

Ce que les corpus nous apprennent sur la langue

Bilger Mireille (Université de Perpignan) - bilger@univ-perp.fr

Cappeau Paul (Université de Poitiers) - Paul.Cappeau@univ-poitiers.fr

La description syntaxique à partir des corpus a pu, dans le passé, susciter de fortes réserves : la principale tenait à la représentativité des résultats. Quand on s'appuyait sur un petit corpus, pouvait-on prétendre accéder au "possible" de la langue ? Cet obstacle est en grande partie levée depuis quelques années (Bilger : 2000) avec la constitution de grands corpus de genres différents, comptabilisant plusieurs millions de mots.

À l'heure actuelle, il paraît difficile de remettre en question l'intérêt d'une exploitation systématique de telles données qui ont permis d'affiner, ou de renouveler, la description d'un grand nombre de phénomènes grammaticaux. Nous développerons trois exemples (les sujets *nous*, les adverbes en *-ment* et la forme *contre*) qui nous permettront d'esquisser les nouvelles perspectives qui s'offrent à la grammaire.

1. Présentation des outils utilisés

1.1. Banques de données

Les données que nous avons exploitées proviennent de plusieurs sources que nous présentons rapidement :

- CA est un corpus d'un million de mots dont la constitution s'est étalée sur 20 ans environ. Il contient des situations de parole variées (interviews, explications techniques, récits de diverses sortes, parodies, etc.). Seules les conversations sont peu représentées.

- CR est un corpus de référence de français parlé de 439 000 mots¹. Il a été recueilli en l'espace de 3 ans dans 37 villes de France. Divers paramètres ont présidé au recueil : l'âge des locuteurs, leur niveau scolaire, trois grandes situations de parole (privée, publique, professionnelle).

- HP est un corpus constitué exclusivement d'interviews d'hommes politiques dans les médias. Il comporte 300 000 mots.

- PE est un corpus de presse écrite (journal *Le Monde*) de un million de mots.

Travailler sur plusieurs banques de données différentes permet de disposer

:

- de plusieurs sortes d'oral (de l'oral formel, non formel, privé, public, professionnel)

- d'un type d'écrit (l'écrit de presse). On s'en est tenu à cette seule dimension pour l'écrit mais il est certain qu'il serait intéressant de disposer de diverses situations.

Une banque de données est un ensemble qui peut se subdiviser en strates. Si l'on prend l'exemple de CA, elle se décompose en trois niveaux différents :

¹ On peut consulter et interroger ce corpus sur le site de l'équipe DELIC : <http://www.univ-mrs.fr/delic/crpf>.

Niveau 1 : La banque de données orale dans sa globalité CA
Niveau 2 : Des situations identifiées à partir de critères externes ex : explications techniques, récits de vie, conversations, récits de voyage...
Niveau 3 : Des genres caractérisés par des propriétés internes ex : récits collectifs, etc.

Tableau 1 : Les strates dans CA (CorpAix)

Selon les faits de langue traités, il est intéressant, comme on le verra, de travailler à un niveau ou à un autre pour dégager des points pertinents.

1.2. Outils d'exploitation

Deux concordanciers ont été utilisés pour relever et étudier diverses formes dans leur environnement syntaxique. Il s'agit de :

- *Xcor* développé par Jean-Pierre Adam
- *Contexte* développé par Jean Véronis

Ces outils offrent la possibilité d'un recensement exhaustif des formes².

Il est intéressant de rappeler à ce sujet que si l'utilisation des concordanciers dans le cadre des recherches lexicographiques est ancienne, elle est d'emploi récent pour l'application aux phénomènes grammaticaux, et cela suscite encore, de la part de certains collègues, une certaine réticence.

2. Quelques exploitations

Dans cette seconde partie, nous donnerons quelques exemples précis qui illustrent l'intérêt du travail sur corpus.

2.1. Le sujet *nous*

La description du sujet *nous* qui ne s'appuierait pas sur des corpus n'aurait probablement pas grand intérêt. En effet, que dire de spécial sur ce sujet ? Nous utiliserons la banque de données CA (français tout venant) pour montrer quels enseignements différents on peut tirer selon que le travail porte sur l'une ou l'autre des strates de la banque de données.

