# Les corpus oraux : situation, exploitation linguistique, bilan et perspectives

Jeanne-Marie Debaisieux

# LES CORPUS ORAUX : SITUATION, EXPLOITATION LINGUISTIQUE. BILAN ET PERSPECTIVES

*Jeanne-Marie DEBAISIEUX*

*Université Nancy2*

RESUME

Après un bref panorama des principaux corpus existant dans le domaine du français parlé, l'article présente au travers des travaux de l'équipe Delic (Université de Provence) comment la prise en compte de données orales informatisées a permis le renouvellement de l'analyse syntaxique du français, en particulier en ce qui concerne les relations morphologie/syntaxe, les liens entre le lexique et la grammaire et l'impact du genre du texte sur la répartition des faits grammaticaux. L'auteur expose ensuite les limites des outils existants et les difficultés inhérentes à la constitution et à l'automatisation de corpus oraux échantillonnés.

L'objectif de la présentation qui suit[1] est de proposer un panorama non exhaustif de l'existant en ce qui concerne les corpus oraux en France et de présenter des perspectives et problématiques touchant tant à la constitution des corpus qu'aux possibilités d'exploitation qu' offrent les outils informatiques. L'exposé s'appuie essentiellement sur les travaux menés par l'équipe Delic de l'Université de Provence[2].

---

[1] Cette présentation a fait l'objet d'un exposé aux journées d'études : « Méthodes en Sciences Humaines : Les linguistiques de corpus » organisées par le Dea des Sciences du langage de l'Université Marc Bloch (Strasbourg) avec le soutien de l'Ecole doctorale des Humanités et de l'UFR de Lettres. Je remercie Catherine Schnedecker pour l'occasion qu'elle me fournit de le publier.

[2] Les travaux de l'équipe portant essentiellement sur ce sujet, il me sera impossible d'en donner une présentation exhaustive. J'espère néanmoins pouvoir en donner un aperçu fidèle en rendant aux auteurs ce qui leur appartient et en m'excusant pour les inévitables oublis que j'ai pu commettre, tant la matière est riche.

**A. La situation : les corpus oraux informatisés disponibles pour le français**

La situation en France en ce qui concerne les corpus de données orales se caractérise par un émiettement des données et un retard certain. Comme le signale J. Veronis (2000) : « Au moment où le British National Corpus propose 100 millions de mots étiquetés du point de vue grammatical et 10 millions de mots de parole transcrite, rien d'équivalent en France ». En effet il n'existe pas en France de corpus comportant à la fois de la langue orale et de la langue écrite, comme il en existe en Espagne, au Portugal, en Italie et bien sûr en Angleterre et aux Etats Unis[3]. La liste fournie en annexe est loin d'être exhaustive mais donne une image de la situation française : les corpus de langue parlée sont beaucoup plus restreints, il sont souvent de taille modeste et rarement consultables par des personnes extérieures à la recherche locale. Ainsi les corpus du Gars, (Groupe Aixois de Recherche en syntaxe - Université de Provence), de Lyon II ou de Paris III, souvent cités dans les travaux ne sont pas directement accessibles. On peut poser qu'il y a sans doute entre quatre ou cinq millions de mots effectivement disponibles mais l'absence de coordination rend l'exploitation de l'ensemble impossible. La Délégation Générale de la Langue Française a lancé récemment une enquête, sous la responsabilité de Paul Cappeau de l'université de Poitiers, qui permettra de mieux connaître l'ensemble des ressources et de fédérer l'existant. Actuellement, les corpus les plus importants en France sont ceux constitués par l'équipe Delic, (Description Linguistique sur Corpus - anciennement Gars). Le premier corpus en taille, plus d'un million de mots, est le corpus oral CorpAix, initié par le Gars. Pour des raisons juridiques et compte tenu des modalités de constitution de ce corpus dont les premières données ont été récoltées dans les années 70, ce corpus n'est pas diffusable. Dernièrement l'équipe Delic a produit « Le Corpus de Référence du Français Parlé » qui constitue le premier corpus francophone aligné et échantillonné et dont le lecteur trouvera une présentation détaillée dans le N° 18 de la revue « Recherches sur le Français Parlé ». Au-delà des ses imperfections, l'outil, qui est encore en cours de révision, constitue un premier témoignage de la langue française parlée aujourd'hui dans les principales villes de France. La présentation avec alignement offre en outre de nouvelles possibilités d'exploration. La plupart des autres corpus francophones importants se trouvent en Belgique et au Canada.

**B. Les corpus : possibilités d'exploitation**

Les recherches menées sur les corpus oraux portent sur différents domaines de la linguistique : on peut citer les travaux de l'équipe de Kerbart Orecchioni à Lyon sur l'analyse des interactions verbales, les travaux de Marie Annick Morel sur l'intonation et la structuration de l'oral spontané, les travaux de sociolinguistique de Françoise Gadet et les études

[3] Cf M. Bilger (2000a) pour une présentation des corpus oraux existants.

grammaticales menées par l'équipe Gars / Delic. C'est à ces dernières que nous attacherons dans cette présentation. On a en effet le sentiment que, dans le domaine de l'analyse grammaticale, la nécessité d'utiliser des corpus oraux n'est toujours pas reconnue par la communauté scientifique. Or il ne fait aucun doute pour qui travaille sur la grammaire d'une langue que la constitution d'un corpus oral est un outil précieux, ne serait-ce que par sa valeur heuristique. C'est à partir de l'écrit que nous avons développé une attitude réflexive et notre intuition de la langue est particulièrement limitée en ce qui concerne le domaine de l'oral. Comme le signalait déjà L. Blomfield (1927), traduit et cité par C. Blanche-Benveniste (1999) : « Nous n'avons pas appris à savoir ce que nous faisons quand nous parlons : structure des sons, grammaire réellement utilisée». Si l'on ajoute à cela le fait que beaucoup d'études linguistiques s'appuient sur des corpus écrits très normatifs et ignorent les productions émises spontanément par les locuteurs, on comprend mieux l'importance des divergences constatées entre notre intuition et la « réalité langagière » de l'oral.

Nous présenterons tout d'abord un premier exemple de ce décalage qui concerne le phénomène de la négation en français dont on sait qu'elle peut apparaître sous deux formes : la négation double « ne…pas » et la négation simple « pas ». On pourrait intuitivement analyser la différence entre les deux en termes de variation stylistique : la seconde serait une forme familière de la première et l'opposition :

*1) je ne la trouve pas méchante vs je la trouve pas méchante*

pourrait être décrite en termes de registre de langue. C'est la conclusion à laquelle aboutissent la plupart des études consacrées à ce sujet.[4] Néanmoins, un examen plus attentif des faits relevés dans les corpus aboutit à poser, au-delà de la différence formelle constituée par la présence ou l'absence du « ne », une différence de portée : la négation double, qui encadre le verbe, porte explicitement sur ce dernier, alors que la négation simple peut donner lieu à deux parenthésages : le premier commun à la négation double, le second soulignant la portée spécifique limitée au constituant qui suit.

je la trouve pas (méchante ) vs je la trouve (pas méchante)

On relève effectivement dans les corpus des exemples pour lesquels la portée « limitée à droite » de la négation est très nette et associée à une position particulière, séparée du verbe. Dans l'exemple qui suit :

*2). j'ai une seule fois dans ma vie (pas eu le trac)* ( Line Renaud – France Inter)

la structure est liée à une interprétation spécifique paraphrasable en « j'ai toujours eu le trac sauf une fois ». Dans ce cas, une formulation avec double négation, d'ailleurs impossible avec cet ordre des mots :

*2') *je n'ai une seule fois dans ma vie pas eu le trac*

[4] Voir à ce sujet l'analyse critique de Francoise Gadet, in Bilger (2000a).

aboutirait à un sens très différent. De tels exemples amènent à une réanalyse de la négation simple et en particulier à une remise en cause du statut de simple variante stylistique qui lui est conférée.

