Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes

Anne Dister

Université de Liège – 1b, Quai Roosevelt, B-4000 Liège – Belgique

Abstract

Many words are ambiguous in their part of speech. We propose to see how it is possible to disambiguate the many frequent words in big corpus (one year of the newspaper *Le Monde*) by the analyse of their local context described in grammars. We will see this method is very efficient, but unrelevant for some forms. In the second part of this paper, we will see the importance of the compound words and the necessity to take them into account for tagging.

Résumé

Beaucoup de mots présentent une ambiguïté de catégorie grammaticale. Après avoir identifié les 100 formes les plus fréquentes dans un gros corpus (une année du journal *Le Monde*), nous verrons comment il est possible de les désambiguïser en analysant leur contexte dans des grammaires locales. Nous verrons que cette méthode est très rentable, mais inefficace pour certaines formes. Dans une seconde partie, nous verrons l'importance des mots composés et la nécessité de les prendre en compte pour la désambiguïsation et notamment l'étiquetage de textes.

Mots-clés : ambiguïté - désambiguïsation - lemmatisation - étiquetage morphosyntaxique – mot composé - Intex

1. Introduction

Les traitements automatiques de la langue sont confrontés au problème de l'ambiguïté, qu'elle soit sémantique, syntaxique ou lexicale. Nous nous penchons ici sur les ambiguïtés morphologiques, qui doivent être levées pour l'étiquetage morphosyntaxique d'un texte ou encore des statistiques sur les lemmes.

Nous aborderons plus particulièrement les homographies les plus fréquentes dans les textes, en nous basant sur une année du journal *Le Monde* (LM94), et nous verrons une méthode simple pour les résoudre.

2. Choix méthodologiques

Intex[1] est un environnement de développement linguistique qui comprend des dictionnaires à large couverture et permet de traiter de très grands corpus, de plusieurs dizaines de millions de mots. Dans ce système, une forme rencontrée dans un texte sera considérée comme ambiguë si elle correspond à plusieurs entrées du dictionnaire[2]. Voici les deux entrées pour la forme *fier*, qui peut-être soit un adjectif au masculin singulier, soit un verbe à l'infinitif :

fier,fier.A:ms

[1] Voir Silberztein (1993).

[2] Pour plus de détails, voir Courtois (1996).

fier,fier.V:W

La constitution du dictionnaire occupe donc une place prépondérante dans la définition même de l'ambiguïté. Blandine Courtois a calculé que plus de 25 % des 665 000 formes qui composent le dictionnaire électronique des mots simples d'Intex sont ambiguës.

3. Les homographies les plus fréquentes

À partir d'une année du journal *Le Monde*, nous avons extrait les 100 chaînes de caractères les plus fréquentes. Parmi ces 100 chaînes se trouvaient des caractères diacritiques que nous avons enlevés. Nous avons ensuite regroupé les différentes graphies correspondant à une même forme : ainsi, dans le tableau suivant, le nombre d'occurrences de *en* équivaut au nombre d'occurrences de *en*, *En* et *EN* additionnées. Néanmoins, ce type de regroupement n'est pas toujours possible. Ainsi, *A* peut correspondre à la forme *à* ou à la forme *a*. De même, nous n'avons pu ramener à une seule forme *DE* (*dé* ou *de*), *DES* (*des*, *dés* ou *dès*), *ENTRE* (*entre* ou *entrê*), *LA* (*la* ou *là*), LES (les ou lès), *MAIS* (*mais* ou *maïs*), *NE* (*ne* ou *né*), *OU* (*ou* ou *où*) et *SUR* (*sur* ou *sûr*). C'est la raison pour laquelle ces formes figurent dans le bas du tableau.

