De l'étiquetage traditionnel au transducteur du texte : la levée d'ambiguïtés par grammaires locales

Anne DISTER

Abstract: Many words are ambiguous in their part of speech. We propose to see a method to disambiguate this words by the analyse of their local context described in grammars. Intex is the system applied to describe this context. We will see the ambiguity in the electronic dictionaries and what kind of results it's possible to have.

Introduction

L'ambiguïté est la pierre d'achoppement de toute étude de la langue et, plus particulièrement, de tout traitement automatique.

En effet, dès lors qu'elle est étudiée ou manipulée à travers le prisme de l'ordinateur, la langue n'est qu'ambiguïtés. Avant même d'aborder la sémantique, des traitements de base tels que, par exemple, la segmentation d'un texte en phrases posent de multiples problèmes : qu'on songe simplement à l'ambiguïté du point (final, de suspension, abréviatif) et de la majuscule (de début de phrase, des noms propres, dans les sigles, etc.).

Le type d'ambiguïtés dont nous traiterons ici est généralement qualifié de morpho-syntaxique : il s'agit d'assigner à chaque forme

✉ Anne DISTER, Université de Liège, 1b, Quai Roosevelt, B-4000 Liège.
Fax : 32 4 366 57 84 e-mail : adister@ulg.ac.be

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXV, 1 à 4, 1999.

d'un texte sa catégorie grammaticale. Le problème qui nous concerne est donc celui des homographes.

L'ambiguïté dans les dictionnaires électroniques

Le système que nous utilisons dans le cadre de notre recherche est le système Intex, élaboré par Max Silberztein[1]. Intex est un environnement de développement linguistique qui comprend des dictionnaires à large couverture. Dans le système, une forme sera considérée comme ambiguë si elle correspond à plusieurs entrées du dictionnaire. Blandine Courtois a calculé que près de 25 % des formes du dictionnaire des mots simples (`delafm`) étaient ambiguës[2].

Voici les entrées extraites du `delafm` pour les mots *la*, *lâche* et *autour* :

```
la,la.N:ms:mp
la,le.DET:fs
la,le.PRO:3fs

lâche,lâche.A+g:ms:fs
lâche,lâche.N:ms:fs
lâche,lâcher.V:P1s:P3s:S1s:S3s:Y2s

autour,autour.N:ms
autour,autour.PREP
```

Chaque ligne constitue une entrée du dictionnaire et contient les informations suivantes :

- la forme telle qu'on peut la trouver dans un texte ;
- une virgule ;
- le lemme : il s'agit du mot-vedette d'un dictionnaire traditionnel, c'est-à-dire, *grosso modo*, de la forme au masculin singulier pour les noms et les adjectifs, de l'infinitif pour les verbes ;
- un point ;

[1] Pour une description détaillée du système, voir Silberztein (1993).

[2] Voir Courtois (1996).

– la catégorie grammaticale : **A** pour adjectif, **N** pour nom, **V** pour verbe, **ADV** pour adverbe, **PREP** pour préposition, **CONJS** pour conjonction de subordination, etc. ;
– éventuellement, après le signe +, des indications sur la transitivité des verbes ou la position des adjectifs ;
– le signe du double point ;
– selon la catégorie, des informations flexionnelles : genre, nombre, temps, mode, etc. [1].

La forme *la* est trois fois ambiguë : il peut s'agir d'un nom masculin singulier ou pluriel (la note de musique), d'un déterminant féminin singulier ou d'un pronom à la troisième personne du féminin singulier. *Lâche* est considéré comme trois fois ambigu : adjectif ou masculin ou au féminin singulier ; nom masculin singulier ; forme (à la première ou troisième personne du singulier de l'indicatif présent **P** ou du subjonctif présent **S**, ou à la deuxième du singulier de l'impératif **Y**) du verbe *lâcher* [2].

Lever les ambiguïtés pour l'étiquetage linéaire de texte

Ces ambiguïtés compliquent considérablement l'étiquetage de texte. Cette tâche consiste à remplacer chaque mot — ou groupe de mots — du texte par une série d'informations qui correspondent peu ou prou à celles figurant dans les entrées du dictionnaire présentées plus haut (lemme, catégorie, informations flexionnelles).

Mais la plupart des ambiguïtés, on le sait, ne sont que virtuelles, et bien souvent, une analyse du contexte immédiat suffit à les résoudre. C'est par l'analyse de ce contexte, décrit dans des *grammaires locales*, que nous levons les ambiguïtés.