Dans cette optique, nous présenterons dans la section qui suit quelques exemples du renouvellement des études grammaticales que peut apporter la prise en compte des données de corpus oraux. Ce renouvellement concerne essentiellement : les rapports morphologie / syntaxe, les relations entre le lexique et la syntaxe et la prise en compte du rôle des « genres » au sens de Biber (1988) dans la description linguistique.

**1. Renouvellement des rapports morphologie /syntaxe**

En ce qui concerne le rapport morphologie / syntaxe, la démonstration portera sur une nouvelle approche des conjonctions dites de subordination au travers l'exemple de *parce que*. Les grammairiens ont relevé depuis fort longtemps à propos de ce morphème, qu'ils définissent par ailleurs comme « conjonction de subordination de cause », des emplois qu'ils ont du mal à analyser et dans lesquels *parce que* aurait plutôt une valeur équivalente à une conjonction de coordination de type « car » ou un adverbe argumentatif comme « en effet » C'est le cas dans l'exemple qui suit :

*3° Il est à la fac j'en suis sûre (parce que , car, en effet ) j'ai vu sa voiture dans le parking*

Ces exemples ne font pas l'objet de description particulière. Ainsi Muller (2002 : 91) après avoir signalé qu' « il faut réserver le statut de connecteur à la classe de mots comme *car* » et que dans ce cas, « le segment connecté forme obligatoirement un schème syntaxique dans son association avec une première phrase, au lieu de former un complément comme il le ferait avec une conjonction de subordination.»[5] se borne à constater :

« Une conjonction de subordination peut être connecteur lorsqu'elle ne construit pas un complément adverbial intégré à la phrase antérieure :
*Il fait beau, parce que Paul se promène* »
Il n'y a pas de lien de cause à effet . […]. Mais la frontière reste difficile à délimiter ».

On pourrait attribuer cette absence de description d'un emploi[6], qui remet pourtant en cause une opposition *parce que* (subordination) vs *car* (coordination) souvent jugée fondamentale dans les grammaires, au fait que ces usages sont minoritaires. Or une étude quantitative (Debaisieux, 2002) a montré que ces fonctionnements dits marginaux sont en fait majoritaires, (ils constituent plus de 85%. des 6000 exemples de CorpAix, et ce dans des situations de communication diverses : conversations mais aussi entretiens assez formels sur France Culture), et qu'ils diffèrent de l'emploi considéré

[5] C'est nous qui soulignons.
[6] On ne trouve mention de ce fonctionnement ni dans la grammaire de Riegel, ni dans celle de Le Goffic.

comme référence à la fois par des caractéristiques formelles et sémantico-pragmatiques. On relève ainsi des emplois

**- en début d'intervention**

4) *L1 papa ferme papa tu fermes*
*L2 (*s'approche et pousse la porte*) parce que tu peux pas le faire toi* (o.d)[7]

**-avec des structures interrogatives**

5) *mais l'attitude des médecins est incompréhensible parce que est-ce qu'ils ne devraient pas eux–mêmes voir comment se passe cet accueil ( MEDSOC)*

Ce qui est impossible dans les cas canoniques de subordination

**- avec des effets d'emboîtements et de portée de la conjonction sur plusieurs énoncés qui s'écartent du schéma canonique Cv parce que CV.**

6) *moi je suis relativement optimiste hein / **parce que** je pense que actuellement nous avons une génération qui est la pire / **mais** ils vont avoir des enfants à leur tour **et** la troisième génération ils seront français* (écrivain - France Culture)

La proposition introduite par la conjonction de subordination couvre trois propositions coordonnées par « mais » et par « et », ce qui ne serait pas possible dans une structure subordonnée. Du point de vue sémantique, on a pu montrer[8] qu'on ne pouvait pas rendre compte de ces emplois en termes de cause :

7) *biner ça désherbe le terrain parce que mine de rien les herbes tels que le liseron ou les mauvaises herbes à grandes feuilles ça ça étouffe le ça étouffe le la graine quoi la la plante (Jardinage)*

Il n'y a pas dans cet énoncé de rapport de cause entre les contenus présentés dans les deux propositions (le fait que les mauvaises herbes étouffent la graine n'est pas la cause du fait que « biner désherbe le terrain ») mais une relation entre deux assertions : celle qu'introduit *parce que* donnant de la pertinence a posteriori à celle qui précède.

La marginalisation par les descriptions classiques d'emplois majoritaires dans les usages attestés est d'autant plus gênante que ce comportement ne paraît pas limité à une langue. Des études comparatives sur l'espagnol et l'italien aboutissent à des résultats similaires[9]. Des faits comparables sont décrits pour l'anglais par J. Miller (1998) qui oppose " les « because CV » antéposables «vraies subordonnées » où l'ordre des mots est contraint, aux « because CV » « principales » où l'ordre des mots est libre. En allemand, la différence est clairement marquée par le non rejet du verbe en fin de proposition postposée introduite par *weil*, comme le montrent les exemples cités par Hannes Scheutz (1998).

---

[7] Observation directe.
[8] Cf. Debaisieux J.M. (à paraître)
[9] cf Debaisieux J.M. et Deulofeu J.(à paraître).

Au-delà de la description des usages du morphème, ce comportement amène à interroger plus largement la relation entre morphologie et syntaxe et en particulier la corrélation obligatoire établie généralement entre la présence d'une marque morphologique et l'existence d'une relation syntaxique. On en arrive en effet à l'idée qu'il y a des *parce que* subordonnants et des *parce que* connecteurs discursifs aux caractéristiques syntaxiques et sémantico-pragmatiques distinctes. L'analyse permet également d'objectiver l'opposition comportement central vs comportement marginal d'un morphème, dimension par rapport à laquelle notre intuition est souvent mise en défaut, compte tenu de l'écart souvent important entre les données trop sélectives de l'écrit normé et les productions spontanées des locuteurs. Ainsi un élément comme *quant à* auquel on accorde[10] un statut d'introducteur prototypique de topique à partir d'énoncés du type :

*8) Quant à Paul, il est parti*

se trouve quasiment absent des corpus oraux et paraît réservé à certains types de discours. On en relève 23 occurrences dans CorpAix, majoritairement dans des plaidoiries d'avocat. En outre les emplois à l'oral ne semblent pas soumis aux contraintes qui sont attachées aux emplois à l'écrit. On relève ainsi des exemples du type :

*9) pour aborder ce sujet i- il convient de + de s'attacher à à deux aspects euh je dirai(s) dans un premier temps la richesse de de la langue française et dans un deuxième temps sa ses faiblesses et et son déclin +* ***tout d'abord quant à la à la richesse de la langue française*** *+ elle se manifeste (euh, à) à plusieurs niveaux + [DROIT]*

dans lequel *quant à* introduit un élément premier d'une liste. Ce qui contredit les résultats des études portant sur l'écrit qui s'accordent sur le fait que le morphème ne peut introduire un premier élément de liste. Compte tenu de la rareté des données, il est bien sûr impossible de tirer de véritable conclusion. Néanmoins, on peut raisonnablement penser que *quant à* constitue en français à un élément à valeur stylistique marquée et non un élément structural indispensable. L'observation des corpus oraux livre d'ailleurs de nombreux exemples où le topique n'est introduit par rien ou par l'élément *pour* du moins dans les discours de type spontanés : Ainsi dans l'extrait suivant :

*10) °on était une soixantaine on avait juste un lit bon une petit table de nuit c'était tout ce qu'on avait comme matériel [..] et autrement* ***la nourriture*** *euh* ***la cuisine*** *euh il y avait des braves euh bonnes soeurs enfin des dames qui s'occupaient de la cuisine + par contre* ***pour le ménage*** *c'était les enfants qui le faisait [...] on se levait très tôt en général et euh* ***le dortoir c'était des lavabos*** *c'était pas comme maintenant c'était des lavabos métalliques* ( Internat)

Deux des trois topiques, *la cuisine* et *le dortoir,* ne sont pas « introduits » : et le troisième, le *ménage* est introduit par *pour*.