Nous donnons le classement par ordre de fréquence décroissant pour LM94 :

1138721 de	169707 qui	63194 ont	36389 ils	24561 président
628021 la	162139 dans	59609 mais	36272 avait	23844 faire
484111 l	153834 pour	58495 aux	35124 ces	20161 C
477784 le	140314 au	56160 sont	35080 nous	785 LA
415407 à	136304 par	53284 cette	34818 tout	440 LES
400600 les	109728 pas	50283 été	34260 fait	419 DE
390457 et	106589 sur	49970 ou	30957 sans	178 DU
349431 des	94713 plus	47502 ses	30059 entre	152 DES
347489 d	93112 qu	47142 comme	29995 était	141 SUR
295357 en	92694 s	45018 sa	29845 lui	42 MAIS
252992 un	86493 ne	42619 elle	28550 France	26 NE
260384 du	85304 ce	42003 m	27299 ans	14 ENTRE
213275 une	84562 n	40687 deux	27293 dont	14 OU
191914 est	83939 se	39399 leur	27262 aussi	
176771 il	74279 son	38382 être	25712 A	
175726 a	67291 avec	38338 même	25527 bien	
169960 que	63225 on	36766 y	26932 où	

Nous proposons de voir dans quelle mesure ces formes les plus fréquentes (qui sont *grosso modo* les mêmes quel que soit le type de corpus analysé) sont ambiguës et, partant, présentons une méthode simple et « rentable » de désambiguïsation.

4. Parmi les mots les plus fréquents, les formes non ambiguës

On constate qu'il y a parmi les mots les plus fréquents 25 formes pour lesquelles il n'y a qu'une seule entrée dans le dictionnaire : *à* (PREP), *ans* (N), *au* (PREPDET), *aux* (PREPDET), *avait* (V), *ces* (DET), *cette* (DET), *dans* (PREP), *dont* (PRO), *elle*[3] (PRO), *et* (CONJC), *était* (V), *faire* (V), *France* (N), *il* (PRO), *ils* (PRO), *ne* (ADV), *nous* (PRO), *ou* (CONJC), *où* (PRO), *qui* (PRO), *sa* (DET), *sans* (PREP), *se* (PRO), *ses* (DET). Ces formes ne posent donc aucun problème pour le travail qui nous occupe ici, qu'il s'agisse

1. de lemmatisation au sens strict : on recherche uniquement le lemme, c'est-à-dire le mot vedette en entrée dans les dictionnaires courants : {a}, {avoir}, etc.
2. d'étiquetage avec étiquette plus ou moins complète :

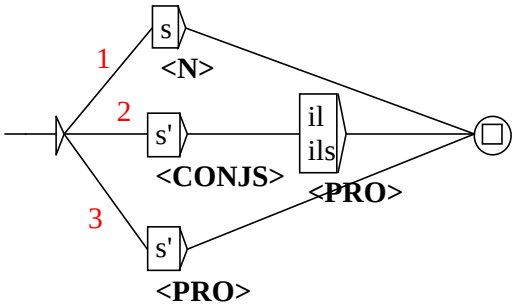
– soit uniquement le lemme et la catégorie grammaticale : ex. {a,N}, {avoir,V}, etc.

– soit le lemme, la catégorie et les informations morphologiques associées : {a,N:ms}, {avoir,V:I3s}[4], {donner,V:P1s:P3s:S1s:S3s:Y2s}[5].

Ces formes totalisent 2 259 850 occurrences, soit 30,36 % des formes les plus fréquentes, ou encore 11,05 % de la totalité de notre corpus.

5. Des grammaires locales pour décrire les ambiguïtés

Restent 56 formes du tableau qui présentent une ambiguïté quant à la catégorie grammaticale (soit 7 443 929 occurrences). En observant le contexte local, parfois très réduit (seulement deux ou trois mots à gauche et/ou à droite) de la forme, il est possible de lever l'ambiguïté. C'est la méthode que nous utilisons, en décrivant ce contexte dans des *grammaires locales*[6].



Grammaire 1

Grammaire1 permet de désambiguïser à 100 % les 92 694 formes *s* de LM94. Il faut lire la grammaire de la manière suivante : si l'on rencontre la forme *s* (chemin1), cette forme est ana-

---

[3] Ici encore, il faut insister sur l'importance des données présentes dans le dictionnaire : on pourrait décider que *elle* est ambigu si l'on choisit de coder différemment le pronom sujet et le pronom complément.

[4] Analyse de la forme *avait* : verbe (V) à la troisième personne (3) du singulier (s) de l'indicatif imparfait (I).