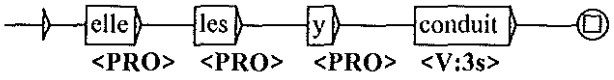
---

[1] Ce type d'information est inexistant pour des catégories telles que adverbe, préposition ou encore conjonction.

[2] Les ressources lexicales qui composent les dictionnaires d'Intex sont issues de l'étude systématique de la langue française entreprise par le LADL (Laboratoire d'Automatique Documentaire et Linguistique), depuis 1967, sous la direction de Maurice Gross.

Une grammaire de levée d'ambiguïtés

Les grammaires locales construites avec Intex sont basées sur le formalisme des automates à état fini. Il s'agit en fait de transducteurs dont la forme graphique est la suivante :



Graphe1

Un transducteur est représenté par un graphe qui contient des nœuds, dont le nœud initial et le nœud terminal, et des transitions. Il s'agit en fait d'un automate à deux bandes : une bande de lecture qui reconnaît une séquence dans le texte (ici, *elle les y conduit*), et une bande d'écriture qui associe à cette séquence préalablement reconnue les indications présentes sous les boîtes (ou nœuds). Ces indications sont appelées contraintes, et correspondent ici à des catégories grammaticales. Elles imposent à *elle*, à *les* et à *y* d'être un pronom, et à *conduit* d'être un verbe à la troisième personne du singulier. Le résultat de l'étiquetage, après l'application de *Graphe1* sur la séquence *elle les y conduit* est le suivant :

```
{elle,.PRO:3fs} {les,le.PRO:3mp:3fp} {y,.PRO}
{conduit,conduire.V+t:Kms:P3s}.
```

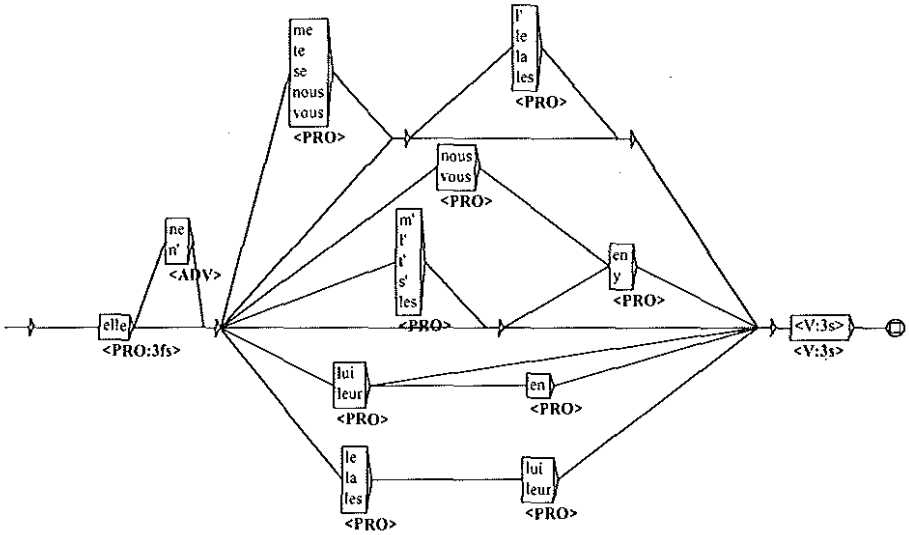
Tous les mots sont étiquetés, donc désambiguïsés, alors que trois d'entre eux étaient ambigus : *les* PRO ou DET, *y* PRO ou N, et *conduit* N ou V.

On le voit dans *Graphe1*, mis à part le nœud initial et le nœud terminal, chaque nœud du transducteur est étiqueté par un symbole. Ici, c'est la forme telle qu'elle apparaît dans le texte. Mais il peut également s'agir :

1) d'une catégorie grammaticale, notée entre crochets et en majuscules `<CAT>` : `<DET>`, `<PRO>`, `<V>`, etc. Il est possible également de prendre en compte des informations flexionnelles : `<DET:ms>`, `<V:F3p>`, etc.
2) d'un lemme, noté entre crochets `<lemme>` : `<conduire>`, `<poisson>`, etc.

3) du symbole vide : <E>.

Graphe1 reconnaît dans un texte *elle les y conduit*, mais non des séquences telles que *elle les conduit*, *elle y conduit*, *Pierre les y conduit*, *elle les y conduira*, etc. En effet, pour qu'une séquence soit reconnue par un automate, il faut, partant du nœud initial, passer par tous les nœuds de l'automate pour arriver au nœud final (on lit de gauche à droite). Dès lors que la séquence n'est pas reconnue, les contraintes ne sont pas appliquées et aucune analyse n'est fournie.



