

Levée d'ambiguïtés sur les mots lexicaux et grammaticaux

ANNE DISTER
Université de Liège

Introduction

De nombreuses tâches dans le traitement automatique des langues naturelles se basent sur des corpus écrits préalablement traités. Il s'agit de partir d'un texte non plus « vierge », mais qui a déjà subi un certain nombre de « transformations » : on a isolé des séquences élémentaires de traitement (mots simples ou composés, syntagmes, phrases), on a rétabli des élisions et des contractions, on a regroupé les négations, etc. Bien souvent, on attend que le texte soit étiqueté : chaque mot (ou groupe de mots) est accompagné d'une étiquette, qui contient un certain nombre d'informations : la catégorie grammaticale, la forme canonique ou encore des informations de genre, de nombre, de flexion, etc.

C'est ce travail d'étiquetage, ou de désambiguïsation, que nous allons présenter ici. L'outil utilisé dans le cadre de notre travail est le logiciel INTEX, qui inclut les dictionnaires à large couverture construits au LADL. En appliquant ces dictionnaires au texte, on obtient pour chaque forme du texte la liste de toutes les possibilités présentes dans le dictionnaire. Il s'agit d'une certaine manière d'un étiquetage initial, mais où un mot du texte peut être accompagné de plusieurs étiquettes. C'est alors que nous intervenons : il s'agit, parmi toutes les hypothèses présentes, de choisir celle qui est pertinente en contexte. Pour le dire autrement, il faut éliminer les solutions parasites générées par la consultation des ressources lexicales — solutions parasites (ou ambiguïtés) dont on sait qu'elles ne sont que rarement effectives lors de la lecture humaine. Ainsi, l'étiquetage initial et la désambiguïsation apparaissent comme deux tâches distinctes, développées dans deux modules différents. Mais on le voit, le dictionnaire occupe une place prépondérante au départ du processus de désambiguïsation. C'est la raison pour laquelle nous allons dans un premier temps centrer notre réflexion sur la manière dont sont codées les ressources lexicales que nous utilisons, et son influence sur le concept même d'ambiguïté.

1. Le concept d'ambiguïté

1.1. Dans les dictionnaires

« Les ambiguïtés dépendent de la qualité et de la finesse des informations associées aux entrées de dictionnaire » (B. Courtois, 1996). Ainsi, le changement des conventions de codage du dictionnaire peut modifier considérablement le nombre des ambiguïtés.

1.1.1. Les formes verbales

Prenons l'exemple des formes verbales. Dans le dictionnaire des mots simples (delafm) d'INTEX (B. Courtois, 1990), toutes les informations sur la conjugaison d'une forme verbale sont regroupées sur la même ligne, constituant ainsi une seule entrée pour la forme¹ :

donne,donner.V:P1s:P3s:S1s:S3s:Y2s

¹ Nous parlons ici uniquement de la forme verbale. Il y a également une entrée pour le nom *donne*.

Cette entrée se lit comme suit : *donne*, dont le lemme est *donner*, est un verbe à la première personne du singulier (1s) de l'indicatif présent (P), à la troisième personne du singulier (3s) de l'indicatif présent, à la première personne du singulier du subjonctif présent (S), à la troisième personne du singulier du subjonctif présent ou à la deuxième personne du singulier (2s) de l'impératif (Y).

Mais on aurait pu imaginer, lors de la construction du dictionnaire, de n'attribuer qu'une seule information de conjugaison pour chaque entrée. Ainsi, la forme *donne* serait associée aux cinq entrées suivantes² :

donne,donner.V:P1s
donne,donner.V:P3s
donne,donner.V:S1s
donne,donner.V:S3s
donne,donner.V:Y2s

Selon ce principe de codage des entrées, tous les verbes du premier groupe verraient leurs entrées multipliées par cinq pour la terminaison *-e* ; et par deux pour les finales *-es*, *-ais* et *-ent*. Pour les verbes des autres groupes, les dédoublements seraient différents mais nombreux aussi.

Par ailleurs, la forme *donne* est associée au lemme — ou base ou forme canonique — *donner*, sans aucun autre type d'informations. Or, on le sait, il existe plusieurs verbes *donner*, qui ont chacun des comportements syntaxiques différents. Par exemple : *Max donne un lit à Ida* (table 9 du lexique-grammaire) ; *la fenêtre donne sur la cour* (table 33); etc. La prise en compte de ces informations syntaxiques multiplie alors les entrées de *donne*, pour l'associer à 12 verbes *donner* différents.

Si en plus on assigne à chacun de ces verbes une seule entrée par information de conjugaison, comme nous l'avons fait ci-dessus, nous obtenons alors 5 x 12 soit 60 entrées pour la seule forme verbale *donne*.

Dans la mesure où nous nous intéresserons uniquement aux ambiguïtés morphologiques, nous ne tiendrons pas compte de ces ambiguïtés qui relèvent de la syntaxe.

1.1.2. *Les adjectifs*

Pour la flexion des adjectifs, différentes possibilités de codage se présentent également. Voici les entrées pour les quatre flexions du lemme <intelligent> :

intelligent,intelligent.A:ms
intelligente,intelligent.A:fs
intelligentes,intelligent.A:fp
intelligents,intelligent.A:mp

Quatre formes différentes qui sont assez logiquement réparties dans quatre entrées lexicales.

Pour <docile> par contre, on ne rencontre plus que deux formes, l'une pour le singulier l'autre pour le pluriel, aucune marque de genre n'apparaissant dans la graphie. Le *delafm* a choisi d'en faire deux entrées et non quatre :

docile,docile.A:ms:fs

² C'est d'ailleurs avec cette présentation — une ligne pour chaque information de conjugaison — qu'est construit le dictionnaire électronique des mots simples de l'italien (voir S. Vietri et A. Elia 2000).

dociles,docile.A:mp:fp

De la même manière, l'adjectif *orange*, qui n'a lui aucune marque au pluriel, possède une seule entrée :

orange,orange.A:ms:fs:mp:fp

Ici aussi, on a choisi de rassembler toutes les informations morphologiques dans une seule entrée, et non d'engendrer quatre entrées différentes. Il en va donc de même pour un grand nombre d'adjectifs de couleur (*abricot, acajou, ambre, amarante*, etc.), mais aussi pour : *afro, antibruit, antichoc, rasoir, pop, puce, punk*, etc. En fait, tous les adjectifs invariables du dictionnaire delafm sont codés de cette façon. Choisir la solution de codage qui démultiplie les entrées donnerait, au lieu des 12 352 entrées actuelles, 25 824 entrées.