Si l'on s'en tient au niveau 1 de la banque de données (c'est-à-dire prise dans sa globalité), le travail sur corpus s'avère assez peu pertinent. Un premier examen de CorpAix (CA) fournit les chiffres suivants :

	<i>je</i> ou <i>j'</i>	<i>nous</i>	<i>on</i>
Nombre d'occurrences	13 640	925	13 732

Tableau 2 : Les sujets *je*, *nous* et *on* dans CA

Que nous apprend ce premier tableau ? Que le sujet *nous* est peu répandu à l'oral. Ces résultats semblent conforter des explications déjà anciennes qui reposent sur la morphologie et considèrent que le pronom *on* s'est développé au détriment du sujet *nous*³ :

² Le résultat se présente sous la forme d'une liste d'énoncés avec : un certain nombre de mots du contexte gauche, la forme recherchée en position centrale et quelques mots du contexte droit.

³ Sur des corpus courts, on peut faire ressortir une corrélation inverse assez forte entre la fréquence de *nous* et celle de *nous on*. La séquence *nous on* est rare quand le sujet *nous* tient une

Dans la langue populaire et enfantine, *on* tend à remplacer *nous*; au facteur psychologique précité s'ajoute une raison de forme : *on* évite une forme plus longue (1^{re} pers. pl. en *-ons*), sans compter que la 3^e pers. sing., sans flexion auditive, est d'un emploi plus commode. [A. Dauzat. (1947). *Grammaire raisonnée de la langue française*. IAC. p. 262]

On peut tout de même creuser quelque peu cette description et rechercher comment les sujets *nous* se répartissent dans les divers enregistrements qui composent la banque de données. C'est ce qu'indique le tableau suivant :

Nombre d'occurrences du sujet <i>nous</i>	0	1	10 ou plus
Nombre de corpus concernés	142	35	30

Tableau 3. Répartition des sujets *nous* dans les corpus de CA

Ces résultats modifient profondément la présentation précédente :

- les sujets *nous* se rencontrent dans 105 enregistrements (sur 247), c'est-à-dire dans 42 % des corpus, alors que les sujets *je* sont présents dans toutes les productions et que la plupart des enregistrements contiennent des sujets *on*. Le sujet *nous* possède une distribution très disparate.

- l'explication morphologique perd une part de sa valeur : si vraiment la forme *nous* était écartée, elle devrait disparaître de façon semblable dans toutes les productions.

Pour mieux cerner les enregistrements (peu nombreux) dans lesquels on trouve de nombreuses occurrences du sujet *nous*, il convient d'en passer au niveau 2. Le travail sur corpus devient dès lors irremplaçable. Le sujet *nous* se maintient :

1) Lorsque les locuteurs s'expriment en tant que porte-parole d'un groupe, d'une entité sociale.

C'est le cas du corpus *Gendarme* où le locuteur expose les attentes des membres de sa profession ou de *Civeaux* qui est un commentaire recueilli pendant la visite guidée d'une centrale nucléaire. Le locuteur précise d'ailleurs souvent la position particulière de sa prise de parole :

1- donc tout cela m'a amené + à parler non pas de moi + ou de problèmes personnels + mais + de réalités sociales (CA)

2) Dans les récits de voyages collectifs.

Cette classification nous fait déjà plonger dans le niveau 3. En effet, si l'on identifie par le contenu les *récits de voyage* (qui correspondraient alors à notre niveau 2), il semble nécessaire sur la base de divers faits de langue de distinguer la "classe" des récits de voyage collectifs caractérisés, entre autres, par la présence de sujets *nous* et *on*. Il n'y a donc pas exclusion d'une forme au profit de l'autre mais plutôt une sorte de complémentarité de ces deux sujets qui présentent des affinités très fortes avec certains temps verbaux. On peut

grande place. Tout semble indiquer que *nous on* est en fait en concurrence avec le sujet *on* plus qu'avec le sujet *nous*. Ce qui renverse quelque peu l'explication morphologique.

notamment dresser le tableau suivant qui indique le nombre de verbes concernés :

	<i>nous</i>	<i>on</i>
imparfait / plus-que-parfait	26	70
passé composé	149	48
autres temps	6	51

Tableau 4. Alternance *nous* / *on* dans les récits de voyages collectifs

Ces résultats montrent que la distribution des sujets recoupe celle de l'opposition entre temps du récit et temps du discours. Ainsi :

- le passé composé domine nettement avec le sujet *nous* qui sert à indiquer les principales péripéties du voyage. L'emploi conjoint de ce sujet et de ce temps permet de relier plus étroitement le locuteur aux événements qu'il raconte :