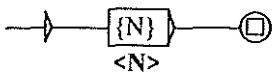
Graphe2

Graphe1 est extrait d'une grammaire plus complète (ici *Graphe2*), qui analyse les formes entre le pronom *elle* sujet et le verbe à la troisième personne du singulier. *Elle les conduit* et *elle y conduit* sont reconnus, ainsi que *elle les y conduira*, puisque l'étiquette n'est plus *conduit* mais <V:3s>. En l'occurrence, l'étiquette avec le lemme noté entre crochets <conduire> aurait également permis de reconnaître la séquence. Notons que *Pierre les y conduit* ne sera toujours pas analysé puisque avant d'arriver à *les*, la grammaire oblige à trouver sur son chemin la forme *elle*.

Les grammaires de levée d'ambiguïtés du français

Actuellement, une cinquantaine de grammaires de levée d'ambiguïtés ont été écrites pour le français[1]. La « philosophie » qui nous a guidée lors de leur élaboration est la suivante : nous préférons ne pas désambiguïser totalement un texte, mais le désambiguïser correctement. En d'autres termes, nous obtenons du silence (dans le cas de l'étiquetage linéaire, tous les mots n'ont pas reçu d'étiquette), mais évitons à tout prix le bruit (le texte ne présente aucune étiquette incorrecte)[2].

Les grammaires d'Intex ont été construites pour tourner en deux passes successives : les grammaires de première passe s'appliquent simultanément sur le texte; le résultat obtenu est sauvegardé, et les grammaires de deuxième passe peuvent alors s'appliquer sur ce résultat. Ce qui différencie fondamentalement ces deux types de grammaires, c'est que celles de passe2 travaillent sur un texte déjà partiellement désambiguïsé. Elles contiennent donc dans leur chemin des mots ou des catégories non ambigus qui se présentent de la manière suivante :



Graphe3

Graphe3 doit reconnaître dans le texte un mot dont on est certain qu'il s'agit d'un nom, puisqu'il a déjà été analysé comme tel lors de la première passe de désambiguïsation. Notons que l'étiquette dans la boîte est entre accolades et non plus entre crochets.

---

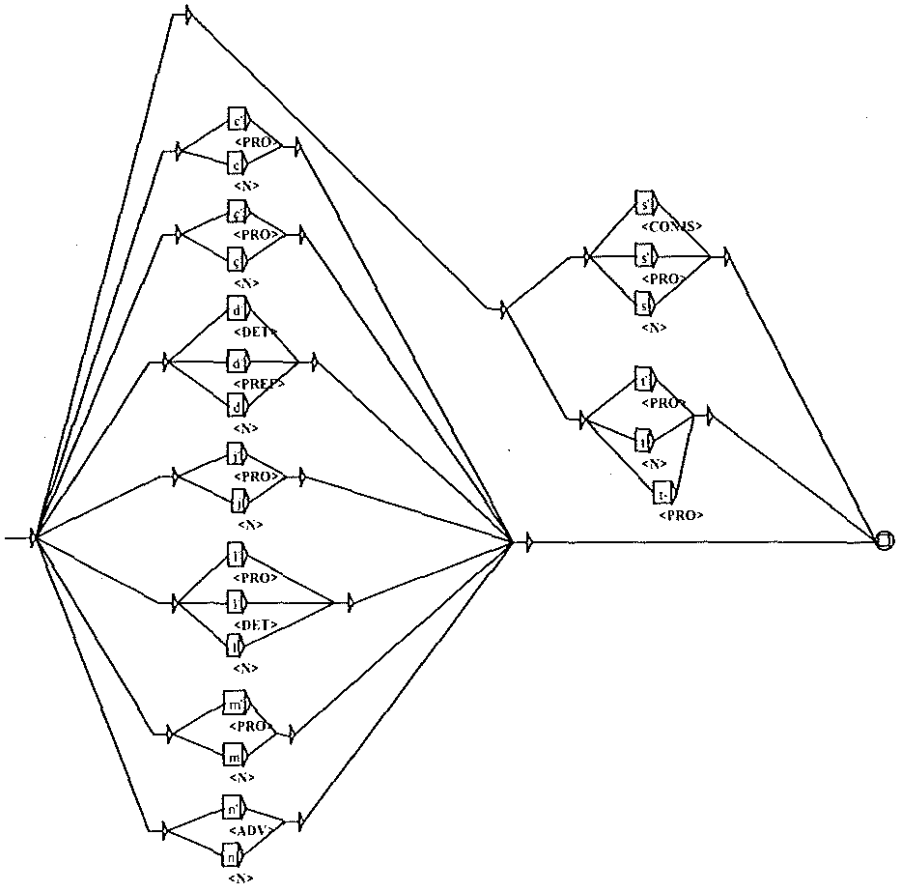
[1] Elles sont disponibles à l'adresse suivante :
http://www.ladl.jussieu.fr/tools/tools.html