Cependant, les choix de codage que nous venons de mettre en évidence et qui réduisent le nombre d'ambiguïtés ne doivent pas donner une fausse idée du dictionnaire. Avec ses 675 249 entrées actuelles (INTEX, version 4.2), c'est une ressource lexicale très complète pour l'analyse automatique de textes. Ses descriptions lexicales très fines mettent en évidence des analyses ignorées — volontairement ou non — de la plupart des systèmes de traitement automatique. Citons entre autres :

- *par*, qui n'est pas seulement la préposition mais également un nom, terme de golf.

par,par.N:ms

par,par.PREP

- *nouvelles* qui est quatre fois ambigu :

nouvelles,nouveau.A:fp

nouvelles,nouveau.N:fp

nouvelles,nouvelle.N:fp

nouvelles,nouvelles.N:fp

Dans le dictionnaire des mots simples delafm, tel que nous venons de le décrire, plus de 12 % des entrées sont ambiguës, c'est-à-dire qu'elles sont homographes avec au moins une autre entrée (B. Courtois, 1996)³.

1.2. Les étiquettes dans le texte désambiguïsé

Si le concept d'ambiguïté dépend de la manière dont est conçu un dictionnaire, celle-ci est intimement liée au degré de précision attendu dans le résultat de la levée d'ambiguïtés. Par précision, nous entendons ici le nombre et le type d'informations souhaitées dans les étiquettes après désambiguïsation. Ainsi, les différents modes d'étiquetage linéaire proposés par le système INTEX donnent des résultats différents et parfois très surprenants.

Voici les quatre modes d'étiquetage possibles, en fonction de la finesse des informations souhaitées dans les étiquettes :

³ Dans cet article de 1996, Blandine Courtois s'intéresse également aux ambiguïtés en termes de catégorie grammaticale. Parmi les combinaisons plus fréquentes du delafm, elle donne (par ordre décroissant) :

Nom Humain / adjectif ; verbe au participe passé / adjectif ; nom masculin / adjectif ; verbe à l'indicatif présent / nom féminin ; nom féminin / adjectif ; verbe au participe passé / nom humain / adjectif ; verbe au participe présent / adjectif ; verbe à l'indicatif présent / nom masculin.

On trouvera encore dans cet article de nombreux exemples des différents types d'ambiguïtés, répartis dans les quatre grandes classes qu'elle isole : ambiguïtés entre deux formes verbales ; ambiguïtés entre une forme verbale et non verbale ; ambiguïtés entre une forme nominale et une forme non verbale ; ambiguïtés restantes.

a. l'entrée lexicale complète (*Lexical entry*) : la forme du texte est remplacée par la totalité de l'entrée lexicale, telle qu'elle est présente dans le dictionnaire.

b. le lemme et la catégorie grammaticale (*Lemma & Category*) : la forme du texte est remplacée par le lemme et la catégorie grammaticale uniquement. Il n'y a donc pas d'informations de mode, temps, personne et nombre pour les verbes ; pas d'informations de nombre, de personne et de genre pour les pronoms personnels ; pas d'informations de genre et de nombre pour les autres pronoms, les noms, les adjectifs et les déterminants.

c. la forme et la catégorie (*Form & Category*) : la forme du texte est accompagnée de sa catégorie grammaticale seule, sans les informations flexionnelles ou de conjugaison (cf. le mode précédent).

d. le lemme (*Lemma*) : la forme du texte est remplacée par le lemme seul.

Les quelques exemples qui suivent montreront aisément que la tâche de désambiguïsation part de données différentes (même si l'on se base sur le même dictionnaire comme c'est le cas ici) en fonction du mode choisi.

La forme *fin* a quatre entrées distinctes dans le delafm :

fin,fin.A:ms

fin,fin.ADV

fin,fin.N:fs

fin,fin.N:ms

Si l'on se place dans le mode d'étiquetage où les étiquettes correspondent à l'entrée lexicale — les étiquettes sont les plus complètes —, *fin* est quatre fois ambigu (puisque nous avons quatre entrées). À l'inverse, si l'on désire obtenir uniquement le lemme, la forme n'est plus ambiguë. Par contre, si au lemme s'adjoint la catégorie grammaticale, mode (b), on a alors trois ambiguïtés :

fin.A

fin.ADV

fin.N

On obtient le même résultat si l'on désire la forme et la catégorie, en mode (c).

On pourrait imaginer que, pour ces deux derniers modes, les informations flexionnelles accompagnent la catégorie grammaticale. Puisque les deux hypothèses pour le nom se distinguent par le genre, on aurait alors de nouveau, bien que l'on choisisse uniquement le lemme ou la forme, quatre ambiguïtés.

Comme nous l'avons vu ci-dessus, *nouvelles* a un degré d'ambiguïté 4 :

nouvelles,nouveau.A:fp

nouvelles,nouveau.N:fp

nouvelles,nouvelle.N:fp

nouvelles,nouvelles.N:fp

Quatre entrées, donc quatre ambiguïtés pour le mode *Lexical Entry*. Si l'on passe dans le mode (b), le lemme et la catégorie, on obtient trois hypothèses pour la forme :

nouveau.N

nouvelle.N

nouvelles.N

Dans le mode (c), forme et catégorie, on n'a plus alors que deux ambiguïtés :

nouvelles.A
nouvelles.N

Les deux exemples qui précèdent montrent que les modes (b) et (c) ne donnent pas nécessairement le même nombre d'étiquettes.

Les trois entrées pour la forme *abandonné* sont les suivantes :

abandonné,abandonner.V:Kms⁴
abandonné,abandonné.A:ms
abandonné,abandonné.N:ms

Trois ambiguïtés en mode (a), entrée lexicale, qui se réduisent à deux lorsque l'on ne considère que le lemme : *abandonner* et *abandonné*.

Restent toujours trois hypothèses pour le mode (b), lemme et catégorie :

abandonner.V
abandonné.A
abandonné.N

ainsi que pour le mode (c), forme et catégorie :

abandonné.V
abandonné.A
abandonné.N

Ce type d'ambiguïtés est particulièrement fréquent puisqu'on le retrouve dans les nombreux cas d'ambiguïtés nom, adjectif, participe passé : *accusé, acquis, agité, brisé, carré, cassé, déterminé, désespéré, fichu, retenue, retourné*, etc. Cette d'ambiguïté entre le nom et l'adjectif se rencontre aussi avec des formes verbales, notamment les première et troisième personnes du singulier (indicatif, subjonctif et impératif présent) : *absente, calme*, etc.

Tous ces exemples tendent à montrer très clairement que la nature même des données lexicales de base (la finesse des informations présentes dans le dictionnaire et les choix de codage), ainsi que la précision des résultats attendus, en termes de complétude des étiquettes, interviennent dans la définition du concept d'ambiguïté. Ainsi, quand nous parlerons dans la suite de cet article d'ambiguïté, c'est en nous basant

1. sur les ressources lexicales du dictionnaire des mots simples où ne sont pas séparés les verbes en fonction de leurs propriétés syntaxiques ;
2. sur le mode d'étiquetage qui donne comme résultat de la désambiguïsation les étiquettes les plus complètes.