[10]Cf. Combettes, B. (1999) et Prevost, S. (2003).

La prise en compte des données orales permet donc de nuancer les résultats obtenus à partir de l'écrit ou des intuitions des locuteurs, qui présentent souvent une image partielle de la langue, caractérisée par l'absence de formes non standard et la marginalisation de structures dont on vient de voir qu'elles peuvent être centrales dans les productions langagières des locuteurs. Un autre apport du recours aux données orales, que nous allons présenter dans la section suivante, et d'avoir mis en évidence la forte corrélation existant entre faits lexicaux et faits grammaticaux. Ce point a fait l'objet de nombreuses publications.[11] Nous ne présenterons ici que quelques exemples.

**2. contraintes lexicales sur les règles syntaxiques**

L'interrogation informatisée de grandes distributions a permis de montrer que de nombreux faits grammaticaux sont en fait contraints par des aspects lexicaux. On a observé[12], par exemple, que l'emploi du « en » marquant la possession pour les choses comme dans l'exemple suivant

*11) j'en vois la cheminée (de cette usine)*

est rare en français parlé et semble restreint à quelques noms : odeur sensation

*12) .j'en ai encore l'odeur dans les narines*

ou à des expressions quasi figées :

*13. j'en voit le bout,*
*il en connaît un rayon*

A propos de *dont*, l'analyse statistique a également permis de montrer[13] que les usages en
conversation étaient limités à un petit nombre de lexèmes verbaux ou nominaux. Pour illustration, l'interrogation d'un corpus personnel de 250.000 mots aboutit aux résultats suivants[14]. Sur 65 occurrences relevées dans le corpus :

- lorsque *dont* est construit par un nom, on relève dans 17 exemples 27, soit 63% des occurrences, le substantif *façon* ou *manière* :

*14) on fait un premier point sur la façon dont les choses se sont passées [job]*
*la façon dont nous laissons faire les vacataires [retraites]*
*la façon dont on réagit face aux problèmes [auto]*
*je suis très très content de la manière dont s- s'est passé le concert [groupeluc]*

- lorsque *dont* est construit par un nom, on relève 11 exemples sur 26, soit 42% des occurrences, avec le verbe *parler* ou un synonyme :

*15) le monstre dont dont je parlais tout à l'heure [job]*

---

[11] (cf. notamment C. Blanche-Benveniste, (1994, 2000b, 2000c).
[12] cf. Blanche-Benveniste (1990, 1997, 2000c)
[13] cf. Blanche-Benveniste (1997, 2000c)
[14] Compte tenu de la taille du corpus interrogé, ces chiffres ne sont fournis qu'à titre indicatif. Ils confirment néanmoins les tendances relevées.

*c'est le prêtre dont je disais /qu'il, qui/ a été renvoyé de Saint P* [seminaire]*
*la déconnexion dont tu parlais c'est aussi le prix de la production [retraites]*

Ces exemples révèlent bien la façon dont certains faits grammaticaux sont liés à des contraintes lexicales. Il y a certainement de nombreux phénomènes à observer selon cette optique, mais l'analyse se heurte au problème de l'insuffisance des données.

**3. Les limites actuelles de l'utilisation des corpus**

Certaines études demanderaient, pour être menées à bien, des corpus plus importants. Ainsi si l'on peut aujourd'hui faire une analyse des usages écrits de *bien que* et *quoique*, grâce notamment au corpus Frantext, aucune étude sérieuse, qui permettrait par exemple de généraliser les observations faites dans le cas de *parce que,* ne peut être menée sur l'emploi des mêmes conjonctions dans différents types de corpus oraux, et ce par manque d'exemples, (moins de 20 exemples dans CorpAix), alors même que l'on a l'intuition d'un décalage important entre les usages de ces morphèmes et les descriptions classiques qui en sont données. M.A. Morel, (1996 : 46) note ainsi à propos des usages de *bien que* :

« l'emploi du subjonctif est obligatoire dans la subordonnée ouverte par *bien que* » […]Les entorses modales […] relevées par les grammairiens se rencontrent le plus souvent lorsqu'il s'agit d'un fait passé […] ( indicatif imparfait derrière bien que) – ou bien lorsqu'il s'agit d'un simple envisagement par la pensée d'une situation fictive (conditionnel ou futur derrière *bien que*).»

Les exemples cités par les grammaires sont effectivement le plus souvent construits sur le modèle où « la concession logique » est exprimée par *bien que* en tête d'une construction verbale au subjonctif :

*16) Bien qu'il ait passé des années dans ce pays il ne sait pas en parler la langue* (M. Riegel, 1994 : 513)

Or la tendance relevée dans les corpus oraux semble être à la postposition de *bien que* introduisant une construction à l'indicatif :

*17) généralement les mâles sont aussi plus beaux et colorés dans la plupart des espèces bien que chez les poissons comme les Trochogaster Leeri ils sont exactement pareils (aquarium)*

Le phénomène semble relever d'une différence syntaxique plus générale que la simple opposition morphologique de mode. En effet, *bien que* avec l'indicatif fonctionne dans ce cas comme un connecteur discursif plus que comme un subordonnant : il introduit une énonciation à valeur de restriction par rapport à l'énonciation précédente et non pas une proposition porteuse d'une valeur de « concession logique ». La confirmation de cette tendance pourrait renforcer l'hypothèse du double fonctionnement des conjonctions dites de subordination (subordonnant et connecteur discursif), mais le manque de données rend cette recherche impossible aujourd'hui. On

pourrait également citer le cas des conjonctions de conséquence *au point que, de telle sorte que* qui sont peu employées dans la plupart des corpus oraux par rapport à *parce que, quand* ou *pour* et qui semblent en outre concurrencées par *ça fait que, ce qui fait que*.

Il ne faudrait pas conclure de ce qui vient d'être évoqué que le fait d'avoir de grands corpus avec des données « tout venant » soit la seule condition permettant de résoudre les difficultés actuelles de la description grammaticale de la langue. B. Habert (2000) dénonce ainsi l'idée qu'il suffirait « d'un élargissement mécanique des données mémorisables [pour produire] inévitablement un échantillon de plus en plus représentatif de la langue traitée ». De nombreux travaux ont en effet montré l'importance de la prise en compte des « genres » de texte pour une meilleure représentativité des données[15].

**5. Distribution des marqueurs linguistiques par « genres »**

Si l'on veut faire la grammaire des usages réels d'une langue, il est nécessaire d'observer des situations diversifiées de communication : les structures syntaxiques et le lexique ne sont pas distribués de la même façon dans une conversation, un discours politique, un écrit juridique, ou un roman. On peut travailler sur des échantillons représentant ces différents types de textes et en explorer les régularités. C'est souvent ce qui a été fait pour l'écrit.[16] Plus récemment, les travaux de D.Biber (1988, 1995) sur l'anglais exposent une démarche plus inductive qui consiste à dégager des convergences de traits syntaxiques et lexicaux pour aboutir à une typologie interne. Un type de texte « se définit alors par la cooccurrence d'un certain nombre de traits linguistiques (et éventuellement par l'évitement systématique d'autres traits).» (Habert op.cit). Pour le français, le travail reste en grande partie à faire d'autant qu'on ne dispose pas de corpus échantillonné important mais les analyses menées par l'équipe aixoise[17] ont montré depuis fort longtemps des pistes allant dans ce sens. Nous en présenterons ici quelques exemples.

Ainsi les enregistrements dans lesquels les locuteurs sont amenés à présenter des explications sur des savoir-faire ou des techniques présentent de nombreuses structures avec des dépendances parfois plus complexes que celles que l'on pourrait relever à l'écrit comme le montrent les exemples suivants, extraits respectivement d'une explication sur le ferrage des chevaux et d'une présentation du métier de journaliste :

---

[15] Cf. Habert (op.cit.) pour une présentation des différentes typologies utilisées dans les travaux récents.