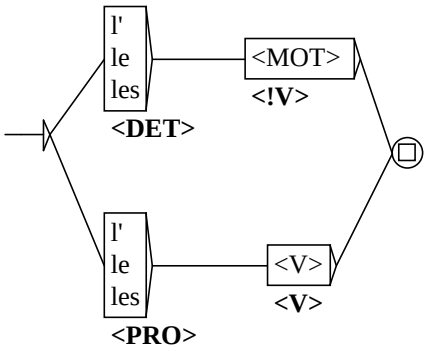
[5] Il s'agit de l'analyse de la forme *donne* : verbe (V) à la première (1) ou à la troisième personne (3) du singulier (s) de l'indicatif présent (P) ou du subjonctif présent (S) ; verbe à la deuxième personne (2) de l'impératif présent (Y).

[6] Les grammaires que nous présentons ici sont en fait des automates à état fini. Pour plus de détails sur le formalisme des automates, voir Silberztein (1997) ; pour plus de détails sur les grammaires de désambiguïsation d'Intex, voir Dister (1999).

lysée comme un nom (c'est ce qu'impose la contrainte <N> sous la boîte); si l'on rencontre la forme *s'* suivie des pronoms *il* ou *ils*, ce *s'* est une conjonction de subordination (chemin2) ; les autres formes *s'* sont des pronoms <PRO> (chemin3).

De la même manière, on peut désambiguïser les formes *n* et *c*, en fonction de la présence ou non de l'apostrophe. Dans ces deux cas également, on a une désambiguïsation totale et correcte à 100 %.

En ce qui concerne le *l*, l'absence d'apostrophe permet d'analyser la forme comme un nom ; la présence de l'apostrophe maintient l'ambiguïté entre les hypothèses pronom et déterminant. Néanmoins, cette ambiguïté peut être levée dans de nombreux cas par la grammaire suivante :



Grammaire 2

Le chemin1 impose à le forme *l'* d'être un déterminant <DET> si le mot qui suit n'est pas un verbe <MOT>/<!V>; le chemin2 impose à *l'* d'être un pronom <PRO> si le mot qui suit est un verbe <V>. Cette grammaire possède un très haut rendement de désambiguïsation, non seulement pour *l'*, mais aussi pour *le* et *les* qui possèdent la même ambiguïté catégorielle <DET> ou <PRO>. Dans le chemin1, il est préférable de ne pas spécifier davantage la catégorie du mot qui suit, car il peut tout autant s'agir d'un nom (« le garçon »), que d'un adjectif (« le beau garçon ») ou encore d'un adverbe (« le très beau garçon »). Pour le *l'*, cette grammaire couvre 481 260 séquences dans lesquelles le *l'* est désambiguïsé dans 84,3 % des cas. Le reliquat correspond à des séquences telles que « les portions » pour laquelle deux analyses parallèles peuvent être données en ce qui concerne la catégorie grammaticale. Il s'agit soit de la suite <DET> <N> (dans « les portions sont équitables », soit de la suite <PRO> <V> (dans « nous les portions »). Il apparaît clairement que cette ambiguïté peut être levée si l'on élargit le contexte[7].

Les grammaires locales que nous avons construites permettent d'étiqueter, notamment, la majorité des occurrences de *en* et *y*.

6. Tenir compte des mots composés pour l'étiquetage

Par ailleurs, la forme la plus fréquente dans les textes, *de*, peut difficilement être analysée par ce type de méthode ; il en va de même de *des*, *d'*, *qu'* et *que*. Les analyses de ces formes nécessitent bien souvent une description syntaxique globale de la phrase. Néanmoins, si l'on envisage une analyse qui tient compte des mots composés, il est alors inutile d'analyser ces formes en tant que telles : elles appartiennent à une séquence plus grande, qui doit elle seule être prise

[7] Notons à cet égard que le nombre de mots pris en considération dans le contexte d'une grammaire locale n'est pas limité.

en considération. Il peut s'agir d'un nom (*point de vue*), d'un adverbe (*de plus en plus*), d'une préposition (*vers la fin de*), d'une conjonction (*afin que*).