Cela posé, une forme sera considérée comme ambiguë si elle possède plusieurs entrées dans le dictionnaire de référence, ici le dictionnaire des mots simples delafm.

2. Les ambiguïtés les plus importantes à lever

Le taux d'ambiguïtés de 12 % que nous donnions ci-dessus a été calculé par Blandine Courtois sur les entrées du dictionnaire. Il s'agit en quelque sorte de chiffres « théoriques »,

⁴ V:Kms = participe passé au masculin singulier.

calculés sur l'ensemble du vocabulaire recensé dans le dictionnaire. Une autre démarche pour comptabiliser les ambiguïtés consiste à se baser sur des données pragmatiques qui relèvent de comptages des occurrences de formes dans des textes.

En effet, « à partir d'un texte donné, l'homographie au niveau des formes graphiques (en tant que réalisations de lexèmes) est beaucoup plus fréquente que celle visible au niveau des vocables dans un dictionnaire » (S. Bolasco 1993) et cela, sans doute dans une large mesure parce que « [les ambiguïtés] concernent bien plus les mots fréquents que les mots rares » (M. Silberztein 1995).

Pour Sergio Bolasco, les homographies les plus intéressantes à lever se répartissent comme suit, en termes de catégorie grammaticale :

Nom Verbe	60 %
Nom Adjectif	16 %
Adjectif Verbe	11 %
Le reste	13 %

Il centre donc son objectif de levée d'ambiguïté sur les mots lexicaux.

Suivant une démarche comparable qui consiste à analyser les données d'un corpus, nous avons extrait les 100 chaînes de caractères les plus fréquentes dans une année du journal *Le Monde* (LM94) ⁵. Après avoir enlevé les caractères diacritiques de cette liste, nous avons regroupé les différentes graphies sous une même forme (*et* pour *et*, *ET*, *Et*).

En termes de formes, on constate que parmi les 70 occurrences les plus fréquentes, très peu participent des types d'ambiguïtés de catégories grammaticales mises en avant par Bolasco. Pour l'**ambiguïté nom / verbe**, nous avons *a*, *être*, *été*, *plus*, *président* et *A*. Pour l'**ambiguïté nom / adjectif**, on trouve uniquement la forme *bien*, qui apparaît également dans la locution conjonctive *bien que*. On trouve seulement deux formes qui possèdent la triple ambiguïté **nom / verbe / adjectif** : *est* et *fait*.

En fait, les formes les plus fréquentes sont principalement des mots grammaticaux. En tête de LM94, nous avons *de* (ambiguïté entre déterminant et préposition), *la* et *l* (nom, déterminant ou pronom), *le* et *les* (déterminant ou pronom).

On peut se demander dans quelle mesure il est intéressant de s'appliquer à lever l'ambiguïté des ces mots grammaticaux, considérés comme vides de sens. Cette tâche ne nous semble pas inutile car, comme nous aurons l'occasion de le voir dans le chapitre suivant, lever l'ambiguïté sur un mot grammatical permet bien souvent de lever l'ambiguïté sur la ou les forme(s) pleine(s) qui l'entourent.

Lorsque l'on parle des ambiguïtés les plus fréquentes, il est possible d'aborder le problème sous deux angles différents : celui de la catégorie grammaticale en général — comme c'est le cas de Bolasco —, ou celui des formes graphiques en particulier.

Dans les grammaires de désambiguïsation qui suivent, nous aborderons simultanément les deux approches mentionnées ci-dessus, l'une qui consiste à se baser sur des formes, et l'autre, traitant plutôt des catégories grammaticales, avec ou sans informations flexionnelles.

3. Désambiguïser grâce au contexte : les grammaires locales

Notre méthode de désambiguïsation se base sur le formalisme des automates à états finis (Silberztein 1997).

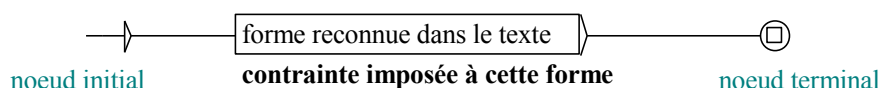
Il s'agit de décrire le contexte local, à gauche et/ou à droite, de la forme qui pose problème, afin de lever l'ambiguïté quant à son lemme — *camée* : *camer*, *camé* ou *camée* —,

⁵ Sur les homographies les plus fréquentes dans un corpus, avec des données chiffrées, voir A. Dister 1999.

sa catégorie grammaticale — *camée* : V, A ou N –, et ses informations flexionnelles — *moule* : ms ou fs. Pour le dire autrement, il faut choisir laquelle des entrées du dictionnaire correspond à la forme en contexte.

Ce contexte n'est pas limité à un empan défini ; néanmoins, il ne peut dépasser le cadre de la phrase (grosso modo, et en simplifiant beaucoup, d'une majuscule à un point, les problèmes d'ambiguïtés dus aux sigles, aux abréviations, etc. étant résolus dans une phase de prétraitement⁶).

Les automates sont représentés sous la forme de graphes, construits avec l'éditeur de graphes d'INTEX. Un graphe se lit de gauche à droite, partant du nœud initial pour arriver au nœud terminal. À l'intérieur des boîtes (ou nœuds) se trouvent les formes que l'on reconnaît dans le texte ; sous les boîtes, les contraintes que l'on impose à ces formes préalablement reconnues.



Comme nous l'avons dit, on peut aussi se baser sur des catégories grammaticales. Mettre <N> (pour la catégorie nom) dans une boîte signifie que l'on reconnaît dans le texte une forme qui peut être un nom, même si elle peut également posséder une autre catégorie dans le dictionnaire, et que c'est cette autre catégorie qui est éventuellement actualisée dans le texte. C'est la contrainte sous la boîte <N> qui assignera l'entrée lexicale dans laquelle la forme est un nom à l'occurrence repérée dans le texte.

De la même manière, on peut travailler sur des lemmes (notés entre crochets), accompagnés ou non d'informations flexionnelles. <être> reconnaît n'importe quelle forme du verbe *être*. Mais la forme du verbe *être* reconnue peut être homographe avec une autre forme : *est* adjectif <est> dans « l'est de la Belgique » ; *suis* verbe <suivre> dans « je suis le chemin indiqué » ; *sommes* nom <somme> dans « de grosses sommes d'argent » ; *soit* conjonction de coordination <soit> dans « soit l'un, soit l'autre » ; *fût* dans « un fût de cognac » ; etc. Là aussi, c'est la contrainte <V> sous la boîte (ou <être>) et la description d'un contexte non ambigu qui étiquetteront la forme reconnue comme une occurrence du verbe *être*.