[16] Cf. en particulier les travaux de J.P. Adam (1990,1992)

[17] Cf. en particulier, C. Blanche-Benveniste, 1994, 1997, 2000, P. Cappeau , 2001, M. Bilger et C. Blanche-Benveniste ,1999)

18) *ben les chevaux ben on les ferre parce que quand on les fait travailler comme on les fait travailler là / qu'ils restent pas au pré à manger normalement à se déplacer normalement / le sabot il s'use et comme le sabot du cheval c'est la partie la plus sensible ben il pourrait plus marcher au bout d'un moment* (Cheval)

19) *maintenant pour euh: pour parler deux minutes de technique(s) journalistique(s) il est vrai que quand on: on doit interviewer quelqu'un euh: c'est-à-dire vraiment aller à la rencontre de quelqu'un en particulier pour lui poser plusieurs questions précises sur une activité qu'il développe sur euh sur sa vie ou sur je ne sais quoi euh: c'est quelq- là c'est quelque chose qui est préparé* ( Journaliste)

Ces exemples s'opposent à la fois à l'idée de la prédominance, dans les phénomènes de variation, de la dimension diastratique, puisque le premier extrait est produit par un locuteur ayant suivi une scolarité obligatoire alors que le second émane d'un locuteur possédant un niveau d'études supérieures et à l'idée d'un oral qui de façon générale ferait appel à la parataxe ou la juxtaposition plutôt qu'à l'hypotaxe ou la subordination. Une autre caractéristique relevée dans les structures explicatives est la présence de sujets complexes ( en gras dans l'exemple) :

20) *.**l'avantage de construire des lignes qui fonctionnent avec un courant alternatif** est de pouvoir augmenter facilement la tension et de la baisser facilement en fonction des besoins (technicien)*

Le sujet est constitué d'un nom constructeur d'un infinitif dont le complément construit une proposition relative. Cette complexité va à l'encontre de certaines généralisations qui postulent que les sujets lexicaux sont quasiment absents à l'oral.

Les textes où les locuteurs exposent des récits récits de vie, quant à eux, se caractérisent par un emploi très fréquent de paroles rapportées au style direct (en gras dans l'exemple suivant) :

21) *nous sommes allés jusqu'à - Corre chez les - Vautrin là chez la belle-sœur - la soeu- la fille de papa - et puis elle était déjà partie elle avec ses gosses - il y avait plus que mon beau-frère tout seul ben - papa lui dit **qu'est-ce que vous faites - oh ben si ça ne vous fait rien je vais partir avec vous** - qu'il dit* **ben d'accord** *- nous voilà repartis - pour descendre dans le sud mais arrivés à Amance - il y avait les autres voitures qui remontaient - qui criaient - **ne descendez pas - c'est pas la peine les Allemands sont là-bas c'est pas la peine de descendre les Allemands sont là-bas** alors - on s'arrête et puis François dit **oh moi je euh je m'en vais je m'en vais il faut que je retrouve Jeanne et puis mes gosses je veux je veux pas rester** - ben - mon papa dit **moi je veux pas descendre hein euh - on r- on retourne et puis c'est tout** - alors il était tard on a couché dans une ferme - et puis on est repartis le lendemain matin* (Guerre)

Ces exemples illustrent bien la nécessité de prendre en compte « les genres » si l'on souhaite aboutir à une meilleure appréhension des faits de langue en dépassant notamment l'opposition souvent simpliste posée entre un écrit, dont on ne sait pas très bien par quoi il est représenté et un oral

réduit à la seule modalité d'oral spontané de conversation. La non prise en compte de la notion de genre peut par ailleurs aboutir à des généralisations abusives : des tendances observées sur un ou deux « genres » sont présentées comme des règles du système grammatical dans son entier. Ainsi dans une étude censée illustrer " [the] increasingly wide difference between spoken and written (European) French, [in particular ] between the every conversational language [...] and the culturally prestigious written language of great texts[...]", B. Fonseca-Greber et L.R. Waught (2003) proposent une analyse des emplois des pronoms personnels en français dans un corpus de « Everyday Conversational European French » comportant 194. 000 mots. Après avoir constaté la disparition du pronom *nous* auquel se substitue dans 99% des cas le pronom *on*, les auteurs s'interrogent sur l'impact de ce phénomène quant à la valeur traditionnelle d'expression de l'indéfini de *on*. Elles notent que cette valeur, qu'elles illustrent par l'exemple suivant :

« M : parce qu'on-nous-avait volé les clés de la maison je-dis y a longtemps »[18]

ne constitue que 5,7% des occurrences. Ces chiffres confirment selon les auteurs "that a radical shift in meaning has taken place» Le pronom aurait fait l'objet d'un glissement sémantique qu'elles expliquent ainsi :

« There as been a change from the indefinite meaning as basic and personal meaning as stylistical marked, to one personal meaning first person plural as the new basic meaning, the indefinite meaning as marginal, and the other personal meaning as still highly stylistically marked ».

Face à ce qu'elles nomment « the decline in the use of *on* for indefinite meaning », les auteurs s'interrogent sur la façon dont les locuteurs expriment la valeur indéfinie et constatent que dans leur corpus, cette valeur est exprimée dans 68% des cas par le pronom *tu*. Ces observations les amènent à conclure en ces termes :

"[an] another radical semantic shift has occurred in how speakers choose to express indefinite meaning. [...] A the present time, the balance has clearly shifted away from *on-* and toward *tu-* as the preferred way of expressing indefinite meaning in ECEF, wathever the social context".

Il existerait donc dans ce domaine « a radical difference between the written and spoken language", ce qui justifierait l'hypothèse d'une « possible diglossia in European french ».

L'article présenté ici très brièvement mériterait une analyse approfondie que nous n'aurons pas loisir de mener. Nous nous contenterons de présenter quelles chiffres tirés d'un corpus personnel et qui nous suggère plus de prudence dans l'analyse du phénomène. Il s'agit tout d'abord de deux corpus de conversations. Dans le premier, (Activpro) deux sujets sont essentiellement abordés qui touchent les activités de deux des trois locuteurs : l'un est chanteur amateur et raconte son dernier concert, l'autre

[18] Nous reprenons ici la transcription des auteurs.

vient de trouver un emploi d'agent immobilier. Dans l'autre corpus (Macdo2), la conversation porte essentiellement sur le récit d'un différent qui a opposé une des interlocutrices à son chef. Les deux corpus sont de taille similaire : 12.115 mots pour le premier, 11.519 mots pour le second et durent environ 10' chacun. Les protagonistes ont sensiblement le même âge, entre 25 et 30 ans. On peut dire que ces corpus relèvent de ce que Fonseca et Wauth nomme « the everyday conversational european french). Nous présentons dans le tableau suivant les différentes occurrences de *on* et de *tu* ainsi que les interprétations, (pronom de 2° personne. ou valeur indéfinie) qui leur sont attachées et des exemples de chaque corpus.