Dans notre corpus de référence LM94 figurent un certain nombre de composés dont l'un des éléments ne se rencontre jamais indépendamment de la composition. La présence de cet élément toujours lié permet ainsi de repérer aisément la présence du composé. Ainsi, *aujourd* et *hui* ne se rencontrent que dans *aujourd'hui* ; *tâtons* et *jeun* nécessitent la présence à gauche de *à*, etc. Dans LM94, la forme *a* entre en composition pour former les mots suivants où elle n'a plus besoin d'être analysée :

a capella	22	*a fortiori*	123	*a posteriori*	108
a cappella	18	*a giorno*	2	*a priori*	529
a contrario	106	*a minima*	13	*a tempera*	1

De la même manière, les formes *que* et *de* entrent dans la constitution de composés non ambigus.

afin que	362	*afin de*	2515
afin qu'	193	*afin d'*	1055
parce que	3265	*de facto*	200
parce qu'	2785	*de guingois*	19
tandis que	2459	*de profundis*	5
tandis qu'	447	*de traviole*	1

Tous ces mots devraient être recensés dans un dictionnaire de composés non ambigus ; ils seraient alors reconnus comme une seule unité de traitement. Cela éviterait ce qui n'apparaît plus dès lors que comme un faux problème d'analyse.

Nous présentons à présent les 100 composés les plus fréquents dans LM94, uniquement pour les catégories nom et adverbe. Ces termes ont été repérés en appliquant au corpus les dictionnaires de mots composés Advs.bin et Noms.bin élaborés au LADL[8]. Les termes (il s'agit ici plus exactement de la forme lemmatisée) sont classés par ordre de fréquence décroissant et accompagnés de leur nombre d'occurrences.

États-Unis (N : 8340); *un peu* (ADV : 6584); *premier ministre* (N : 6507); *en effet* (ADV : 951); *sans doute* (ADV : 4905); *d' abord* (ADV : 4429); *peut-être* (ADV : 4366); *d' ailleurs* (ADV : 3930); *par exemple* (ADV : 3830); *plus tard* (ADV : 3501); *au moins* (ADV : 212); *en revanche* (ADV : 3109); *à la fois* (ADV : 2823); *de plus en plus* (ADV : 2784); *Union européenne* (N : 2563); *affaires étrangères* (N : 2439); *au-delà* (N : 2392); *par ailleurs* (ADV : 2376); *en fait* (ADV : 2333); *envoyé spécial* (N : 2226); *de l' intérieur* (ADV : 2141); *pour la première fois* (ADV : 2108); *en outre* (ADV : 2091); *secrétaire général* (N : 2050); *Nations unies* (N : 2008); *à travers* (ADV : 1985); *porte-parole* (N : 1971); *en particulier* (ADV : 1952); *chiffre d' affaires* (N : 1944); *la nuit* (ADV : 1943); *de même* (ADV : 1898); *New-York* (N : 1887); *bien sûr* (ADV : 1865); *sur ce* (ADV : 1815); *à peine* (ADV : 1813); *Assemblée nationale* (N : 1787); *autre part* (ADV : 1785); *Grande-Bretagne* (N : 1777); *président de la République* (N : 1744); *non plus* (ADV : 1724); *élection présidentielle* (N : 1691); *un jour* (ADV : 1683); *projet de loi* (N : 1595); *à nouveau* (ADV : 1570); *directeur général* (N : 1559); *vice-président* (N : 1556); *conseil général* (N : 1552); *et plus* (ADV : 1528); *d' ici* (ADV : 1515); *non seule-*

[8] Laboratoire d'Automatique Documentaire et Linguistique, sous la direction de Maurice Gross (Paris 7).