Nous allons maintenant nous pencher en détail sur quelques grammaires locales qui lèvent l'ambiguïté entre déterminant et pronom.

3.1. L'ambiguïté entre déterminant et pronom

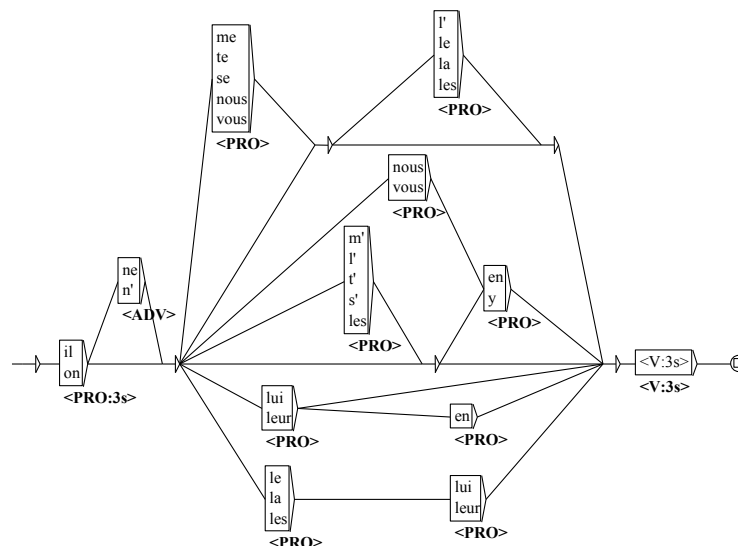
En effet, dès lors que l'on s'attache aux formes et non aux catégories, on constate que parmi les plus fréquentes des formes d'un corpus se trouvent celles ambiguës entre déterminant et pronom personnel : *la*, *l'*, *le*, *les*, *ce*, *le*, *les* et *leur*. Il faut noter que la forme reprise dans le tableau ci-dessus était *l* et non *l'*, et possédait trois hypothèses : DET PRO et N. La présence de l'apostrophe permet d'emblée de réduire ces hypothèses aux deux DET et PRO.

Une première manière de lever l'ambiguïté de ces formes est de décrire leur contexte d'apparition dans les séquences où ce sont des pronoms préverbaux, comme dans : *je l'acquis*, *je la porte*, *il l'interdit*, *ils les parent*, etc.

Voici un exemple de ce type de grammaire, très simple⁷ :

⁶ Voir A. Dister (1998), M. Silberstein (1998) et N. Friburger, A. Dister et D. Maurel (2000).

⁷ Certaines de nos grammaires s'inspirent de celles construites par M. Gross et M. Silberstein. Pour le graphe 1, voir M. Gross (1989).



Graphe 1

Le graphe 1 reconnaît et analyse des séquences comme *il le donne, il lui donne, il ne le donne, il ne le lui donne, il ne lui donne, il le lui donne, il le donnait, il lui donnait, il ne le donnait, il ne le lui donnait, il ne lui donnait, il le lui donnait*, etc.

Le graphe 1 décrit une séquence relativement contrainte ; il a l'avantage de se baser sur des pronoms sujets non ambigus. Dans cette grammaire, toutes les formes présentes entre le sujet *il* ou *on* et le verbe sont désambiguïsées. Mais de plus, l'ambiguïté éventuelle sur le verbe est elle aussi levée. Ainsi, dans les phrases que nous donnions ci-dessus, *porte* et *interdit* étaient ambigus :

interdit,interdire.V:Kms:P3s:J3s
interdit,interdit.A:ms
interdit,interdit.N:ms

porte,porte.A:ms:fs
porte,porte.N:fs
porte,porteur.V:P1s:P3s:S1s:S3s:Y2s

et sont maintenant étiquetés :

{il,.PRO+z1:3ms} {la,le.PRO+z1:3fs} {porte,porteur.V+z1:P3s:S3s}
{il,.PRO+z1:3ms} {l,le.PRO+z1:3ms:3fs}' {interdit,interdire.V+z1:P3s:J3s}

Sur le modèle du graphe 1 nous avons construit des grammaires comparables avec les pronoms sujets non ambigus *je, j' et ils*.

Nous avons pris certaines précautions pour les autres pronoms sujets *elle, nous, vous et elles* qui peuvent également être des pronoms compléments. En fait, si nous avions appliqué pour *elle* une grammaire trop étroitement calquée sur le graphe 1, nous aurions étiqueté erronément les phrases suivantes (extraites de LM94) :

Et puis elle revenait chez elle entre deux repas, s'embaillant dans un manteau (...)
Mais il y avait chez elle l'envie de sortir de la classe, de faire au (...)
Il est en droit d'attendre d'elle écoute et respect.
Entre-temps, il y eut pour elle la rencontre avec un professeur de philosophie
(...) combattre une violence qui portait en elle la remise en cause de ses privilèges.

Ces cas constituent à l'évidence une très faible minorité d'occurrences. Mais dans l'optique où l'on souhaite une désambiguïsation correcte, il est essentiel de les prendre en considération lors de l'écriture d'une grammaire.

Par ailleurs, une telle grammaire, quoique conçue pour décrire une séquence comprise entre un pronom sujet et le verbe accompagnant ce pronom, reconnaîtrait un autre type de segment ayant d'autres fonctions syntaxiques et l'étiquetterait correctement :

Tout chez elle remonte à l'enfance, une sorte de subconscient (...).

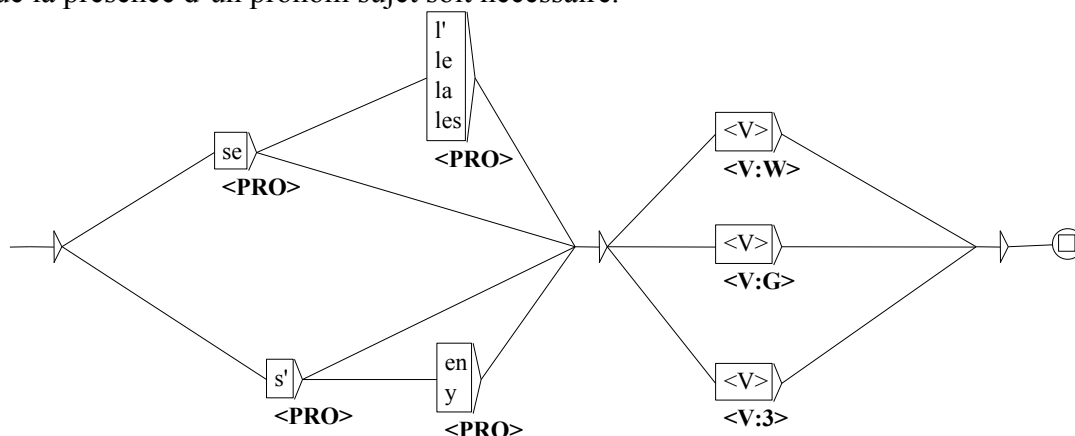
Cette squadra azzura si sûre d'elle aura servi d'aiguillon et de modèle à l'équipe (...).