***On* et *tu* dans deux corpus de conversation**

| | **Macdo2** | **ActivPro** |
|---|---|---|
| nb. total de *on* | 30 | 21 |
| *on* pro.2°pers. | 17 | 21 |
| *on* indéfini | **13** | **0** |
| nb. total de *tu* | 38 | 44 |
| *tu* pro.2° pers. | 37 | 34 |
| *tu* indéfini | **1** | **10** |

*On* pronom 2°personne :

22) *déjà ce manager on s'entendait pas [macdo2]*
*on est on est payé à peu près de la même façon alors [macdo2]*
*alors on a joué pendant pendant une heure et demi comme ça [activpro]*
*on était deux sur la scène [activpro]*

*On* valeur « indéfini » :

23) *quand on me dit tu fais ça je fais rien du tout [macdo2]*
*j'ai parlé j'ai parlé on me calmait [macdo2]*

*Tu* pronom 2°personne

24) *il me dit non il y a rien à faire tu restes en caisse [macdo2]*
*alors il a insisté euh non tu vas en pause [macdo2]*
*je sais pas si tu connais rue de la Hache angle rue de la Hache [activpro]*
*et donc tu travailles en en en comment dire en en en partenariat [activpro]*

*Tu* valeur « indéfini »

25) *c'est le genre quand tu provoques par exemple quelqu'un de ma famille euh moi aussi je renforce [macdo2]*

*tu peux avoir des rendez-vous à six heures à huit heures [activpro]*
*la seule contrainte c'est que tu es coincé du du lundi au samedi [activpro]*

On constate que sur l'ensemble des deux corpus, les emplois indéfinis de *on* constituent 25% des occurrences relevées alors que les emplois indéfinis de *tu* ne représentent que 13 %. La conversation qui tourne autour du récit de l'incident de travail ne comporte en effet qu'une seule occurrence de *tu* à valeur indéfinie, alors qu'elle comporte 13 occurrences de *on* avec cette même valeur, qui sont liées le plus souvent à des commentaires de l'interlocutrice du type « alors en fait elle aime qu'on lui marche dessus aussi quoi ». Les nombreuses séquences de paroles rapportées entraînent quant à elles l'apparition du *tu* de 2°pers. Le second corpus se distingue du premier puisque les *on* de type indéfini en sont absents et que l'on relève 10 occurrences de *tu* qui ont cette valeur. L'hypothèse que nous proposons est que cette différence peut être interprétée en termes de type d'interaction. Une rapide recherche dans deux autres corpus aboutit en effet à constater une tendance assez similaire : dans un corpus de 35 minutes de type récit de vie, mené par une locutrice de plus de 60 ans auprès de sa petite fille, on relève, pour un total de 132 *on,* 87 occurrences à valeur de pronom de 2° pers contre 45 à valeur d'indéfini et ce selon une répartition nette : les premiers introduisent des verbes d'action ou de description à l'imparfait :

*26) alors là c'était agréable on faisait des bonnes parties de - de jeux de cartes les soirs[Ginette]*

les seconds, des commentaires généraux au présent :

*27) alors que aujourd'hui on ne fait plus tout ça parce que les les jeunes vont travailler plutôt dehors et puis euh on a plus le caractère [Ginette]*

Les 18 occurrences de *tu* relevées dans ce même corpus correspondent toutes à une valeur de pronom de 2° pers. A l'inverse, dans un corpus de durée similaire, mais dans lequel un locuteur également âgé, explique en quoi consiste le jardinage, on relève un total de 95 *tu* dont 66 à valeur « indéfini » présentant essentiellement des séries de verbes d'action :

*28) et au printemps tu passes le la moulinette tu rassites tu euh rass- tu ratisses tu sèmes p- tu tu plantes tu bines tu arroses [jardinage]*

qui correspondent aux explications fournies par le locuteur quant aux taches à accomplir dans un jardin. Le même texte présente 27 occurrences de *on,* dont 19 à valeur d'indéfini qui introduisent soit des commentaires d'ordre général (cf27) soit des commentaires métalinguistiques( cf.28).

*29) ah on ne le répétera jamais assez [jardinage]*
*30) comment on appelle ça [jardinage]*

On ne peut, sur ce petit échantillon, parler de déclin de l'usage de *on.* On note par contre que l'emploi de *on* ou de *tu* à valeur « indéfini » n'est pas insensible au type de texte, l'usage de *tu* semblant être lié à l'exposition d'explications techniques, pour lesquels le locuteur semble rechercher l'implication de son interlocuteur. Il n'est pas question de remettre en cause

l'ensemble de l'analyse de B. Fonseca-Greber et L.R. Waught sur la base de données aussi succinctes. Il s'agit plutôt pour nous de montrer que la catégorie « everyday conversational » établie par les auteurs est trop vaste et qu'elle ne permet pas de rendre compte des divergences constatées dans la répartition de certains phénomènes, qui sont liées à la nature « textuelle » de l'interaction. Il faudrait bien sûr compléter l'analyse mais en l'absence de précisions quant à la typologie des corpus utilisés par les auteurs, il nous paraît difficile de parler d'une évolution générale à l'oral. Il semble d'abord nécessaire d'analyser précisément certaines corrélations entre traits grammaticaux et genre de texte afin de mieux saisir la portée des phénomènes relevés.

L'établissement d'une typologie interne des textes s'appuyant sur la convergence de traits lexicaux et grammaticaux nécessite des analyses fines. En effet les constructions qui sont prises comme indicateurs de genre doivent être soigneusement définies. Ainsi, une forte proportion de passif et de nominalisation pourrait être considérée comme un trait permettant de situer un texte dans un genre narratif éloigné d'un oral spontané. Or dans une étude de 1997, Claire Blanche-Benveniste montre que l'on trouve dans des textes oraux de nombreux verbes psychologiques employés à la forme passive dont l'agent est introduit par « de »

*31) j'ai été choqué de cela, il a été surpris de sa réaction , j'ai été frappée ect.*

Ces passifs en *de* « sont très nombreux dans tous les passages où il est question de sentiment ». Un comptage fait à partir d'un ensemble de récits et de conversation et d'une collection du Monde Diplomatique fait apparaître 25 occurrences d'emplois passifs du verbe *frapper* dans les récits, 2 dans les conversations alors qu'on ne relève aucun emploi dans les écrits journalistiques. Ce passif-là est donc un phénomène fortement conditionné par le contexte où il apparaît et étroitement lié aux caractéristiques lexicales des verbes. En ce qui concerne les nominalisations, C. Rouget (2000) montre qu'il est nécessaire de distinguer deux types de nominalisations prédicatives (ie un déverbal construisant un argument) et que seules celles du type 1 tel que :

*32) cela justifie <u>la construction</u> de grands télescopes »*

qui se caractérisent par les déterminants qui les introduisent et le fait d'apparaître dans toutes les positions syntaxiques, sont significatives quant au style « écrit technique ». Pour aboutir à une classification fine des textes, il faut donc considérer les faits grammaticaux, non pas dans leur globalité, mais de façon « morcelée » (C. Blanche-Benvensite, 2000a) en fonction des relations qu'ils entretiennent avec certains traits lexicaux.

**7. L'héterogéneité des pratiques langagières**

Un autre phénomène révélé par les analyses sur corpus est lié à ce qu'on a constaté depuis longtemps, c'est-à-dire la diversité de la compétence des locuteurs. Pour reprendre la citation de Berrendonner (1983) : « Tout se

passe comme si un locuteur donné était capable de changer de compétence en fonction des diverses situations de discours auxquelles il se trouve confronté.» Une manifestation de cette diversité apparaît dans ce que Blanche Benveniste (1990, 1999) appelle « la langue du dimanche », c'est-à-dire la recherche par les locuteurs d'un registre soigné, qui se caractérise par toute une série de traits : les *parce que* transformés en *car* les *quand* en *lorsque*, la présence de négation double etc. On trouve par exemple ce type de phénomènes au début des enregistrements. On relève ainsi dans le passage suivant extrait d'un récit de voyage :

*33) et justement nous avons demandé comment ces gosses savaient -ils déjà si tôt travailler le bois et le rendre si beau / parce que ce qu'ils faisaient c'était de magnifiques figures sur le bois exactement comme on en voit sur nos plus beaux meubles et bien il nous a été répondu que l'usine ne formait pas les enfants / pour travailler à l'usine il fallait qu'ils sachent travailler déjà* (homme 49 ans - scolarité obligatoire)

- une inversion du sujet : « *comment ces gosses savaient -ils déjà si tôt travailler le bois* » qui aboutit d'ailleurs à un énoncé non standard, puisque l'inversion se situe dans une interrogation indirecte.
- une forme passive inattendue dans ce registre : « *il nous a été répondu* »
- une négation double : « *l'usine ne formait pas les enfants* »
- des adjectifs et superlatifs antéposés : « *de magnifiques figures, nos plus beaux meubles* ».

