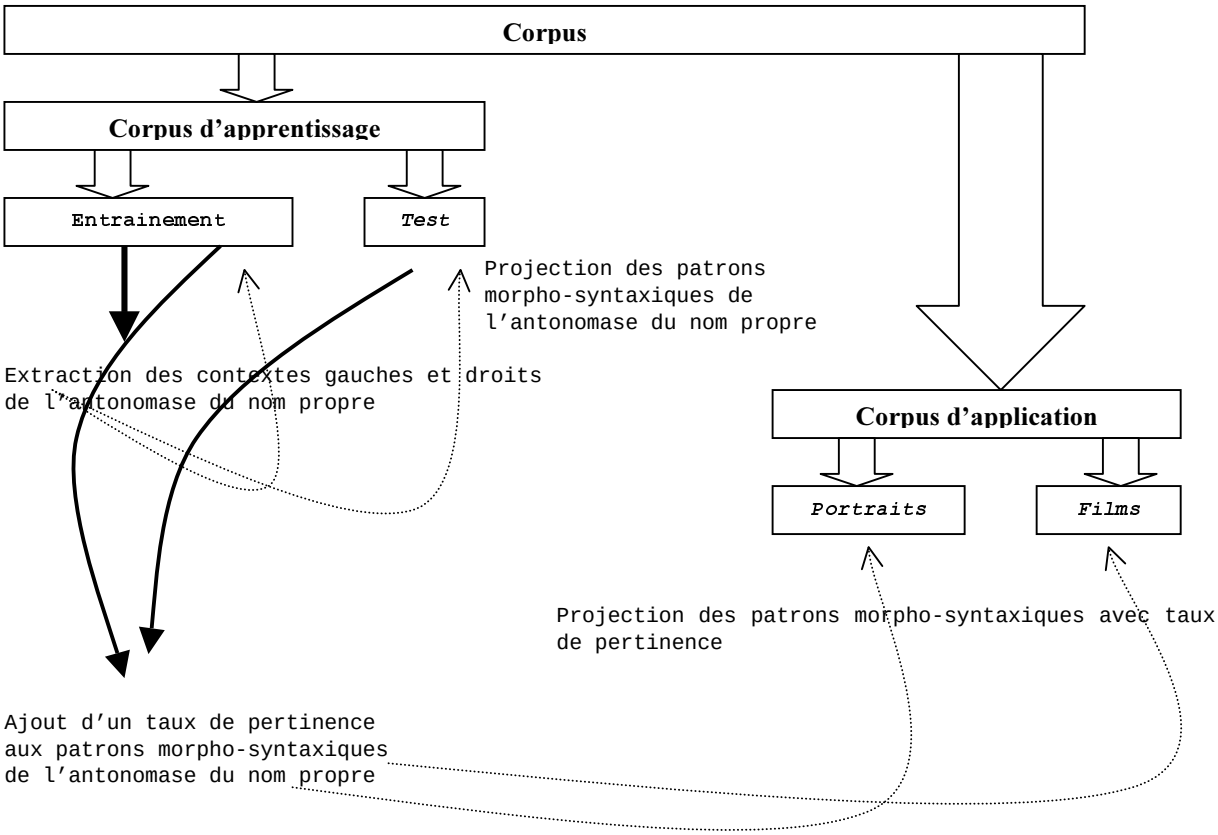
Leroy S. (2000). Repérage des GN antonomasiques dans un corpus de presse. Rapport technique. U.M.R. C.N.R.S. 5475-Praxiling – Montpellier III.

Morin E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement automatique des langues* 40-1 : 43-166.

Poibeau T. et Nazarenko A. (1999). L'extraction d'information, une nouvelle conception de la compréhension de texte ? *Traitement automatique des langues* 40-2 : 87-115.

Sinclair J. (1996). Preliminary Recommendations on Corpus Typology. Rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards).

Wolinski F., Vichot F. et Dillet B. (1995). Automatic Processing of Proper Names in Texts. In *Proceedings of the 7th Conference on Computational Linguistics (EACL'95)*, Dublin, pp. 23-30.




*Représentation de la chaîne des traitements.*