(...) avec la menace communiste, voter pour elle était un engagement nécessaire.

A priori, cela ne semble pas gênant puisque l'analyse morphologique n'est pas faussée. Mais cela implique alors de bien garder présent à l'esprit que ce type de grammaire, dont on pense qu'il décrit une séquence commençant par un pronom personnel sujet, ne peut être utilisé sans précautions dans le cadre de l'analyse syntaxique.

Dans la mesure où l'on ne s'occupe ici que de désambiguïsation d'homographes et que nos annotations n'ont pas d'informations syntaxiques proprement dites, le problème ne se pose pas. Notons que cela ne serait pas le cas si le dictionnaire que nous utilisons distinguait deux pronoms *elle*, l'un préverbal (PRO+PPV) et l'autre non.

Voici une autre grammaire grâce à laquelle on peut lever l'ambiguïté de *l'*, *le*, *la* et *les* sans que la présence d'un pronom sujet soit nécessaire.



Graphe 2

Se est non ambigu ; *s'* est ambigu, mais n'est une conjonction de coordination que lorsqu'il est suivi de *il* ou de *ils*. Dans tous les autres cas, c'est un pronom. La séquence qui suit ces formes est relativement contrainte : un verbe à l'infinitif, un verbe au gérondif ou un verbe à la troisième personne du singulier ou du pluriel. Entre les deux, *l'*, *le*, *la* et *les*, ou *y* et *en*, désambiguïsés également.

Ici encore, on lève l'ambiguïté sur le pronom et éventuellement sur la forme verbale, sans qu'il soit nécessaire de définir davantage le contexte.

La grammaire qui suit, décrite dans le graphe 3, permet de lever l'ambiguïté des formes *le* et *l'* lorsqu'elles suivent les prépositions *à* ou *de*.



Graphe 3

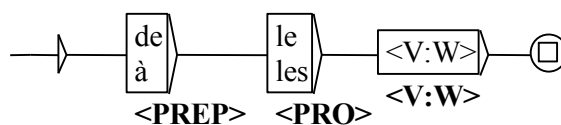
En effet, après les prépositions *à* et *de*, *le* et *les* sont des pronoms. Lorsqu'ils sont déterminants, *le* et *les* ne se rencontrent dans cette position que sous la forme contractée avec la préposition *au*, *aux*, *du* ou *des* (codé PREPDET pour "préposition déterminant") :

*à le chat mais au chat ; *de le jardinier mais du jardinier ; *à les enfants mais aux enfants ;
 *de les femmes mais des femmes.

Une grammaire semblable ne peut être construite pour *la* et *l'* qui ne se contractent pas. On a ainsi des séquences à *la*, *de la*, à *l'* ou *de l'* dans lesquelles *la* et *l'* sont soit déterminant soit pronom :

à l'école ; de la bière ; de la regarder ; à la chercher ; etc.

On peut encore aller plus loin et élargir la séquence à droite du pronom, où une forme lisible a priori comme verbe à l'infinitif est effectivement un verbe à l'infinitif, et non un nom ou un adjectif :



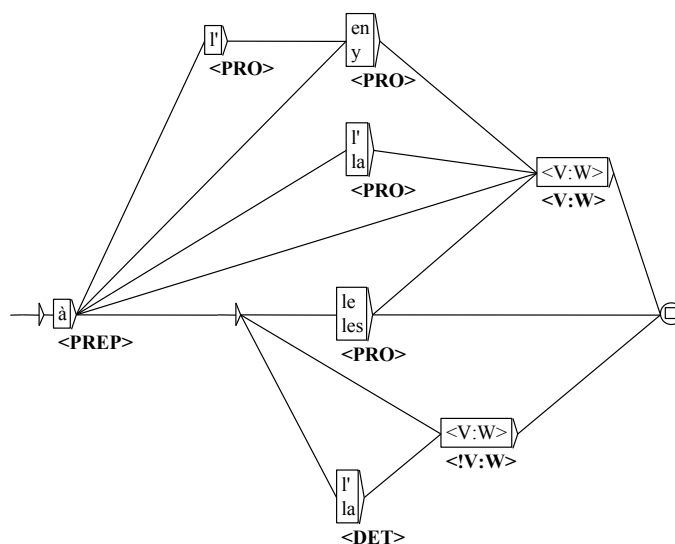
Graphe 4

Dans cette position, cela permet de lever l'ambiguïté sur un certain nombre de formes de verbes à l'infinitif homographes avec des noms et/ou des adjectifs. La plupart sont homophones : *le conseiller*, *le toucher*, *le goûter*, *le lever*, *le boulanger*, *le foutre*, *le lancer*, etc. ; certaines ne le sont pas : *le boxer*, *le scanner*, etc.

Dans le cas où l'on aurait affaire au nom ou à l'adjectif et non au verbe à l'infinitif, le déterminant serait contracté : *du justicier*, *du passager*, *du rappeur*, etc.

Même en élargissant le contexte de la sorte, les formes *la* et *l'* ne peuvent toujours pas être soumises aux mêmes contraintes que *le* et *les* : en effet, on peut rencontrer (*de+à*) (*la+l'*) suivi d'une forme à l'infinitif sans que cette séquence doive être analysée comme une séquence PREP PRO V:W, et ce toujours à cause de la non-contraction et de l'ambiguïté des formes infinitives : *l'aller*, *l'avoir*, *l'être*, *l'écailler*, *l'oranger*, *la tendre*, *la lire*, *la reporter*, etc.

Le graphe 5 est le graphe général qui traite ce type de séquences.



Graphe 5

Lorsqu'il n'y a pas de pronom, la forme qui suit immédiatement la préposition a de fortes chances d'être un verbe à l'infinitif (<V:W>/<V:W>); néanmoins, il peut également ne pas s'agir d'un verbe à l'infinitif (<V:W>/<!V:W>, qui doit être lu : « une forme étiquetable a

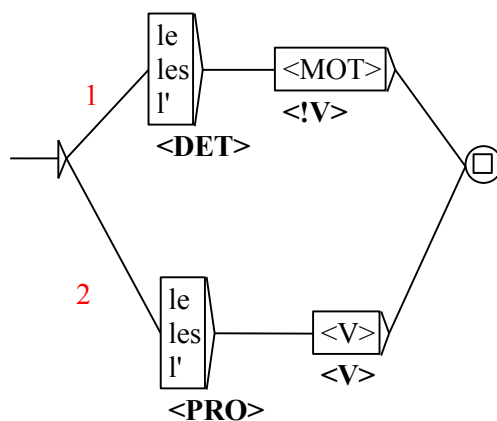
priori comme verbe à l’infinitif n’est pas un verbe à l’infinitif »)⁸. On ne peut en effet exclure l’hypothèse adjectif ou nom, comme dans :

une église à clocher gothique

Pour la grammaire comparable avec *de* — que nous dissociions de celle de *à* pour des questions de lisibilité, notamment en raison de l’élision en *d’* —, les exemples sont très nombreux : *une place de conseiller* ; *mon travail de boulanger* ; *elle va de dîner en dîner* ; *de fier qu’il était il est devenu plus chaleureux* ; etc.