Ce que cet exemple permet également de constater, c'est le caractère hétérogène de cette mini-production : le lexique sujet du verbe dans la construction avec inversion, « les gosses » relève, lui, d'un registre peu soutenu. Autant de points qui montrent que la constitution de corpus en une typologie inductive basée sur des critères internes doit s'appuyer sur des analyses extrêmement fines.

Devant la richesse des perspectives ouvertes par l'analyse de corpus, on peut s'interroger sur les raisons du retard dans la constitution de corpus de données orales dans le domaine français. Nous fournirons ici, sans prétention explicative, quelques éléments de réponse. Outre le poids des représentations sur la langue de la part de la communauté des linguistes qui font que l'oral est encore trop souvent jugé comme peu digne d'analyse, la situation est liée en partie aux difficultés de constitution de corpus oraux.

### C. Problèmes d'établissement des données

Le premier problème est lié au recueil de données et à la difficulté d'avoir accès à des situations diversifiées : on manque par exemple de situations « agoniques »,[19] les conversations à plusieurs participants présentent des difficultés techniques, certains usages professionnels, par exemple les plaidoiries d'avocats, sont, pour des raisons de confidentialité,

[19] Le terme est repris de Andre LarocheBouvy (1984 )

peu accessibles. F. Gadet (2000b) signale ainsi « il est un ensemble dont le recueil pose des problèmes spécifiques, souvent difficiles à résoudre : ce sont les énoncés de vernaculaire où prédomine le parler familier, recueillis dans des contextes ordinaires ». Outre ces difficultés, se pose le problème de la transcription des données orales. Il existe aujourd'hui des logiciels d'aide à la transcription[20] mais l'essentiel du travail ne peut être automatisé et la transcription des données reste une activité extrêmement coûteuse en temps[21], d'autant qu'elle doit être vérifiée par plusieurs personnes pour atteindre un bon degré de fiabilité. En outre, compte tenu de la complexité de la relation graphie phonie en français, la transcription demande un minimum de savoir sur la langue et sur les pièges à éviter. Nous présentons ici rapidement quelques exemples de ces difficultés qui ont fait l'objet de nombreuses publications. [22]

Certaines séquences sont particulièrement sujettes à confusion et demandent beaucoup de rigueur de la part du transcripteur : il s'agit par exemple des séquences en « qui /qu'il »[23], illustrées par l'exemple suivant transcrit par un étudiant :

*34) ben les chevaux ben on les ferre parce que quand on les fait travailler comme on les fait travailler là / **qui restent** pas au pré à manger normalement à se déplacer normalement / le sabot il s'use*

Le transcripteur s'en est tenu à la prononciation entendue [qui] qui est une prononciation banale. Cette transcription revient à proposer une analyse syntaxique qui n'est pas la seule possible et certainement pas la plus vraisemblable puisque qu'elle analyse la séquence comme une relative prédicative, alors même que ce type de relative est réservée à certains verbe du type : « je le vois qui arrive ». Une analyse plus plausible consiste à poser que l'on a affaire à deux morphèmes : la conjonction *que* qui reprend la subordonnée en *quand* et le pronom clitique *il*. On proposerait donc pour le passage la transcription suivante :

*quand on les fait travailler comme on les fait travailler là / **qu'ils restent** pas au pré à manger normalement*

On voit comment une erreur de transcription peut aboutir à la production d'un énoncé à la grammaticalité douteuse. Les difficultés de la transcription sont également liées au fait qu'il ne s'agit pas d'une simple activité de discrimination et/ou de segmentation. La transcription est une activité d'interprétation qui amène le transcripteur à reconstruire le sens et il est

---

[20] On peut citer : Trancriber : www.ldc.upenn.edu/mirror/Transcriber/ et Winpitch : http://www.winpitch.com/

[21] A titre indicatif, on considère qu'il faut une demi-heure de transcription pour 1 minute d'enregistrement.

[22] De nombreux articles traitent des problèmes de transcription de façon détaillée. Cf en particulier le n° 14 de *Recherches sur le Français Parlé.*

[23] cf.Bilger (2000c)

parfois impossible de trancher entre deux interprétations. C'est le phénomène que l'on nomme « double écoute » et qui est systématiquement conservé dans les transcriptions aixoises. On relève des faits de double écoute très fréquents sur des distinctions impossibles

- entre deux consonnes

*35) c'est **/les, des/** sages-femmes qui vont accueillir ben des gens qui ont des problèmes de de d'infertilité*

- entre deux voyelles

*36) pour un premier - et ben euh cinq jours et ben on va /**les, le**/ suivre tous les jours*

On trouve également des phénomènes de double écoute pour des segments à première vue très différents :

*37) enfin ce procédé voilà c'est simplement euh c'est simplement une fenêtre en fait qui se promène donc c'est dans un train et l'idée **/derrière, de regarder/** le train c'est quand même : aussi de de montrer un espace* ( japon)

Et il y a bien sûr les cas où il est impossible de trancher entre deux homophones :

*38) donc /**c'est, ces**/ deux filles qui font ça elles tournent donc elles recréent une danse( japon)*

L'interprétation d'une séquence se fait également en fonction des connaissances sur le domaine, les savoirs ou les attentes du transcripteur. On transcrit selon ce que l'on croit entendre ou selon ce que l'on croit probable. Ainsi un transcripteur débutant d'un interview radio ne sachant pas que « Les piliers de la terre » « est le titre d'un roman célèbre, transcrit un vocable phonétiquement proche :

*38) mais derrière **l'épillet** de la terre il y a toute votre expérience*

## PERSPECTIVES

### Les progrès souhaitables dans les outils d'exploitation

L'exploitation de corpus informatisés nécessite des outils de requête. Les recherches en syntaxe présentée plus haut s'appuient par exemple sur l'utilisation d'un concordancier qui permet d'interroger la distribution d'un élément sur un grand nombre d'exemples. Mais l'outil a ses limites. Le concordancier travaille en effet sur une chaîne de caractères sans tenir compte de la nature grammaticale des éléments. Ainsi pour une interrogation concernant *bien que* l'outil retiendra les séquences à locution conjonctive et des séquences comme

*40) tout le monde savait très bien que ça portait préjudice*

constituée de l'adverbe *bien* et de *que,* introducteur de complétive. Pour résoudre ce type de problème et sélectionner les séquences pertinentes, on a donc besoin aujourd'hui de constituer des corpus étiquetés, c'est à dire des corpus où chaque mot se voit attribuer un étiquetage morphologique qui indique essentiellement à quelle partie du discours il appartient. Mais comme le signale Valli et Veronis (1999), la technique d'étiquetage morpho-

syntaxique est une technologie relativement bien développée mais « pratiquement inexploré pour l'oral ». On se heurte dans ce cas à des problèmes d'analyse auxquelles les catégories grammaticales traditionnelles n'apportent pas de solution : les auteurs citent notamment, les emplois comme « particules discursives » des éléments comme *quoi, tu sais , tu vois, comment dire*, *bon*, les emplois adverbiaux d'éléments comme *pareil*, les répétitions de mots et la présence de « réalisations grammaticales non normatives ».

En outre si l'étiquetage par partie du discours peut résoudre des problèmes d'ambiguité lexicale, il ne peut suffire lorsque les requêtes doivent être formulées en termes de fonction grammaticale. Pour prolonger par exemple un travail sur les formes sujets à l'oral, thème dont les enjeux en termes typologiques sont importants[24], on aurait besoin de formuler des requêtes en termes de « sujets clitiques, sujets nominaux etc ». L'analyse doit donc découper les chaînes en termes de fonction syntaxiques , opération nommée en TAL, le parsing.