[2] Cette approche de travail est tout à fait particulière dans un domaine où le but ultime semble être de tout étiqueter, avec des taux d'erreur annoncés à 5 %, 3 % voire moins. Le problème est évidemment que l'on ne sait jamais quand on rencontre un mot faisant partie de ces 5 % ou ces 3 %. De plus, de tels résultats sont la plupart du temps obtenus par des systèmes dont les dictionnaires sont lacunaires (moins d'hypothèses de départ, donc moins d'ambiguïtés à lever) et ne tiennent aucun compte des mots composés ou des expressions figées. Certains systèmes n'utilisent aucun dictionnaire, faisant de l'apprentissage sur corpus. D'autres créent des catégories mixtes regroupant certains types d'ambiguïtés.

Les grammaires que nous avons construites couvrent une grande variété d'ambiguïtés et abordent des problèmes relativement différents. Nous en présentons ici quatre, trois de première passe et une de deuxième passe.

Désambiguïser les lettres de l'alphabet

Onze lettres de l'alphabet sont ambiguës; *Graphe4* lève l'ambiguïté pour neuf d'entre elles :



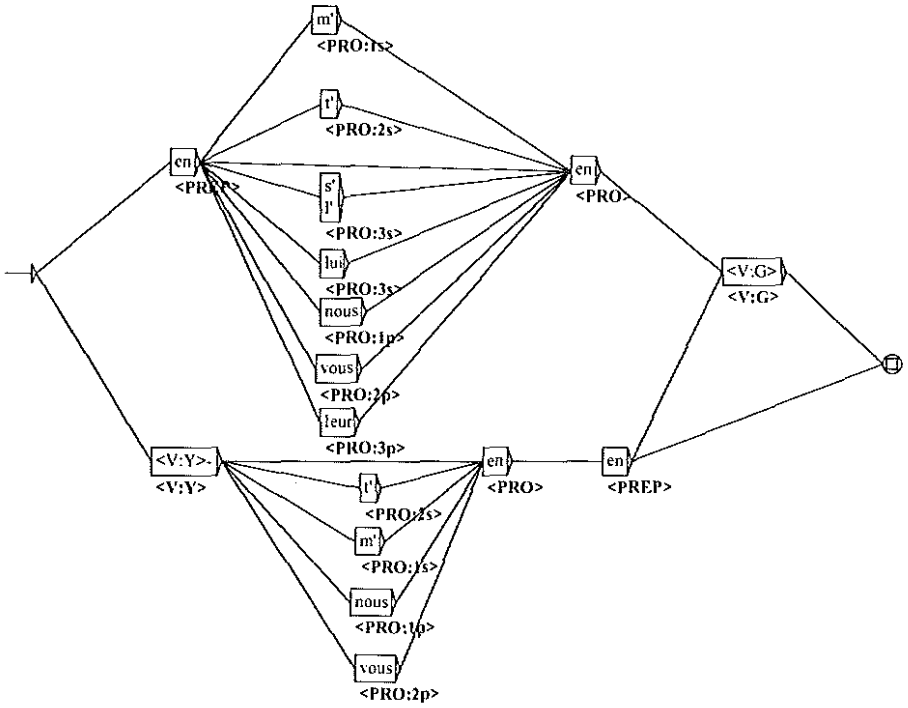
Graphe4

La présence ou non de l'apostrophe après les lettres *c*, *ç*, *d*, *j*, *l*, *m*, *n*, *s* et *t* nous permet de faire le départ entre l'hypothèse « nom » (c'est-à-dire la lettre de l'alphabet en tant que telle), et une ou plusieurs autres hypothèses. Mais certaines lettres suivies de l'apostrophe restent ambiguës : *d'* déterminant ou préposition, *l'* déterminant ou pronom. En revanche, il est possible de lever l'ambiguïté du *s'* en élargissant le contexte droit : suivi de *il* ou de *ils*, *s'* est conjonction de subordination. Dans tous les autres cas, c'est un pronom.

