Recherche de critères formels pour l'identification automatique des particules discursives

Sandra Teston & Jean Véronis

Equipe DELIC, Université de Provence
29 Avenue Robert Schuman - 13100 Aix-en-Provence
{sandra.teston, jean.veronis}@up.univ-aix.fr

Nous nous intéresserons dans cette communication à un certain nombre de « marqueurs » d'unités discursives, (l'unité discursive sera définie ici comme une nouvelle construction grammaticale ou une nouvelle construction non canonique constituant une unité de communication complète, telle *bof* ou *si je pouvais*…) qui jouent un rôle important dans l'organisation du discours : *particules discursives* (*bon, ben, là, quoi,* etc.), *connecteurs* (*mais, donc, parce que, aussi,* etc.) et un certain nombre de locutions telles que *de toute façon, en fin de compte, en définitive,* etc. Chanet (2004) recense la terminologie extrêmement abondante qui a été proposée dans la littérature, et souligne le flou (voire l'aspect contradictoire) des définitions proposées.

Il est connu que ces marqueurs (particulièrement les particules discursives -- ci-après PDI) ont une fréquence particulièrement élevée à l'oral, mais leur répartition exacte entre genres textuels est mal connue. Nous montrerons dans la première partie de cette communication que ce serait une erreur de les considérer comme marques exclusives de l'oralité ou des représentations écrites de l'oralité (dialogues), et que ce qui diffère entre genres, c'est leur distribution. En particulier, les *nouvelles formes de communication écrite* (NFCE), que sont les forums, chats, e-mails, comportent une proportion importante de marqueurs, mais nous verrons qu'ils ne correspondent pas nécessairement à ceux de l'oral. De plus, particulièrement dans le cas des PDI, de multiples formes sont homonymes de catégories grammaticales traditionnelles (par exemple, *bon, quoi*), et le traitement automatique des langues (TAL) est relativement démuni face à leur détection : elles sont purement et simplement ignorées par la plupart des systèmes, tels que les étiqueteurs morpho-syntaxiques (qui se contentent par exemple de *bon* adjectif ou *quoi* pronom en toutes occurrences). Cette situation, théoriquement peu fondée, mais pratiquement acceptable dans certains genres d'écrits, n'est viable ni pour l'oral, ni pour les NFCE. La deuxième partie de la communication indique que des progrès substantiels peuvent être espérés dans le traitement de ces particules, pour peu qu'on leur accorde l'attention qu'elles méritent.

Nous avons constitué quatre corpus de même taille (440 000 mots chacun), bien différenciés du point de vue du genre :

- le *Corpus de référence du français parlé,* réalisé dans notre équipe (DELIC, 2004), et qui comporte la transcription d'environ 37 heures de parole spontanée, enregistrée dans des situations diverses ;
- un corpus composé de messages diffusés sur le forum Usenet *fr.soc.divers* (sujet généraux de société), préalablement nettoyé (suppression des en-têtes, des citations et des signatures) ;
- un corpus d'œuvres littéraires du domaine public (fin XIXème-début XXème), comportant volontairement environ un quart de pièces de théâtre (représentation de l'oral), le reste étant composé de contes, extraits de romans (comportant donc une part importante de dialogues) , correspondances (forme de « dialogue » en temps différé) et poésie ;
- un ensemble d'articles du journal *Le Monde.*

Nous avons tout d'abord calculé la fréquence des formes correspondant à des marqueurs dans chacun des corpus, en nous fondant sur la liste de 85 formes proposées par Chanet (2004). Il va de soi que nombre de ces formes sont ambiguës (*bon, quoi*, etc.) et que toutes les occurrences observées ne sont pas seulement des marqueurs (ce qui fait l'objet de la deuxième partie de notre présentation). Les chiffres ci-dessous ne doivent donc être pris que comme des tendances, à affiner lorsque des procédures de traitement automatique fiables seront disponibles.

La fréquence moyenne des formes sur l'ensemble des quatre corpus est de 3,9%, et culmine à 6,9% pour l'oral. La Figure 1 montre clairement la gradation qui existe entre les quatre genres : oral > forum > littérature > presse. Ceci confirme la place importante des marqueurs dans les NFCE, mais il est peut-être étonnant de ne pas en trouver une proportion plus importante dans le corpus littéraire, étant donné sa composition (théâtre, dialogue, correspondance). Sans doute la représentation de l'oral est-elle assez différente de l'oral lui-même.

Il est très intéressant de se pencher sur la distribution des différentes formes correspondant à des marqueurs dans les quatre corpus. Nous avons appliqué un test de χ^2 à chacune des formes ; le Tableau 1 donne pour chacun des corpus la liste des formes qui sont caractéristiques de chacun des corpus (contribuant à plus de 25% du χ^2). Les formes sont classées par fréquence décroissante.