A l'oral, le parsing se heurte à un certain nombre de difficultés liées aux modes mêmes de production. Le premier est celui de la présence d'énoncés parenthétiques, en gras dans l'exemple suivant, qui provoquent notamment des ruptures de construction :

*41) l'encadrement d'une équipe de journalistes ce sont des fonctions où je suis servi en la matière parce que /* ***en tout cas c'est comme ça que je vois les choses*** */ être journaliste c'est faire des choix en permanence [...] puisque on ne peut pas tout mettre on on doit euh* ***comment expliquer ça clairement*** */ euh hiérarchiser l'information* (journaliste)

Pour trouver le verbe qu'introduit *parce que*, il faut sauter par dessus la séquence « *c'est comme ça que je vois les choses* » pour aller chercher « *être journaliste c'est faire des choix* » et reconstruire l'enchaînement : « *l'encadrement d'une équipe de journalistes ce sont des fonctions où je suis servi en la matière parce que être journaliste c'est faire des choix en permanence* ». Si l'on n'isole pas la parenthèse, le parseur analysera mal la séquence en faisant porter le *parce que* sur le verbe qui suit et aboutira au découpage suivant : « *l'encadrement d'une équipe de journalistes ce sont des fonctions où je suis servi en la matière parce que en tout cas c'est comme ça que je vois les choses* » qui ne correspond pas à ce qui a été produit par le locuteur. Il faut donc construire un analyseur qui « ignore » les parenthèses ou qui les traite à un deuxième niveau pour établir les relations syntaxiques, ce qui suppose qu'on ait des critères clairs pour repérer leurs limites. Leur traitement automatique nécessite une analyse fine des indices récurrents qui les accompagnent, notamment les indices prosodiques et certaines marques de balisage.[25] Il reste beaucoup de choses à faire dans ce domaine.

---

[24] cf. Deulofeu,J. (2000) pour une présentation détaillée de cette problématique.
[25] Cf. Delormier, D., et Morel, M., (1986),

Le deuxième phénomène est celui de la présence des bribes, amorces et répétitions[26]. Ces phénomènes ne sont pour l'instant pas intégrés dans les systèmes de reconnaissance automatique et rendent ces derniers inopérants comme le montre l'expérience suivante menée par l'équipe Delic avec un système de reconnaissance automatique dans le domaine du renseignement, dont nous reprenons ici la présentation [27] :

« L'énoncé suivant

*non non non non je veux pas de Pa- de de Paris gare d'Austerlitz*

a par exemple été reconnu comme

*Nancy dans nonante jours à le Havre de Paris gare d'Austerlitz*

Le système n'intègre pas la notion de répétition ou d'amorce de mot inachevé, et fournit donc de façon erronée, une approximation lexicale au mieux de ses possibilités.»

Or les répétitions ou les amorces constituent « les traces du discours en cours d'élaboration » (C. Blanche- Benvensite, 1997) et montrent comment le locuteur construit progressivement les syntagmes au fur et à mesure du déroulement de son discours. Cette progression sur l'axe paradigmatique n'apparaît pas dans la présentation linéaire de la transcription mais peut être mise en valeur par une présentation qui en visualise en quelque sorte les empilements. Comparons la présentation linéaire en (42) qui peut paraître d'une difficulté impossible à traiter par des systèmes automatiques

*42) alors pour les ferrer ben il y a + le principal si tu veux c'est de c'est de mettre bon le coup de mettre le fer c'est juste pour protéger le empêcher l'usure de de la corne mais le le truc c'est de de rester dans les aplombs du cheval tu vois*

et la présentation qui suit où l'on a signalé l'énoncé parenthétique (encadré) et visualisé en liste les bribes.»

**alors pour les ferrer** ben il y a +
le principal si tu veux c'est de
c'est de mettre bon
le coup de mettre le fer c'est juste pour *protéger le*
empêcher l'usure *de*
*de* la corne
*le*
**le truc** **c'est de rester dans les aplombs du cheval tu vois**

On perçoit ainsi la construction progressive de la séquence « alors pour les ferrer le truc c'est de rester dans les aplombs du cheval ». Pour automatiser le repérage de ces structures en listes, il faudrait là encore travailler sur les éléments récurrents qui permettent de les repérer. Deux attitudes sont ensuite

[26] cf Blanche-Benveniste (1990, 1997)

[27] cf. Benzitoun ,C., Campione, E., Deulofeu, J., Henry, S., Teston, S., Valli, A., Veronis, J., (à paraître).

possibles : soit on considère ces amorces et répétitions comme des phénomènes inutiles et on les supprime, soit on considère à la suite des spécialistes de l'oral, que ces éléments sont des observables qui permettent de mieux saisir la constitution progressive du sens. On a dans ce cas tout intérêt à les conserver et à les intégrer à l'analyse. On pourrait pour ce faire s'inspirer de ce qui existe pour le traitement automatique de l'écrit, comme par exemple « les grammaires spécialisées pour le traitement des listes et des énumérations »que propose Nuria Gala[28].

L'analyse de ces phénomènes à l'oral est loin d'être terminée mais elle a permis d'ores et déjà de révéler que les locuteurs traitent différemment le déroulement syntaxique de l'énoncé qui peut faire l'objet de longues mises en mémoire et la recherche de dénomination qui semble se construire au fur et à mesure du discours. Le défi lancé aux parseurs est donc de produire automatiquement des analyses malgré la complexité des phénomènes traités.

Un autre problème et qui n'est pas le moindre est lié au fait que le parsing suppose un découpage des textes en unités, ce qui pose à l'oral de multiples difficultés.[29] Pour l'écrit en effet les programmes de traitement s'appuient sur la ponctuation pour aboutir à une segmentation mais ce principe de découpage n'est pas pertinent à l'oral. D'une part, comme cela a été abondamment montré, en particulier par C. Blanche-Benveniste (op.cit), il n'y a pas d'équivalence entre les marques prosodiques de l'oral et des marques graphiques de ponctuation : on ne saurait dans l'exemple suivant remplacer les pauses par des virgules :

*43) on a réalisé un système + où il y a une pompe ++ électrique ++ qui fonctionne à l'énergie solaire*
(Pompes)

D'autre part, on relève à l'oral de nombreux éléments dont les points de rattachement ne sont pas évidents. Ainsi le segment *sauf s'il pleut ou s'il fait froid* de l'exemple (44), semble être rattaché et au segment qui le précède et au segment qui le suit :

*44) je vais souvent au centre Ville Saint Sebastien tous les trucs là quoi je fais les boutiques j'y vais à pied sauf s'il pleut ou s'il fait froid je reste chez moi (shopping)*

Sabio (1997) signale de nombreux exemples de ce qu'il nomme « les éléments flottants » et qui résistent au découpage : Ainsi pour l'exemple (45) que nous reprenons de sa démonstration,

*45) ça y est il coulait le mois d'après il était au chômage ( baral, 21,8)*

ponctuer reviendrait forcement à rattacher le segment « le mois d'après » à un des noyaux verbaux et donc à préjuger de l'analyse avant que cette dernière soit menée. De façon générale, il est admis chez les spécialistes de

---

[28] CF. GALA, N., (à paraître).

[29] cf. Blache 2000 pour une présentation des problèmes liés aux choix théoriques de la démarche et leurs implications sur l'analyse dans le domaine de l'écrit.

l'oral[30] que la notion de phrase ne permet pas de définir des unités de segmentation cohérente, comme on peut l'observer au travers de l'exemple qui suit, repris de 6.