On le voit assez clairement dans tous les exemples qui précèdent, lever l’ambiguïté sur la forme pronom ou déterminant nous permet également de lever l’ambiguïté verbe, nom ou adjectif de la forme qui suit.

Sous INTEX, on formalisera cela de la manière suivante :



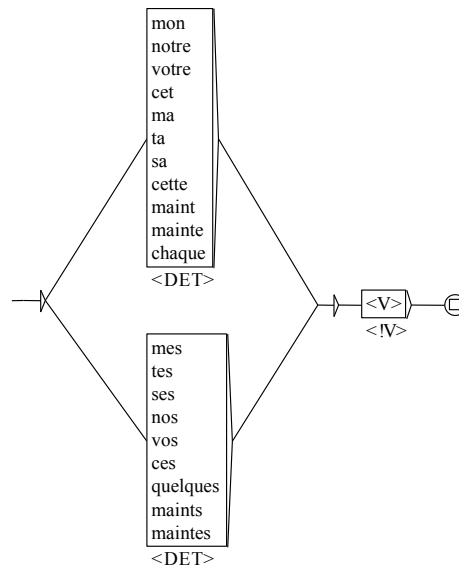
Graphe 6

Le chemin 1 dit que les formes *le*, *les* et *l’* sont des déterminants et que le mot qui suit n’est pas un verbe ; le chemin 2 impose à *le*, *les* et *l’* d’être des pronoms et à la forme verbale qui suit d’être effectivement un verbe. Le fait que les deux chemins figurent en parallèle permet de conserver les deux hypothèses.

3.2. L’ambiguïté entre nom, adjectif et verbe

Certaines de nos grammaires permettent de lever l’ambiguïté sur les trois grandes catégories nom, adjectif et verbe en se basant sur une liste de déterminants non ambigus. Ainsi, le graphe 7 :

⁸ Le point d’exclamation exprime la contrainte « n’est pas ». Ici, les deux chemins parallèles ont la même longueur et commencent au même endroit (à la forme *à*). Dans le cas où l’on rencontre dans le texte une forme à l’infinitif qui suit *à*, les deux hypothèses (infinitif et non-infinitif) seront conservées. Cf. le graphe 6.



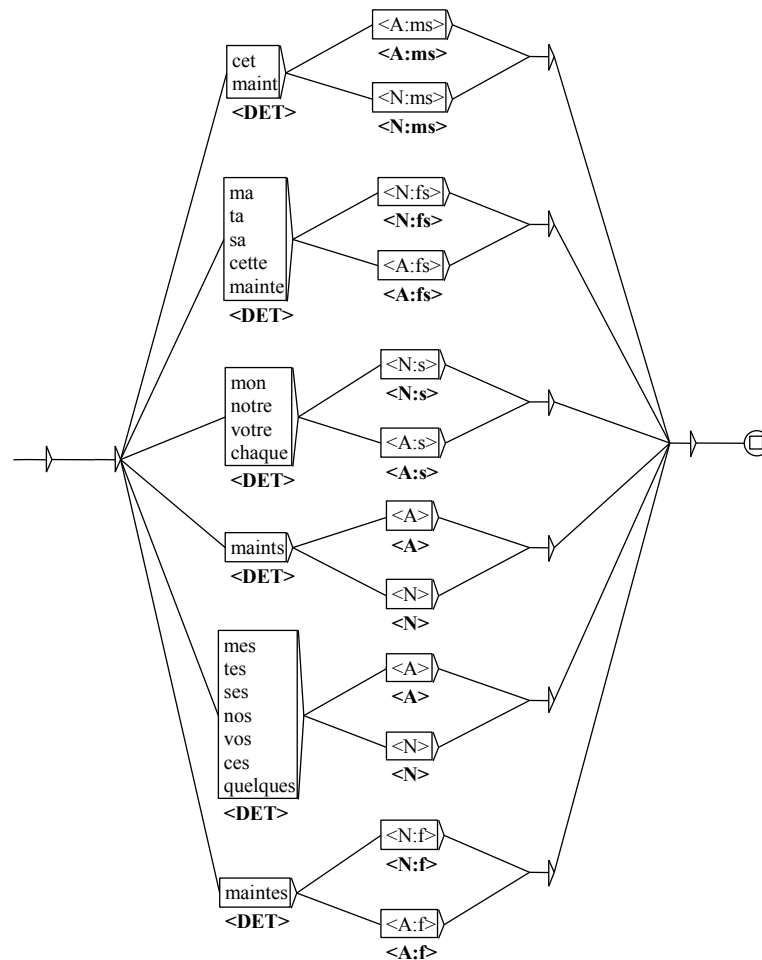
Graphe 7

La forme qui suit l'un des déterminants repris dans cette liste, et regroupés dans deux boîtes selon leur nombre, ne peut être un verbe (<V>).

Nous allons développer davantage et ne pas utiliser une règle négative (« n'est pas un verbe »). Il s'agit alors de donner explicitement la catégorie de la forme, et l'on pense aux possibilités nom ou adjectif. Auxquelles devrait s'ajouter une troisième hypothèse : adverbe, comme dans :

(...) chargé de remettre le pied à l'étrier à ces encore mal nommés (...)
 (...) de ces éternellement dangereux confins maroco-algériens (...)
 (...) il me semble que ce serait mon plus beau titre de gloire.

Nous n'indiquons pas ici l'hypothèse adverbe afin de faciliter la lecture du graphe. En revanche, nous rendons la grammaire plus explicite, ce qui permettrait notamment de lever l'ambiguïté quant au genre du nom ou de l'adjectif.



Graphe 8

Ici, les déterminants sont ventilés dans 6 boîtes, selon le double critère de genre et de nombre⁹.

Avec les formes *cet*, *maint*, *ma*, *ta*, *sa*, *cette* et *mainte*, l'adjectif ou le nom est au singulier, masculin avec les deux premières formes, féminin avec les autres. Avec les formes *mon*, *votre* et *chaque*, l'adjectif ou le nom qui suit peut être aussi bien masculin que féminin.