Pour les lettres *a* et *y*, il est impossible de lever l'ambiguïté de cette façon puisque *a* verbe et *y* pronom ne prennent pas d'apostrophe.

La désambiguïsation de la forme *en*

Graphe5 analyse des séquences dans lesquelles se trouvent consécutivement, ou séparées par un pronom, deux formes *en*.



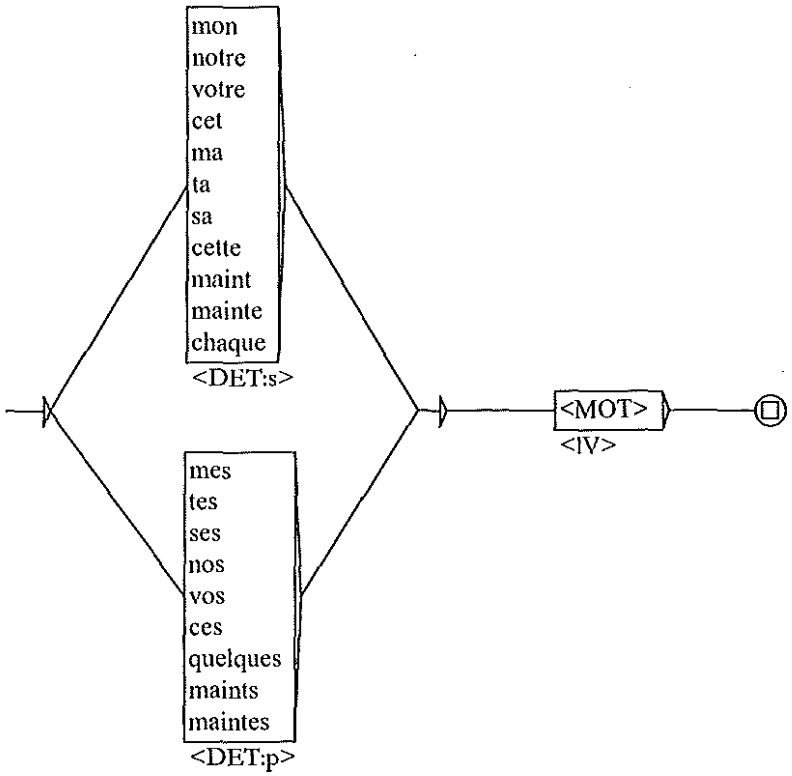
Graphe5

Cette grammaire permet d'analyser correctement des séquences comme *en en voulant*, *en lui en donnant*, *allons-nous en en (Corse)* [1] *!*, *prenez-en en marchant*, *donne-m'en en (échange)*, etc.

Graphe5 désambiguïse non seulement toutes les formes *en*, mais également tous les mots composant la séquence reconnue.

Or, ce n'est pas toujours le cas, comme le montre *Graphe6*.

Grammaire sur base de déterminants non ambigus



Graphe6

[1] Les mots notés entre parenthèses n'appartiennent pas à la séquence reconnue. Nous les indiquons néanmoins pour plus de lisibilité.

Graphe6 se base sur une liste de 20 déterminants non ambigus : à la suite de l'un de ces déterminants, on ne peut trouver un verbe (<!V>).

Si ce le mot est ambigu verbe/une autre catégorie, l'ambiguïté sera levée et le texte sera étiqueté. C'est le cas de :

- {chaque,.DET:ms:fs} {dîner,.N:ms} : *dîner* est aussi une forme du verbe dîner, à *l'infinitif.*
- {ses,son.DET:mp:fp} {affiches,affiche.N:fp} : *affiches* est aussi une forme du verbe *afficher* à la deuxième personne du singulier de l'indicatif ou du subjonctif présent.
- {ces,ce.DET:mp:fp} {portions,portion.N:fp} : *portions* est aussi une forme du verbe *porter* à la première personne du pluriel de l'imparfait.