ment (ADV : 1506); *y compris* (ADV : 1491); *dans le monde* (ADV : 1460); *casque bleu* (N : 1440); *une fois* (ADV : 1434); *à deux* (ADV : 1432); *cette fois* (ADV : 1383); *droits de l'homme* (N : 1353); *en tout cas* (ADV : 1322); *au total* (ADV : 1293*); Ile-de-France* (N : 1291); *à l'époque* (ADV : 1243); *mise en place* (N : 1243); *l'an dernier* (ADV : 1234); *ancien ministre* (N : 1223); *à l'origine* (ADV : 1221); *point de vue* (N : 1202); *tout à fait* (ADV : 1188); *en attendant* (ADV : 1169); *conseil d'administration* (N : 1161); *sur le terrain* (ADV : 1156); *Crédit lyonnais* (N : 1152); *dès lors* (ADV : 1144); *jusqu'à présent* (ADV : 1136); *pour l'instant* (ADV : 1128); *collectivité locale* (N : 1124); *un coup* (ADV : 1109); *rendez-vous* (N : 1105); *d'après* (ADV : 1096); *conseil régional* (N : 1058); *service public* (N : 1055); *parti socialiste* (N : 1015); *pouvoir public* (N : 1015); *aménagement du territoire* (N : 1014); *depuis longtemps* (ADV : 985); *la semaine dernière* (ADV : 983); *encore plus* (ADV : 980); *ce jour* (ADV : 970); *la journée* (ADV : 966); *avant tout* (ADV : 963); *le jour* (ADV : 958); *de nouveau* (ADV : 951); *en moyenne* (ADV : 950); *du moins* (ADV : 947); *plus ou moins* (ADV : 926); *Conseil de sécurité* (N : 919); *communauté internationale* (N : 917); *Front national* (N : 888); *parti communiste* (N : 872); *plus tôt* (ADV : 864); (ADV : 841).

6.1. Les composés non ambigus

Parmi ce termes, nous pouvons d'emblée en isoler certains qui sont non ambigus : *États-Unis, peut-être, Union européenne, Nations unies*[9]*, porte-parole, New-York, Assemblée nationale, Grande-Bretagne, élection présidentielle, vice-président, Ile-de-France, conseil d'administration, Crédit Lyonnais, dès lors, jusqu'à présent, collectivité locale, depuis longtemps, la semaine dernière, Front National, etc.*

Partant de cette liste, on pourrait croire que tous les mots formés avec un trait d'union sont non ambigus. Or, le recensement des composés non ambigus doit se faire avec la plus grande prudence car il faut constater, parmi ces 100 mots les plus fréquents, la présence de deux contre-exemples :

rendez-vous peut également être analysé comme la suite V-PRO ;

au-delà peut faire partie des prépositions *au-delà de, au-delà d', au-delà du* ou *au-delà des*. Ces quatre suites totalisent 1 961 occurrences sur les 2 392 de *au-delà*. Néanmoins, il ne faut pas conclure que la suite *au-delà de* est d'office à analyser comme une préposition. Pour preuve ces deux phrases extraites de notre corpus :

« (…) de retrouver cette préexistence de l'écriture et d'entendre en écho l'au-delà de l'écrit. »

« N'est-elle pas trop silencieuse sur l'irrationnel, sur l'au-delà de la mort ? »

Ni *au-delà*, ni *au-delà* de ne peuvent donc être répertoriés dans un dictionnaire de composés non ambigus.

6.2. Ambiguïté composé / suite libre

Le deuxième type de composés que nous pouvons mettre en évidence participe du phénomène de l'ambiguïté sémantique. Soit le terme présent dans le texte est effectivement un mot composé, soit il s'agit simplement d'une séquence libre (suite de plusieurs mots simples). Dans notre corpus, c'est le cas de *casque bleu.* Il peut s'agir du nom humain (les soldats de la paix), soit d'un casque de couleur bleue, sans aucune référence au militaire. Le même type d'ambiguïté se retrouve dans *premier ministre*. Ce terme fait généralement référence à la personne qui occupe la fonction de « premier ministre » (en 1999, Lionel Jospin pour la France, Guy

[9] En tout cas, avec une majuscule à *Nations*. Mais on peut imaginer une séquence comme « des nations unies contre l'agresseur » dans laquelle *nations* et *unies* recevraient chacun une analyse.

Verhoofdstad pour la Belgique). Mais on peut très bien parler d'un ministre qui est le premier à faire quelque chose, et *premier ministre* est alors une suite libre comme dans : « Il est le premier ministre à tenir les engagements pris pendant la campagne électorale. » Il peut s'agir alors du ministre de l'agriculture ou du ministre des finances.

Ce type d'ambiguïtés est fréquemment mentionné dans les travaux sur l'étiquetage, avec les exemples bateau de *carte bleue*[10] ou *cordon bleu*.