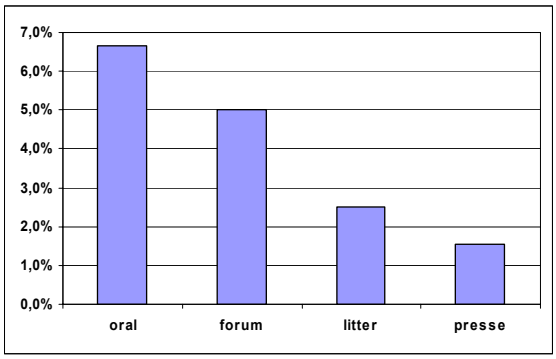

Figure 1. Fréquence des formes correspondant à des marqueurs discursifs

Oral	*donc, là, bon, alors, parce que, quoi, après, puis, enfin, voila, ben, en fait, quand même, puisque, par exemple, c'est-à-dire, ensuite, justement, disons, du coup, malgré tout, étant donné que, en l'occurrence*
Forum	*mais, bien, aussi, car, surtout, d'ailleurs, de plus, seulement, plutôt, en plus, pourtant, en effet, sinon, finalement, certes, par contre, en tout cas, à propos, bref, tout de même, au contraire, du moins, au fait, néanmoins, par ailleurs, toutefois, à cause de, concernant, après tout, en somme, en ce qui concerne, n'empêche, autrement dit, en réalité, en fin de compte, de ce fait, de toute manière, somme toute, inversement*
Littérature	*ainsi, or, cependant, malgré, tandis que, au fond, par conséquent, si bien que, de sorte que*
Presse	*notamment, en revanche, précisément, de fait, en outre, au total*

Tableau 1. Formes (MD et PDI) les plus fréquentes par Corpus

Si l'on observe des attendus (par exemple, *donc, là, bon,* etc. à l'oral, et *notamment, en revanche,* etc. dans *Le Monde*), on ne peut qu'être frappé par le comportement du corpus littéraire et du forum. Il est surprenant que la littérature ne contienne pas ces formes de manière plus importante, même dans sa représentation de l'oral (mais il est vrai que la tranche temporelle concernée est ancienne : nous entreprendrons bientôt une étude comparative avec de la littérature contemporaine et des dialogues de cinéma). Le forum, quant à lui, se distingue de façon tout à fait étonnante, à la fois par le nombre des formes, mais aussi par leur nature : ce ne sont pas les formes caractéristiques de l'oralité qui y sont massivement significatives (même si celles-ci sont aussi présentes), mais plutôt des formes parfaitement normatives, que l'on rencontre parfois très peu à l'oral (par exemple *car* ou *certes*). La nature très argumentative des forums explique sans doute cette particularité.

Comme nous l'avons dit plus haut, ces chiffres sont des estimations, car nombre de formes correspondant à des marqueurs (particulièrement des PDI) sont ambiguës. C'est le cas, par exemple, de *bon,* que nous avons choisi pour une étude détaillée, qui peut être adjectif ou adverbe (et très rarement nom). La Figure 2 donne la proportion de *bon* dans chacun des corpus et la répartition PDI/non-PDI. On voit que c'est surtout *bon* PDI qui entraîne une différence importante entre genres. Si cette tendance se confirme pour les autres PDI, elle justifierait pleinement l'approximation que nous venons d'utiliser, mais il est évident que la mise au point de programmes automatiques permettant un tri, même imparfait, constituerait une amélioration appréciable.

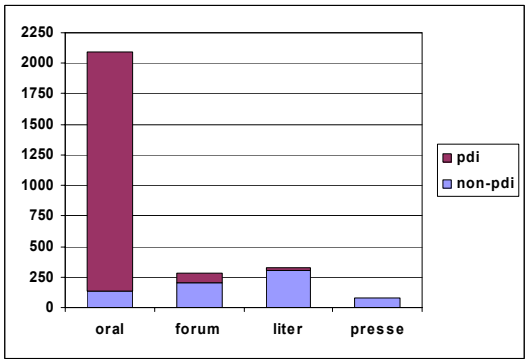

Figure 2. Fréquence de bon dans les quatre corpus

Bon particule discursive présente des propriétés autres que celles de l'adjectif ou de l'adverbe dont il est homonyme, puisque, notamment, on ne peut ni le rattacher à une tête nominale, ni le modifier (**très bon*), ni le faire commuter avec un adjectif, comme dans les exemples suivants:

> ***bon*** *j'ai été gentil parce que je lui ai dit oui [Oral]*
> *les fruits sont très (sic) chers ici mais* ***bon*** *c'est vrai que le plus économique (sic) c'est les burgers ou*

pizzas. [Forum]