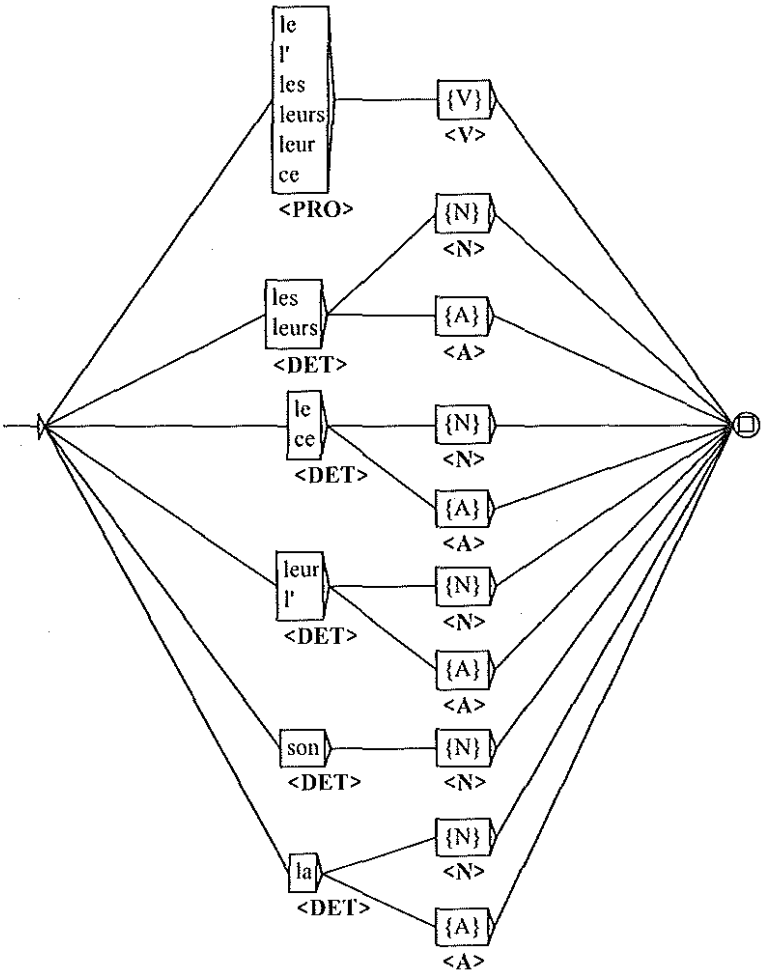
En revanche, s'il existe, outre l'hypothèse « verbe », deux autres hypothèses (voire plus) pour ce mot, il ne sera pas étiqueté. C'est le cas de séquences comme *cette innocente*, *mon devant* ou *vos caches*. Voici les entrées du `delafm` pour ces mots :

```
innocente,innocent.A+g:fs
innocente,innocent.N:fs
innocente,innocenter.V+t:P1s:P3s:S1s:S3s:Y2s

devant,devant.ADV
devant,devant.N:ms
devant,devant.PREP
devant,devoir.V:G

caches,cache.N:fp
caches,cache.N:mp
caches,cacher.V:P2s:S2s
```

Vos caches n'est pas désambiguïsé dans la mesure où nous choisissons un étiquetage dont les étiquettes très complètes reprennent la totalité des informations présentes dans l'entrée lexicale. Mais si on souhaite faire de la lemmatisation au sens strict (on ne désire que le lemme, ou que le lemme et la catégorie, sans information flexionnelle), le texte peut alors être étiqueté : soit {mon} {cache} dans le premier cas, soit {mon,.DET} {cache,.N} dans le second.

On n'y insistera pas ici, mais il apparaît clairement que les performances de la levée d'ambiguïtés dépendent non seulement, comme nous l'avons déjà dit (cf. n. 2, p. 6), de la complétude (ou plutôt, pour une grande performance, de l'incomplétude) des dictionnaires, mais aussi du degré de précision des étiquettes demandées.

Nous présentons maintenant une grammaire de deuxième passe :



Graphe7

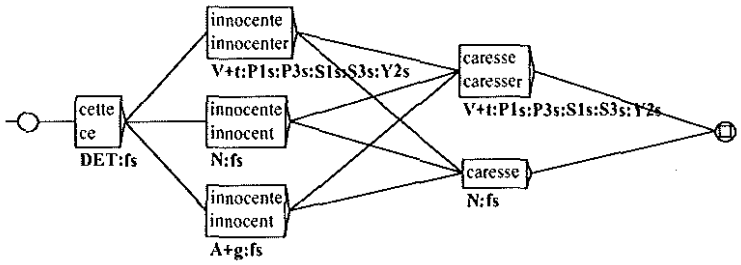
Graphe7 désambiguïse le contexte gauche de noms, de verbes ou d'adjectifs déjà désambiguïsés lors de l'application de la première passe de grammaires locales. *Le*, *l'*, *les*, *leurs*, *leur* et *ce*, devant un verbe, sont des pronoms. Tandis que placés devant un nom ou un adjectif, ils appartiennent à la catégorie déterminant. *La* devant un nom ou un adjectif est un déterminant. On ne peut tirer aucune conclusion lorsqu'il précède un verbe : il peut s'agir du pronom (*elle la donne*) ou du nom (*sur mon piano, le* la *sonne faux*). Quant à *son*, déterminant suivi d'un nom, on ne peut rien dire lorsqu'il est à gauche d'un adjectif : déterminant (*son beau chapeau*) ou nom (*un son agréable*).