Pour les déterminants pluriels par contre, on ne peut rien dire du nombre du mot qui suit. En effet, « dans le cas, relativement rare, où un seul déterminant sert pour plusieurs noms, l'accord se fait comme pour l'épithète : le déterminant se met d'ordinaire au pluriel ; le genre est celui des noms, si ceux-ci ont le même genre ; sinon, le déterminant est au masculin » (A. Goosse 1993, § 562). On peut citer l'exemple suivant :

Ces samedi 30 et dimanche 31 juillet

Par ailleurs, développer la grammaire de la sorte permet de désambiguïser la forme qui suit alors que le graphe 7 ne le permettait pas. C'est le cas par exemple pour les séquences comme *cette auxiliaire*, *mon fils*, *ma pub* ou *votre vers*, où l'on ne peut lever l'ambiguïté que grâce au genre ou au nombre du déterminant. Voici les entrées du delafm pour ces formes :

auxiliaire,auxiliaire.A:ms:fs

⁹ On peut se contenter de noter simplement <DET> sous les boîtes, sans aucune information flexionnelle dans la mesure où ces déterminants sont non ambigus. En effet, les contraintes peuvent être moins complètes que les informations présentes dans le dictionnaire.

auxiliaire,auxiliaire.N:ms
auxiliaire,auxiliaire.N:ms:fs
fils,fil.N:mp
fils,fils.N:ms:mp
pub,pub.N:fs
pub,pub.N:ms
vers,ver.N:mp
vers,vers.N:ms:mp

Voici les séquences étiquetées après application de la grammaire :

{mon,.DET:ms} {fils,.N:ms}
{ma,mon.DET:fs} {pub,.N:fs}
{votre,.DET:ms} {vers,.N:ms}

Par contre, on ne peut pas lever toutes les ambiguïtés sur *auxiliaire*, qui peut être nom ou adjectif féminin¹⁰.

{cette,ce.DET:fs} auxiliaire

De plus, il est alors possible d’attribuer à un grand nombre de formes, non ambiguës selon notre critère d’une seule entrée lexicale — qui peut paraître bien insatisfaisant dans de nombreux cas —, une information flexionnelle plus précise. Ainsi, des formes homographes

au singulier et au pluriel

brebis,brebis.N:fs:fp
cas,cas.N:ms:mp
choix,choix.N:ms:mp

au masculin et au féminin

athlète,athlète.N:ms:fs
cinéaste,cinéaste.N:ms:fs

aux deux genres et nombres ensemble

albinos,albinos.A:ms:fs:mp:fp
albinos,albinos.N:ms:fs:mp:fp

mais rattachées au même lemme, constituent une seule entrée lexicale. Avec le graphe 7, les étiquettes dans le texte désambiguïsé avaient donc la forme suivante :

ma{brebis,brebis.N:fs:fp} ; mes{cas,cas.N:ms:mp} ; cette{athlète,athlète.N:ms:fs} ;
cet{albinos,albinos.A:ms:fs:mp:fp}.

¹⁰ Notons néanmoins que si l’on souhaite obtenir le résultat non plus sous forme d’étiquettes dans un texte linéaire mais sous la forme d’un transducteur (ayant donc le même format que les grammaires de désambiguïsation), on observe qu’une partie de l’ambiguïté est levée puisque l’hypothèse <auxiliaire,auxiliaire.N:ms> a été éliminée des chemins. Pour plus de détails, voir A. Dister 1999. Pour un système comparable, qui élimine des chemins dans le transducteur du texte, voir É. Laporte et A. Monceaux 1999.

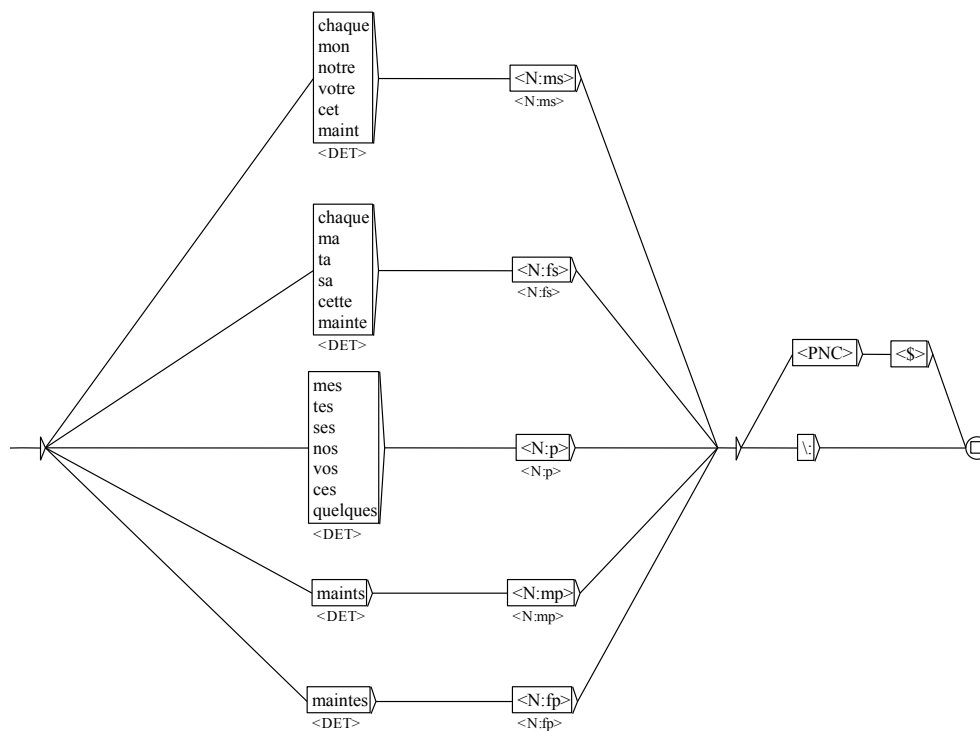
Avec le graphe 8, nous avons maintenant :

ma{brebis,brebis.N:fs} ; mes{cas,cas.N:mp} ; cette{athlète,athlète.N:fs} ; maint{cinéaste,cinéaste.N:ms}

qui est d'une certaine manière plus satisfaisant.

3.3. Lever l'ambiguïté entre adjectif et nom en fin de phrase

Le graphe 9 permet quant à lui de lever l'ambiguïté entre nom et adjectif, toujours en se basant sur la même liste de déterminants non ambigus.



Graphe 9

Ici, une contrainte supplémentaire est posée : la séquence reconnue doit se trouver en fin de phrase <PNC> <\$>. <PNC> représente une ponctuation et <\$> le symbole de fin de phrase inséré par INTEX dans le texte lors d'une phase de prétraitement (avant l'application des dictionnaires). Le double point n'étant pas séparateur de phrases dans cette phase de prétraitement, nous l'ajoutons dans le chemin du graphe.