***Bon, bon** ! le duc est jeune, Marquise, et gageons que cet habit coquet des nonnes lui allait à ravir. [Litter]*

*"...Cette année, j'ai critiqué l'organisation du tournoi. **Bon**, passons à autre chose..." [Monde]*

De telles propriétés distributionnelles caractérisent toutes les PDI, et nous soutiendrons le point de vue selon lequel les PDI relèvent d'une analyse syntaxique et pas seulement pragmatique. La comparaison est intéressante avec les éléments autonomes fonctionnellement (AF)**,** *c'est-à-dire* non gouvernés par le verbe recteur principal de l'énoncé (« non régis ») : certains adverbes (dits « de phrase » comme *franchement*), des connecteurs (*de toute façon*, *finalement*…), des « circonstants » (*par bonheur* dans *par bonheur je pars en vacances*), des appositions, des constructions verbales incises, des syntagmes prépositionnels (*à vrai dire*…), etc. Sur le plan syntaxique, les PDI ont des fonctionnements très analogues à ces éléments, qui, eux, sont reconnus par la tradition normative et décrits dans les grammaires (incomplètement, il est vrai). Sur le plan de l'analyse en catégories, les PDI demandent toutefois une analyse distincte, car leur rattachement à une catégorie connue, adjectif, verbe, adverbe, syntagme prépositionnel etc., présente beaucoup plus de difficultés. On peut cependant montrer :

1. que les PDI ne peuvent constituer un groupe avec une syntaxe interne développée :

*je suis amoureux, **quoi**, *je suis amoureux **eh quoi**, *je suis amoureux **quoi encore**, ***comme quoi**, ***de quoi***

2. que cette syntaxe interne est affectée de contraintes à l'instar des AF, parfois même plus importantes que celles des AF eux-mêmes, ainsi l'unité PDI *tu sais* dans cet exemple :

*tu as de beaux yeux, **tu sais**, ?**tu sais bien**, ***tu ne sais pas**, ***tu sais de source sûre**, ***je sais***

à comparer avec un AF comme *à mon avis* ***:***

*tu as de beaux yeux **à mon avis**, **à notre avis**, **à mon humble avis***

*tu as de beaux yeux ***à l'avis de Jean**, ***à cet avis**, ***à l'avis que toute l'équipe partage.***

Par contraste, ces contraintes n'affectent pas le syntagme prépositionnel régi :

*il se rend **à mon avis, à l'avis de Jean, à cet avis, à l'avis que toute l'équipe partage.**.*

Le domaine des PDI peut donc être intégré à celui des AF décrits par les grammaires, mais pas strictement assimilé à celui-ci. Nous avons donc entrepris d'analyser comparativement les différentes aptitudes, capacités et action sur l'énoncé de tous les AF, y compris les PDI. Une première approche de classement des AF peut être tentée à partir des sites d'*insertions*, des *contraintes* qu'ils génèrent au niveau de la syntaxe externe et/ou de la syntaxe interne, et de leur *portée* particulière. En observant les PDI sous ce nouvel angle, leurs caractéristiques formelles peuvent être observées et déterminées avec minutie. D'ailleurs, des micro-distributions et une syntaxe propre à chaque forme peuvent apparaître, comme à l'intérieur de l'organisation grammaticale où des éléments d'une même catégorie présentent souvent des contraintes spécifiques, ainsi :

*tu me l'as demandé **donc** je viendrai, tu me l'as demandé je viendrai **donc***

*je viendrai à la fête **car** tu me l'as demandé, *je viendrai à la fête tu me l'as demandé **car***

***c'est évident** que tu peux venir, **évidemment** que tu peux venir*

***c'est génial** que tu viennes, ***génialement** que tu viennes*

***l'important** est que tu sois remercié, ***le grand** est que tu sois remercié*

*une excellente carte, *une routière carte*

Après analyse, les résultats obtenus permettent d'identifier les formes dans leur emploi discursif, par exemple *bon* peut être spécifié par des propriétés formelles communes (aux AF-PDI en général) et par des propriétés formelles particulières à la forme, qui l'opposent notamment aux autres catégories avec lesquelles elle est ambiguë (pour *bon* : adj. ou adv.).

Nous nous sommes fixés pour objectif d'écrire un programme qui implémente les critères discriminants de *bon*, avec des contraintes méthodologiques très précises. On sait en effet que, sur un cas particulier, on peut multiplier les règles ad hoc jusqu'à arriver à une précision aussi grande que l'on désire. De plus, considérant qu'il s'agit d'une étude pilote destinée à tester la pertinence de l'approche, il s'agissait de faire appel le moins possible aux ressources humaines et linguistiques. Nous nous sommes donc fixé pour limites que (1) l'écriture du programme ne devait pas prendre plus de deux heures, (2) le programme ne devait pas faire appel à plus d'une quinzaine de règles, et (3) le programme ne devait pas faire appel à un dictionnaire de plus d'une centaine de formes. Le Tableau 2 résume les différentes règles implémentées (le signe + marque des règles qui indiquent des ruptures de construction fréquentes). Dans les contextes où aucune règle n'est applicable, le programme a été pourvu d'une stratégie par défaut, consistant à appliquer l'étiquette la plus fréquente dans le corpus traité.