2a- *nous sommes arrivés à Calais après avoir roulé pratiquement euh + depuis depuis Aix-en-Provence (CA)*

2b- *à Ouagadougou euh nous avons cherché pour nous loger (CA)*

- l'imparfait est plus souvent employé avec le pronom *on* et va donc servir à introduire des commentaires, décrochés par rapport à la trame événementielle :

3a- *enfin on (n') était pas là pour manger quoi + on était là pour pour observer*

3b- *nous sommes venus là habiter on avait + une grande pièce qui servait de tout quoi (CA)*

- parmi les autres temps utilisés, le présent (préférentiellement couplé avec *on*) servira pour les commentaires généraux :

4a- *et puis quand on a sommeil c'est + c'est assez + assez euh reposant (CA)*

4b- *la mairie + avait fait des préfabriqués ce qu'on appelle eh tu vois (CA)*

On voit ici comment, dans un type de production très particulier (les récits de voyage en *nous*), la réalisation des formes sujets apparaît comme un indicateur intéressant et très sensible.

2.2. L'exemple des adverbes en *-ment*

Dans les grammaires, les adverbes en *-ment*, dits *adverbes de manière*, sont souvent présentés comme étant, par vocation, des adverbes qui dépendent syntaxiquement du verbe et qui, dans ce sens, en modifient le sens. Le poids de l'étymologie joue sans aucun doute un rôle prépondérant dans cette présentation.

Quels enseignements nous apportent l'exploitation des corpus sur l'emploi des adverbes ? Nos observations (Bilger : 2003) porteront dans un premier temps sur deux banques de données : l'une orale (CR) et l'autre écrite (PE) afin de tirer certaines informations quantifiées sur ces formes. Dans un deuxième temps, nous regarderons de façon plus précise la répartition de ces adverbes en tenant compte des situations de parole. C'est alors au niveau 2 que nous nous situerons.

- a) Si on compare le corpus oral CR et le corpus écrit PE, on relève :

• d'une part que les 10 adverbess les plus fréquents ne sont pas identiques dans les deux types de corpus :

	Oral (CR)	Ecrit (PE)		Oral (CR)	Ecrit (PE)
1-	vraiment	seulement	6-	également	largement
2-	justement	notamment	7-	énormément	largement
3-	évidemment	également	8-	tellement	particulièrement
4-	forcément	actuellement	9-	complètement	simplement
5-	effectivement	vraiment	10-	finaleme nt	rapidement

Tableau 5. Les 10 adverbess les plus fréquents à l'oral et à l'écrit.

• d'autre part que les fréquences comparées de ces formes sont également différentes :

- à l'oral, le premier adverbe (*vraiment*) est trois fois plus fréquent que la seconde forme *justement*, suivie de près par la troisième forme *évidemment*.
- à l'écrit les trois premiers adverbess (*seulement*, *notamment*, *également*) ont des fréquences d'apparition quasi équivalentes.

On retrouve là des tendances maintes fois signalées (Halliday : 1985) concernant les différences de répartition : à l'oral l'écart est plus grand entre les premières formes et les suivantes, l'écrit présente une plus grande homogénéité.

On note que contrairement aux présentations fournies par les manuels de grammaire, les adverbess les plus fréquents, notamment à l'oral, sont majoritairement employés comme adverbess de "phrase" ou d'"énonciation" et non comme adverbess modifiant le sens du verbe (ou adverbess de manière).

b) Tous les adverbess relevés dans la liste du corpus écrit sont inclus dans la liste des 100 adverbess les plus fréquents du corpus oral, ce qui s'oppose à l'idée que l'oral serait pauvre mais bien entendu ils apparaissent à des rangs différents :

	Oral (CR)	Ecrit (PE)		Oral (CR)	Ecrit (PE)
seulement	1 ^{er}	15 ^e	finaleme nt	6 ^e	10 ^e
notamment	2 ^e	16 ^e	largement	7 ^e	58 ^e
également	3 ^e	6 ^e	particulièrement	8 ^e	31 ^e
actuellement	4 ^e	24 ^e	simplement	9 ^e	14 ^e
vraiment	5 ^e	1 ^{er}	rapidement	10 ^e	27 ^e

Tableau 6. Les dix adverbess les plus fréquents à l'oral et leur rang à l'écrit

Cette approche globale semble orienter vers l'hypothèse que la différence de distribution des adverbess peut être traitée en terme d'opposition oral / écrit.