Le graphe 7 indiquait qu'un verbe ne pouvait suivre un déterminant. Dans une séquence en fin de phrase, on peut aller plus loin puisque le mot qui suit le déterminant ne peut être qu'un nom. Ici, l'éventuelle ambiguïté entre nom et adjectif est résolue. Voici quelques occurrences issues de *La Femme de trente ans* de Balzac. Le mot qui suit le déterminant est aussi codé comme adjectif dans le delafm :

(...) le coude que la route fait en cet endroit.

(...) nous prescrivont de témoigner à nos proches.

(...) la Société fait, à tort ou à raison, sa morale.

On le voit, partant de la même liste de déterminants, il a été possible d'affiner de plus en plus l'analyse et de contraindre davantage le contexte afin d'améliorer la désambiguïsation.

4. Conclusion

Nous avons montré que le concept d'ambiguïté était variable. Dans notre méthode de désambiguïsation, il est doublement lié : à la constitution du dictionnaire et au choix de codage des informations d'une part ; au résultat attendu lors de la désambiguïsation et notamment à la finesse associée aux informations présentes dans les étiquettes d'autre part. Nous avons vu que modifier l'un ou l'autre de ces paramètres peut parfois changer considérablement le décompte des ambiguïtés et partant, la tâche de désambiguïsation (notamment, en termes de résultats obtenus).

Par ailleurs, nous avons vu que les grammaires de désambiguïsation construites avec INTEX offraient l'avantage de se baser aussi bien sur des catégories grammaticales en général que sur des formes en particulier. Désambiguïser une forme en particulier (*le* ou *les*) peut avoir des répercussions en termes de catégories grammaticales (ici, les trois grandes catégories qui concentrent la majorité des ambiguïtés : verbe, nom ou adjectif). Affiner la description dans la grammaire, en contraignant davantage le contexte ou en explicitant les traits flexionnels par exemple, permet d'améliorer le taux de désambiguïsation. Cette méthode simple, qui consiste à analyser le contexte local de la forme à désambiguïser, n'en demande pas moins la plus grande prudence. On l'a vu, on ne peut traiter toutes les formes appartenant à une même catégorie, même finie, de la même manière (*la* n'est pas *le*, *elle* n'est pas *il*). Pour intuitives qu'elles sont, ces grammaires demandent à être sans cesse testées, notamment lorsque la bibliothèque de graphes s'enrichit.

Anne Dister
Université de Liège
Quai Roosevelt 1b (Bât. A4)
4000 Liège
Belgique
adister@ulg.ac.be

Références

- Bolasco, Sergio. 1993. Choix de lemmatisation en vue des reconstructions syntagmatiques du texte par l'analyse des correspondances. In *Actes des Secondes Journées Internationales d'Analyse Statistique de Données Textuelles (JADT 1993)*, Paris : Telecom Paris.
- Courtois, Blandine. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, 87, *Dictionnaires électroniques du français*, Paris : Larousse, pp. 11-22.
- Boons, Jean-Paul ; Alain Guillet ; Christian Leclère. 1976. *La structure des phrases simples en français. Constructions intransitives*. Genève : Droz.
- Courtois, Blandine. 1996. Les formes ambiguës de la langue française. *Lingvisticæ Investigationes* XX:1, Amsterdam : John Benjamins, pp. 167-202.
- Dister, Anne. 1998. Problématique des fins de phrase en traitement automatique du français. In *À qui appartient la ponctuation ? Actes du colloque international et interdisciplinaire de Liège (13-15 mars 1997)*, Bruxelles : Duculot (Champs linguistiques), pp. 437-447.
- Dister, Anne. 1999. De l'étiquetage traditionnel au transducteur du texte. La levée d'ambiguïtés par grammaires locales. *RISSH, Revue Informatique et Statistiques dans les Sciences Humaines* 35, Liège : CIPL, pp. 9-24.
- Dister, Anne. 2000. Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes. In *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles*, Lausanne, pp. 131-138.

- Dister, Anne. 1999. Construire des grammaires de levée d'ambiguïtés pour INTEX. *Analyse syntaxique et lexicale. Le système INTEX*, C. Fairon (ed.), *Lingvisticae Investigationes*, Amsterdam/Philadelphia : John Benjamins, pp. 231-245.
- Friburger, Nathalie ; Anne Dister ; Denis Maurel. A paraître. Améliorer la reconnaissance automatique des fins de phrases. In *Actes des troisièmes journées INTEX (Liège, 13-14 juin 2000)*, dans *Revue, Informatique et Statistiques dans les sciences humaines* 36, Liège : CIPL.
- Goosse, André. 1993. *Le Bon usage*. Louvain-la-Neuve : Duculot.
- Gross, Maurice. 1975. *Méthodes en syntaxe*. Paris : Hermann.
- Gross, Maurice. 1989. The use of finite automata in the lexical representation of natural language. In *Lecture Notes in Computer Science*, M. Gross & D. Perrin. Springer, pp. 34-50.
- Laporte, Éric. 1994. Levée d'ambiguïtés par grammaires locales. *Lingvisticae Investigationes Supplementa*, Amsterdam/Philadelphia : John Benjamins, pp. 97-114.
- Laporte, Éric ; Anne Monceaux. 1999. Elimination of lexical ambiguities by grammars : the *ELAG* system. *Lingvisticae Investigationes XXII*, Amsterdam/Philadelphia : John Benjamins, pp. 341-367.
- Pierrel, Jean-Marie (ed.). 2000. *Ingénierie des langues*. Paris : Hermès.
- Senellart, Jean. 1999. *Outils de reconnaissance d'expressions linguistiques complexes dans de grands corpus*. Thèse non publiée.
- Silberztein, Max. 1993. *Dictionnaires électroniques et analyse automatique de textes, Le système INTEX*. Paris : Masson.
- Silberztein, Max. 1995. Dictionnaires électroniques et comptage de mots. In *Actes des Troisièmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1995)*, Rome.
- Silberztein, Max. 1996. Expérience d'étiquetage avec levée d'ambiguïtés. In *Actes des Premières Journées INTEX (21-22 mars 1996)*, Paris : LADL, Université Paris 7, CNRS.
- Silberztein, Max. 1997. The lexical analysis of natural languages. In *Finite-state language processing*, E. Roche, Yves Schabès (eds.), Cambridge, Mass./London : MIT Press.
- Silberztein, Max. 1998. Normalisation des textes. In *Actes des Troisièmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1998)*, Nice, pp. 601-614.
- Van Halteren, H. (ed.). 1999. *Syntactic Wordclass Tagging*. Dordrecht : Kluwer Academic Publishers.
- Vietri, Simonetta ; Annibale Elia. 2000. Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora. In *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles*, Lausanne, pp. 179-188.