Le programme a été testé sur les quatre corpus, et les résultats ont été vérifiés manuellement. Le programme a fait appel à la stratégie par défaut dans 6,5% des cas seulement, ce qui montre la *bon*ne couverture des règles. La précision obtenue, sur l'ensemble des quatre corpus a été de 97.6% (Tableau 3). L'étiquetage de base (baseline) qui consisterait à attribuer à toutes les occurrences d'un corpus l'étiquette la plus fréquente du corpus ne produirait que 91,5% d'étiquettes correctes. Le programme réduit donc le nombre d'erreurs de 72,3% par rapport à cet étiquetage naïf (colonne Δ Err), ce qui est un résultat honorable étant donné la rusticité (voulue) du programme. Le corpus pour lequel l'étiquetage naïf est le moins performant est le forum, à cause de la répartition de *bon* PDI/non-PDI, qui y est plus équilibrée que dans les autres corpus.

Contexte	***PDI***	***NON-PDI***
Locution		•*à quoi bon, pour de bon, ce que bon (me, lui...) semble*
Gauche	•Déterminant fém. ou plur. $^{+}$ •Clitique $^{+}$ ou pronom disjoint •Début d'énoncé •*Euh*, pause, amorce (oral), ponctuation (écrit) •PDI, connecteurs, etc. : *ah, alors, sinon, mais, parce que...* •*Il y a* •Verbe : *dire* (discours rapporté)	•Modifieur : *très, vraiment...* •Prépositions: *à, de, en, pour* •Verbes : *être, sembler, paraître, tenir, sentir, juger, trouver, il fait*
Droite	•Déterminant •PDI et connecteurs : *alors, ben, sinon, mais...* •Clitique sujet (*je, tu, il...*) •*Euh* •*Il y a, c'est*	•Expressions sans déterminant : *bon voyage, bon week-end, bon gré, bon an (mal an), bon sang...*

Tableau 2. Principaux critères discriminants pour bon

	Prec.	***Base***	***Δ Err***
oral	98,4%	93,6%	-75,0%
forum	95,5%	73,1%	-83,3%
litter	94,8%	92,2%	-33,3%
monde	97,6%	98,8%	+100%
Total	**97,6%**	**91,5%**	**-72,3%**

Tableau 3. Résultats de l'étiquetage automatique

C'est cependant sur le corpus littéraire que notre programme se comporte le moins bien, ce qui n'est guère étonnant, étant donné la grande complexité des structures et la richesse du lexique (et même des problèmes typographiques dans les poésies, *bon* en début de vers ne marquant pas nécessairement un début d'énoncé!). L'analyse des erreurs montre des améliorations évidentes. Par exemple, nous avons considéré à tort une virgule précédant *bon* à l'écrit comme révélatrice d'une PDI, faisant une analogie trop rapide avec la pause à l'oral. En fait, elle est la plupart du temps la marque d'une apposition, avec *bon* adjectif :

*Matisse, **bon** élève du symbolisme [Monde]*

Il restera toutefois environ 1% de cas très difficiles à traiter de façon automatique, liées, à l'oral, à des répétitions ou ruptures de constructions, telles que :

*il est vrai **ce bon** ce vin ressent aucun boisé [oral]*

*ce n'est pas ma faute **si bon** + et puis c'est un choix [oral]*

ou à l'écrit, à des problèmes typographiques et orthographiques très fréquents dans les forums (coupures des lignes, oublis d'accents, etc.).

Cette expérience montre donc que, d'une part, des critères distributionnels précis caractérisent bien (au moins) certaines particules discursives, et que, d'autre part, des résultats tout à fait honorables peuvent être atteints en matière de traitement automatique avec des efforts et des ressources très limités. Nos résultats semblent donc ouvrir des perspectives encourageantes pour un meilleur traitement de ces particules, certainement utile pour l'analyse de l'architecture et de l'organisation discursive des documents numériques, y compris les transcriptions d'oral et les nouvelles formes de communication écrite (e-mails, chats, forums, etc.).

Références

Chanet, C. (2004). Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. *Recherches sur le français parlé, 18* (sous presse).

Equipe DELIC (2004). Présentation du *Corpus de référence du français parlé. Recherches sur le français parlé.* 18 (sous presse).