En fait, les choses ne sont pas si simples, dans la mesure où si l'on creuse quelque peu les critères d'analyse et que l'on se situe au niveau 2 (une subdivision du corpus qui repose sur des critères externes), la distribution se présente sous un nouvel aspect.

La simple opposition écrit / oral n'est donc qu'en partie satisfaisante, tant il est vrai que la situation de parole va avoir une incidence sur les résultats. Ainsi, selon que nous avons affaire à un corpus de "parole privée", de "parole

professionnelle” ou de “parole privée”, on va pouvoir noter plusieurs phénomènes :

- Les adverbes les plus fréquents dans chacun de ces sous-ensembles ne sont pas identiques :

	Privée	Professionnelle	Publique
1-	vraiment	vraiment	évidemment
2-	justement	forcément	vraiment
3-	évidemment	énormément	effectivement
4-	tellement	effectivement	également
5-	finaleme nt	justement	justement
6-	énormément	également	notamment
7-	forcément	complètement	forcément
8-	complètement	essentiellement	finaleme nt
9-	autrement	simpleme nt	simpleme nt
10-	absolument	exactement	autrement

Tableau 7. Les adverbes les plus fréquents

- A l’intérieur de chaque sous-ensemble la part qu’occupe chaque adverbe est très variable. Si l’on prend l’exemple de l’adverbe *vraiment*, il représente 25 % des emplois d’adverbes en *-ment* dans la “parole privée”, 18 % dans la “parole professionnelle” et seulement 9 % dans la “parole publique”.
- Le corpus de “parole publique” se caractérise par une répartition beaucoup plus homogène des formes adverbiales, alors que celui de “parole privée” privilégie fortement l’emploi de *vraiment*.

Les adverbes en *-ment* apparaissent comme un critère intéressant qui permet de caractériser une production. Ainsi, le corpus de “parole publique” se rapproche du corpus de presse écrite avec lequel il partage cinq adverbes (sur les dix plus fréquents) : *vraiment*, *également*, *notamment*, *finaleme nt* et *simpleme nt*. Le corpus de “parole professionnelle” n’en partage que trois : *vraiment*, *également* et *simpleme nt*. Enfin, le corpus de “parole privée” n’en conserve que deux : *vraiment* et *finaleme nt*.

Ce résultat confirme ce que d’autres paramètres (le nombre de sujets lexicaux, le pronom relatif *dont*, le *ne* de négation, etc.) permettent de vérifier : la “parole publique” partage de nombreux traits avec les productions écrites formelles.

Le travail sur corpus permet ici de faire ressortir que les adverbes en *-ment* sont très sensibles aux situations de parole.

2.3. L’exemple de contre

Nous avons conduit notre analyse (Bilger et Cappeau : 2003) en explorant :

- CA (le français tout venant)
- HP (l’oral des hommes politiques)
- L’écrit de presse (*Le Monde*)

Le tableau ci-dessous présente les grandes tendances qui se dégagent :

Parlé divers	Hommes	Presse écrite (PE)
--------------	--------	--------------------

	(CA)	politiques (HP)	
Nombre de mots	1 000 000	300 000	1 000 000
Occurrences de <i>contre</i>	291	213	800 (400 retenues)
<i>par contre</i>	53 %	6 %	1 %

Tableau 8. La forme *contre* dans diverses banques de données

Ce premier tableau permet de révéler des faits de distribution qui sont masqués dans les ouvrages fondés essentiellement sur des corpus écrits. Deux points importants se dégagent :

a) La forme *contre* est bien moins attestée dans le français tout venant (corpus CA). Dans les deux autres corpus (HP et PE) le nombre des occurrences est assez proche. Cette forme se rencontre donc plus dans la langue formelle.

b) Si l'on regarde de plus près l'environnement de *contre* une première différence majeure se manifeste : c'est la locution *par contre* qui occupe la majorité des emplois dans le français tout venant. Dans le corpus écrit (PE), cette locution est quasi absente (d'ailleurs le TLF, qui s'appuie sur *Frantext* - corpus écrit littéraire - n'y accorde qu'une place marginale dans son article). Dans l'oral politique, cette expression est également faiblement utilisée.