6.3. Les ambiguïtés de segmentation

Mais un autre problème nous semble aussi gênant, qui affecte un plus grand nombre de formes. Il s'agit de toutes les ambiguïtés liées à des difficultés de découpage, mais dont les limites dépassent la forme même du composé. Entre autres, nous pouvons citer les exemples suivants :

droits de l'homme : « Il a reçu ses droits de l'homme le plus puissant d'Arabie Saoudite »;

parti socialiste : « Parti socialiste, il est revenu communiste »;

service public : « Il a rendu ce service public[11] »;

en fait : « Si le personnel des hôpitaux publics en fait déjà les frais (…) »;

mise en place : « En 1993, une nouvelle direction a été mise en place et une réorganisation lancée »;

sur le terrain : « Sur le terrain de l'emploi, il ne peut être question (…) »;

de l'intérieur : surtout dans les occurrences *ministre de l'intérieur* ou *ministère de l'intérieur*.

On peut également ranger dans cette catégorie les formes *au-delà* et *au-delà de* mentionnées plus haut, mais aussi *sur ce* (« sur ce chemin »), *à l'origine* (« à l'origine de mon malheur »), *autre part* (« d'autre part »), etc.

S'il est vrai que les cas de figure illustrés par les trois premiers exemples sont plutôt rares, ceux-ci ne nous semblent pas devoir être écartés a priori. De plus, on le voit dans les exemples repris ensuite, le problème de découpage est bien réel pour un grand nombre de candidats composés rencontrés dans les textes. Il serait d'ailleurs intéressant de se pencher également sur les locutions prépositives ou conjonctives qui nous semblent de ce point de vue assez productives.

S'ils s'avèrent incontournables dans toute étude de la langue, la prise en compte des mots composés apporte de nouvelles perspectives mais aussi son lot de problèmes pour l'analyse automatique.

7. Conclusions

Décrire le contexte immédiat de certains homographes dans des grammaires locales afin de lever les ambiguïtés de catégorie grammaticale est une méthode particulièrement rentable, notamment pour les formes les plus fréquentes des textes. Mais on l'a vu, certaines formes échappent à ce type d'analyse trop contextuel. De plus, la prise en compte d'unités composées comme unités minimales, qui nous semble être la seule approche satisfaisante, change totalement la perspective de travail. Avant d'analyser chaque forme comme une entité autonome, il

---

[10] En Belgique, cet exemple n'est d'ailleurs absolument pas parlant puisque *carte bleue* n'est jamais figé (on utilise *carte de crédit* ou *carte bancaire*).

[11] Ces trois premiers exemples sont de nous.

faudrait observer systématiquement dans quelle mesure elle entre ou non dans la construction d'un mot composé. Il faudrait également coder les composés de manière à obtenir des dictionnaires fiables de composés non ambigus.

Références

Adda G., Mariani J., Paroubek P., Rajman M. and Lecomte J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. in *Actes de TALN99, 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles. 12-17 juillet 1999. Cargèse, Corse*, pages 15-24.

Courtois B. (1996). Formes ambiguës de la langue française. *Lingvisticæ Investigationes (*XX:1), John Benjamins : 167-202.

Dister A (1999). Construire des grammaires de levée d'ambiguïtés pour Intex. *Lingvisticæ Investigationes* (à paraître).

Dister A. (1999). De l'étiquetage traditionnel au transducteur du texte. La levée d'ambiguïté par grammaires locales. *RISSH* (à paraître).

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

Illouz G., Habert B., Fleury S., Folch H., Heiden S. and Lafon P. (1999). Maîtriser les déluges de données hétérogènes. In *Actes de TALN99, 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles. 12-17 juillet 1999. Cargèse, Corse*, pages 37-46.

Laporte É. (1995). Levée d'ambiguïtés par grammaires locales. *Linguisticæ Investigationes Supplementa*, John Benjamins : 97-114.

Silberztein M. (1990). Le dictionnaire électronique des mots composés. *Langue française* (n°87), *Dictionnaires électroniques du français*, Larousse : 71-83.

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

Silberztein M. (1995). Dictionnaires électroniques et comptage des mots. In *Actes de JADT' 95*.

Silberztein M. (1997). The lexical analysis of natural languages. In Emmanuel Roche and Yves Schabès editors, *Finite-state language processing*, MIT Press.