Certaines des grammaires précisent également la place que doit occuper la séquence que l'on veut reconnaître : en début ou en fin de phrase, après un signe de ponctuation, etc. Tenir compte de la position dans la phrase permet de lever un nombre d'ambiguïtés non négligeable.

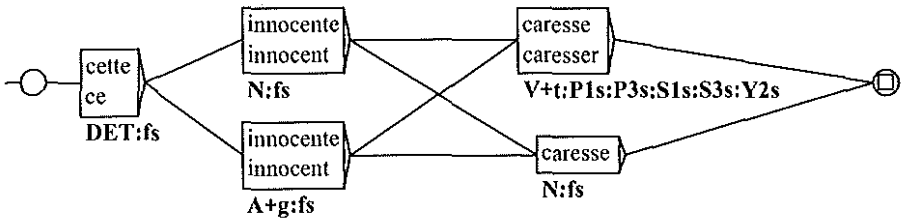
La levée d'ambiguïtés appliquée à l'automate du texte

Nous avons vu que *Graphe6*, qui s'appuie sur 20 déterminants non ambigus, permet d'étiqueter correctement des séquences comme *chaque dîner*, *ses affiches*, *ces portions*, etc., dans lesquelles le mot qui suit le déterminant est soit un verbe, soit autre chose (mais une seule hypothèse). Dès lors qu'il y avait plus de deux hypothèses pour ce mot, celui-ci n'était pas étiqueté. C'était le cas de *cette innocente*, *mon devant* et *vos caches*. Alors que dans ces trois séquences, la grammaire impose au second mot de n'être pas un verbe, aucune trace visible de cette désambiguïsation n'apparaît dans le texte lors de l'étiquetage linéaire.

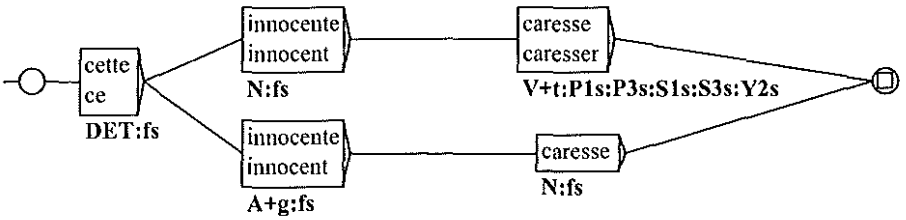
Or, l'une des possibilités offerte par le système Intex est d'obtenir les résultats de la désambiguïsation sous la forme d'un transducteur. Voici le transducteur de la séquence *cette innocente caresse* sans levée d'ambiguïtés :



Voici maintenant le transducteur de cette même séquence, mais après application de *Graphe6* :

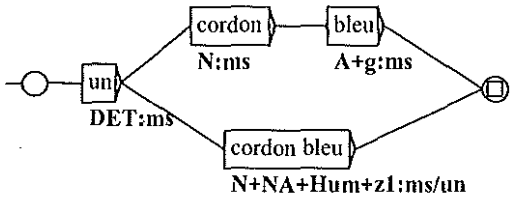


On le voit, le nœud `innocente/<V>` a été supprimé, ne laissant plus que les deux hypothèses « nom » et « adjectif » pour le mot. Une désambiguïsation encore plus poussée devrait permettre d'arriver au transducteur suivant :



On trouve alors parallèlement les 2 chemins `DET_N_V` et `DET_A_N`. Ces analyses correspondent respectivement à *cette innocente caresse la statue du Christ* et *il s'offusque de cette innocente caresse*.

Travailler directement sur le transducteur du texte offre aussi le grand avantage de permettre une levée d'ambiguïtés qui tient compte des mots composés. Or, la plupart des mots composés peuvent également être analysés comme une séquence libre : *un cordon bleu* est un cordon de couleur bleue ou une bonne cuisinière. Pour trancher face à ce type d'ambiguïtés, il faut bien souvent des connaissances sémantiques ou pragmatiques que n'ont évidemment pas les systèmes d'étiquetage. Lors d'un étiquetage linéaire, soit on considère que ce type de séquence n'est jamais figé (ce qui est faux), soit qu'il l'est toujours (ce qui est tout aussi faux). Une représentation du texte par transducteurs offre alors la possibilité d'envisager les 2 séquences, représentées dans des chemins parallèles, l'un pour la suite libre, l'autre pour le figement.