Ce premier relevé montre donc, à nouveau, l'intérêt du travail sur corpus :

- la répartition dans les différentes banques de données ne pouvait être anticipée;
- il est nécessaire de disposer de données variées (l'oral formel des politiques se différencie de l'oral divers de CA) pour ne pas avoir une représentation déformée, fragmentaire ou réductrice notamment de l'oral (tout ne peut pas être ramené à la conversation familière).

Si l'on regarde avec plus d'attention les environnements syntaxiques dans lesquels se trouve *contre*, des tendances intéressantes se dégagent. Une différence sensible se manifeste entre les cas où *contre* introduit le complément d'un verbe et ceux où il introduit le complément d'un nom :

5- *c'est vrai que je aujourd'hui je peste contre contre la route hein*
(CA)

6- *comment pouvez-vous concilier aujourd'hui votre circulaire contre le tchador et le libéralisme* (HP)

Le tableau suivant permet de comparer ces deux emplois dans les trois banques de données examinées :

	CA	HP	PE
V+ <i>contre</i>	72 %	53 %	50 %
N + <i>contre</i>	28 %	33 %	49 %

Tableau 9. Les deux fonctionnements de *contre*

Le corpus écrit présente un équilibre entre ces deux fonctionnements (V + *contre* et N + *contre*). C'est pour le corpus du français tout venant que l'écart est le plus important et que l'emploi verbal domine. Le corpus des hommes politiques apparaît comme intermédiaire entre ces deux tendances, ce qui

montre là encore qu'une distinction trop grossière entre oral et écrit ne peut suffire à rendre compte de cette différence de distribution. On retrouve les tendances observées précédemment sur une certaine proximité entre "parole publique" et écrit formel.

Le tableau ci-dessous permet d'approfondir encore un peu plus les différences de réalisations syntaxiques de *contre* :

	Parlé divers (CA)	Hommes politiques (HP)	Presse écrite (PE)
V + <i>contre</i> SN	72 %	53 %	50 %
	<i>être</i> 21 %	<i>être</i> 38 %	<i>lutter</i> 17 %
	<i>avoir rien</i> 11%		<i>protester</i> 7 %
	<i>lutter</i> 5,5 %	<i>lutter</i> 16 %	<i>mettre en garde</i> 7 %
			<i>être</i> 5, 5 %
N + <i>contre</i> SN	28 %	33 %	49%
	<i>lutte</i> 25 %	<i>lutte</i> 43 %	<i>lutte</i> 34 %
			<i>crime</i> 19 %
			<i>attaque</i> 5 %

Tableau 10. Détail des deux fonctionnements de *contre*

Les associations préférentielles occupent une place importante dans les corpus. Ainsi *être contre*, *n'avoir rien contre* ou *lutter contre* représentent près de la moitié des emplois avec une tête verbale :

7- *je ne suis pas du tout contre une procédure de désignation d'un candidat unique (HP)*

8- *je n'ai rien contre Marguerite Yourcenar (CA)*

9- *le premier c'est lutter contre la pollution de la mer (CA)*

Les verbes constructeurs présentent une forte disparité entre le corpus de presse et le corpus d'oral tout venant. Le verbe *être* notamment passe de 5,5 % des cas à l'écrit à 21 % à l'oral. L'écrit possède une certaine spécificité de ce point de vue.

Pour les têtes nominales, l'écart est moins sensible entre ces productions. C'est le nom *lutte* qui est le plus utilisé (dans toutes les banques de données) :

10- *je souhaiterais qu'en dehors de la lutte contre l'inflation on prenne en considération la lutte contre le chômage à l'échelle à l'échelle européenne (HP)*

Enfin, il est intéressant de noter un emploi qui, semble-t-il, est réservé à la presse : celui dans lequel *contre* peut s'interpréter comme une sorte d'élément de coordination (un joncteur de liste). L'exemple suivant illustre cet emploi :

11- *le prix retenu est de 205F par action contre 185F au départ. (Presse)*

Cet emploi n'est pas attesté dans le corpus oral tout venant, et on en trouve très peu dans l'oral des hommes politiques (seulement 1%), alors que cela correspond à 24% des emplois de la presse écrite. Il est intéressant de noter que cet emploi n'est d'ailleurs pas recensé dans le TLF.

3. Pour conclure

Les relevés faits à partir de banques de données fournissent des informations irremplaçables sur les tendances de la langue (les dictionnaires

actuels qui ne s'appuient pas de manière systématique sur les données de corpus, ou seulement sur un certain type de corpus ne fournissent pas ce type d'indication).

Ce type de description pourrait d'ailleurs servir à renouveler la présentation de certaines unités grammaticales selon plusieurs directions :

- les données orales attestées permettraient de revenir sur certains *a priori* trop vite érigés en règles : on a vu que l'absence du sujet *nous* devait être pondérée par sa présence dans certains types de productions. D'autres travaux (Jeanjean : 1981, Blanche-Benveniste : 1994 et Blasco-Dulbecco : 1999) ont remis en cause l'idée que les formes sujets dédoublées comme *mon frère il* étaient massivement utilisées à l'oral alors qu'elles représentent en moyenne 12 % des sujets.

- des données variées permettraient aussi de mieux faire ressortir les liens entre les types de production et l'utilisation de certaines formes. La distribution des adverbes en *-ment* a permis de vérifier qu'il existait une différence dans les listes des adverbes les plus fréquents à l'oral et à l'écrit mais aussi que cette différence se retrouvait à l'oral quand on tenait compte du type de production.

- les collocations occuperaient une part plus importante et pourraient être organisées en tenant compte de leur fréquence d'apparition. Cette perspective évoquée dans la description de *contre* soulève la question des indications qui devraient figurer dans des ouvrages nouveaux à mi chemin entre les grammaires (un recensement des environnements syntaxiques) et les dictionnaires (les associations lexicales dominantes notamment).

De telles indications pourraient servir à mieux dégager les priorités à enseigner et à tenir compte de façon plus fine de ce qui est attesté.

La prise en compte des données de corpus renouvelle la conception que l'on peut avoir de la grammaire de la langue. De fait, on note régulièrement que les phénomènes grammaticaux sont limités à la fois par le lexique et par le genre. En ce sens, comme le dit M.A.K. Halliday (1991), il n'y aurait pas d'un côté un "système abstrait" et de l'autre des "réalisations individuelles", mais bien un système linguistique constitué du cumul de l'ensemble des usages.

Les grammaires actuelles sur le français présentent un inconvénient majeur : celui de ne fournir aucune information sur la productivité des phénomènes cités. Faire appel aux exemples attestés est certes une avancée, mais cela ne donne aucune garantie sur la qualité de la description elle-même, si celle-ci ne s'accompagne d'aucune donnée quantitative.

Références

BIBER , Douglas et alii. (1999). *Longman Grammar of Spoken and Written English*. Harlow. Pearson Education Limited.

BILGER, Mireille (Ed). (2000). *Corpus - Méthodologie et applications linguistiques*. Paris. Honoré Champion.

BILGER, Mireille (2003), " Quelques données sur les adverbes en *-ment* dans le corpus de référence de Français parlé ", *RSFP*, 18. Université de Provence.

- BILGER Mireille, et CAPPEAU, Paul. (2003). “ Les emplois de *contre* dans les corpus de français parlé et de presse écrite ”, *Recherches linguistiques*, 26 : 91-111.
- BLANCHE-BENVENISTE, Claire. (1994). “ Quelques caractéristiques grammaticales des ‘sujets’ employés dans le français parlé des conversations ”, in Yaguello, M. (éd), *Subjecthood and subjectivity. The status of the subject in linguistic theory*, 77-107. Paris : Ophrys; London : Institut français du Royaume Uni.
- CAPPEAU, Paul. (A paraître). *La réalisation des sujets syntaxiques dans les corpus oraux*, Actes du colloque international Variation, catégorisation et pratiques discursives, Université de Paris III (2002).
- HALLIDAY , M.A.K. (1985). *Spoken and written Language*. Oxford. Oxford University Press.
- HALLIDAY, M.A.K. (1991) “ Corpus studies and probabilistic grammar ” in Aijmer & Altenberg (eds), *English Corpus linguistics*, London-New-York, Longman. 30-45.
- JEANJEAN, Colette. (1981). “ L’organisation des formes sujets en français de conversation : étude quantitative et grammaticale de deux corpus ”, *RSFP*, 3 : 99-134.
- MOREAU, M-L. (1986). “ Les séquences préformées: entre les combinaisons libres et les idiomatismes. Le cas de la négation avec ou sans ne ”. *Le Français Moderne*. 3-4 : 137-160.