Mais outre ce problème où pour la même séquence on hésite entre figement et suite libre, il faut encore mentionner les innombrables cas où le mot composé est formellement ambigu. Ainsi, *gardes du corps* est le pluriel du nom composé *garde du corps* mais aussi un segment de la phrase *ce que tu gardes du corps de ton texte est négligeable*[1]. Là aussi, l'ambiguïté est représentée par des chemins parallèles dans le transducteur du texte.

[1] Ce petit extrait surréaliste pousse le problème de découpage à l'extrême : « Tout tourne à l'eau de boudin quand le pire est évité de justesse. Mais celui qui tombe dans un sac de nœuds coulants tire le diable par la queue de cerise sur le gâteau d'anniversaire de la libération de la femme de chambre à coucher » (Cédrick Fairon).

Conclusions

L'originalité des grammaires de levée d'ambiguïtés d'Intex est double : d'une part, leur caractère « parfait » les distingue de la plupart des systèmes d'étiquetage : pas de désambiguïsation totale, mais une désambiguïsation correcte. D'autre part, représenter le résultat de la levée d'ambiguïtés sous la forme de transducteurs offre l'avantage de conserver ce caractère parfait en tenant compte des mots composés. Un système qui ne tiendrait pas compte des mots composés, qui analyserait indépendamment chacune des formes composant *à la fin de*, *grâce à*, *tout à coup*, etc., passerait à côté d'une des aspects prépondérants de la langue que sont les figements.

Bibliographie

COURTOIS (B.) : 1990, « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française*, n° 87, *Dictionnaires électroniques du français* (Paris : Larousse), p. 11–22.

COURTOIS (B.) : 1996, « Les formes ambiguës de la langue française », *Linguisticæ Investigationes*, 20, 1 (Amsterdam : John Benjamins), p. 167–202.

DISTER (A.) : 2000, « La levée d'ambiguïtés morphologiques par grammaires locales. Le système Intex », *Actes des XIII^es Journées de Linguistique, 25 et 26 mars 1999* (Québec) ; à paraître.

DISTER (A.) : 2000, « Construire des grammaires de levée d'ambiguïtés pour Intex », dans FAIRON (Cédrick) : 2000, *Analyse syntaxique et lexicale. Le système INTEX. Linguisticae Investigationes*, 22, (Amsterdam/Philadelphia : John Benjamins).

GROSS (M.) : 1975, *Méthodes en syntaxe* (Paris : Hermann).

GROSS (M.) : 1968, *Grammaire transformationnelle du français. 1) Syntaxe du verbe* (Paris : Cantilène).

GROSS (M.) : 1990, *Grammaire transformationnelle du français. 3) Syntaxe de l'adverbe* (Paris : ASSTRIL).

LAPORTE (É.) : 1995, « Levée d'ambiguïtés par grammaires locales », *Linguisticae Investigationes Supplementa* (Amsterdam-Philadelphia : John Benjamins), p. 97–114.

LAPORTE (É.), SILBERZTEIN (M.) : 1996, « Ambiguity rates. Automatic analysis of French text corpora and computation of ambiguity rates for different tagsets », in LAPORTE (É.) ed., *GRAMLEX Deliverables, October 1995 – June 1996* (Paris : LADL).

SILBERZTEIN (M.) : 1990, « Le dictionnaire électronique des mots composés », *Langue française*, n° 87, *Dictionnaires électroniques du français* (Paris : Larousse), p. 71–83.

SILBERZTEIN (M.) : 1993, *Dictionnaires électroniques et analyse automatique de textes, Le système INTEX* (Paris : Masson).

SILBERZTEIN (M.) : 1996, « Expérience d'étiquetage avec levée d'ambiguïtés », *Actes des Premières Journées INTEX, 21-22 mars 1996* (Paris : LADL, Université Paris 7, CNRS), p. 17–31.

SILBERZTEIN (M.) : 1997, « The lexical analysis of natural languages », in ROCHE (E.) and SCHABÈS (Y.) eds, *Finite-state language processing* (Cambridge, Mass. – London : MIT Press), p. 175–203.

VAN HALTEREN (H.) éd. : 1999, *Syntactic Wordclass Tagging* (Dordrecht : Kluwer Academic Publishers).