*46) moi je suis relativement optimiste hein / parce que je pense que actuellement nous avons une génération qui est la pire / mais ils vont avoir des enfants à leur tour et la troisième génération ils seront français (MACE. 12, 3)*

Si l'on applique à cet exemple un découpage en phrase canonique, le respect des normes de ponctuation obligeant à mettre un point juste avant le *mais*, on aboutit à énoncé pour le moins contradictoire du type : « moi je suis relativement optimiste hein / parce que je pense que actuellement nous avons une génération qui est la pire ». C'est ce qui se passe dans l'exemple suivant, extrait de la presse et pour lequel l'auteur a voulu transcrire les propos d'un interview tout en se tenant au plus près de la norme écrite :

*47) Il m'a engagée d'abord comme secrétaire pour taper un scénario pus comme script girl pour Fanny. J'ai beaucoup hésité parce qu'il me proposait 300 francs par semaine ce qui était beaucoup par rapport à ce que je gagnais. Mais c'était un métier précaire et j'ai décidé de prendre le risque. (Presse)*

La transcription isole en effet une séquence *« J'ai beaucoup hésité parce qu'il me proposait 300 francs par semaine ce qui était beaucoup par rapport à ce que je gagnais »* qui peut induire le lecteur en erreur. On voit au travers de ces quelques exemples à quel point la segmentation des textes à l'oral n'est pas aisée. Elle nécessite une recherche quant aux critères à adopter pour définir l'unité de l'analyse syntaxique. La réflexion est cours au sein de l'équipe DELIC sur cette « unité maximale » qui est à l'heure actuelle définie « à partir des constructions verbales, nominales, adjectivales ou adverbiales, regroupant un élément tête ainsi que tous les éléments qui sont sous sa dépendance ». Nous ne mentionnerons que deux des problèmes auxquels les chercheurs sont confrontés.

Le premier, déjà relevé, concerne la non congruence entre marques morphologiques et relations syntaxiques. D'une part les morphèmes dits de subordination n'introduisent pas toujours de vraies subordonnées, mais peuvent construire des structures dont on voudrait faire des unités maximales. Il en est ainsi dans les exemples de *parce que* mentionnés plus haut, ou dans l'exemple suivant :

*48) le figmag est un pavé qui pèse presque aussi lourd que les journaux anglais le dimanche parce que tu sais les journaux anglais pèsent très lourd le dimanche (voyage)*

L'exemple peut être analysé en deux unités : «le figmag est un pavé qui pèse presque aussi lourd que les journaux anglais le dimanche » et « parce que tu sais les journaux anglais pèsent très lourd le dimanche ».

---

[30] cf.CBB et Berrendonner 90.

Il n'y a pas de lien de dépendance syntaxique entre les deux unités, contrairement à ce que pourrait laisser penser la présence du morphème *parce que.* Celui ci fonctionne comme connecteur de discours et non pas comme un « marqueur de rection » et introduit une assertion à valeur de commentaire de ce qui précède. Il en est de même des exemples de *bien que* avec l'indicatif mentionnés plus haut.

D'autre part on a montré (Cf. J. Deulofeu, 1989) que des énoncés non marqués morphologiquement peuvent être regroupés pour former une seule unités maximale : Ainsi dans l'exemple :

*49) des fois ils mangaient c'était minuit*

les deux constructions ne peuvent être analysées comme des constructions indépendantes mais constituent une seule unité maximale qui repose sur une relation de dépendance syntaxique relevant du domaine de ce que l'on nomme la macrosyntaxe[31]. La difficulté est donc double : il s'agit de faire reconnaître à un analyseur dans quel cas intégrer dans l'unité maximale les structures introduites par des morphèmes et dans quel cas les exclure et par ailleurs de repérer, en l'absence de marques mophologiques des structures organisées en unités maximales.

Le deuxième problème vient du fait que l'on ne peut pas toujours s'appuyer sur les critères prosodiques pour délimiter les unités Ainsi si nous reprenons le début de l'exemple 6

*50) moi je suis relativement optimiste hein /parce que je pense que actuellement nous avons une génération qui est la pire/ / mais ils vont avoir des enfants à leur tour et la troisième génération ils seront français (MACE. 12, 3)*

La prosodie nous amènerait à découper la séquence en deux unités : [moi je suis relativement optimiste parce que .. une génération qui est la pire] ° [mais ils vont avoir des enfants à leur tour….]. On note en effet une rupture importante avec une intonation de fin d'énoncé avant mais. Or le fait d'isoler cette dernière séquence de ce qui précède aboutit, on l'a vu à un énoncé incohérent du point de vue du sens, On doit donc poser que dans cet exemple, l'unité maximale est constituée par l'ensemble de la séquence. Il n'y a pas congruence entre prosodie et structure syntaxique.

**Conclusion**

Le recours à de grands corpus oraux a d'ores et déjà permit d'enrichir la description syntaxique du français dans des domaines qui ne sont pas restreints à l'oral : ainsi l'observation de grandes distributions amène à interroger le rapport entre marques morphologiques et relations syntaxiques et met également en évidence l'importance des contraintes lexicales sur certaines tournures grammaticales. De façon plus générale les

[31] Cf. Blanche-benveniste (1990,2002),Berrendonner (1991,2002) et Deulofeu (2003) pour une présentation des analyses macrosyntaxiques du français.

recherches sur corpus revèlent comment l'analyse des distributions des formes linguistiques selon les genres permet de mieux les analyser. Le travail doit être poursuivi en termes quantitatifs et qualitatifs pour aboutir à la constitution de grands corpus échantillonnés et faire avancer l'analyse syntaxique. Le traitement automatique de ces corpus se révèle indispensable mais nécessite, compte tenu de la nature même de l'oral, une étroite collaboration entre linguistes et informaticiens. Il est regrettable que dans ce domaine, aucune programmation institutionnelle cohérente ne soit mise en place alors même que les analyses menées montrent l'utilité d'un tel investissement pour une meilleure description de la langue.

**Eléments bibliographiques**

ADAM, J. M., (1990) : *Eléments de linguistique textuelle,* Liège, Mardaga, coll. Philosophie et langage.

ADAM, J. M. (1992) : *Les TEXTES : Types et prototypes*, Paris, Nathan Université.

ANDRE-LAROCHEBOUVY D., (1984) : *La conversation quotidienne,* Credif, coll. "Essais", Paris.

ASHBY, W. (1982) : « The drift of french syntax », Lingua n°57.

BENZITOUN,C., CAMPIONE, E., DEULOFEU, J., HENRY, S., TESTON, S., VALLI, A., VERONIS, J., (à paraître) : «L'analyse syntaxique de l'oral : Problèmes et méthode », in *Evans* : *Méthodes et outils pour l'évaluation des analyseurs syntaxiques* Journées d'Etudes de l'ATALA- Paris, mai 2004.

BERRENDONNER, A., LE GUERN, M., PUECH, G., (1983) : "Principes de grammaire polylectale",Lyon, P.U.L.

BERRENDONNER, A., (1991) : « Pour une macro-syntaxe », in Dominique Willems (éd.), *Données orales et théories linguistiques,* Paris -Louvain Duculot, pp.25-31.

BERRENDONNER, A., (2002) : « Les deux syntaxes », in M. Charolles, Le Goffic et M.A. Morel (coord), *Y a-t-il une syntaxe au-delà de la phrase ?* Verbum, Tome XXIV, P.U.N, Nancy,pp. 23-36.

BIBER, D., (1988) : *Variety across speech and writing*. Cambridge University Press.

BIBER, D., (1995) : *Dimensions of register variation. A cross-linguistic comparison*. Cambridge. Cambrige University Press.

BILGER, M. (éd.), (1999) : *L'Oral spontané. Revue Française de Linguistique Appliquée*, vol. IV-2.

BILGER, M. (éd.), (2000a) : *Corpus. Méthodologie et Applications linguistiques*, Champion, Paris.

BILGER,M. (coord), (2000b) : *Linguistique sur corpus. Etudes et réflexions,* Cahiers de l'université de Perpignan., n°31. Presses Universitaires de Perpignan.

BILGER, M., (2000c) : « Petite typologie des conventions de transcription de l'oral »., in BILGER, M. (coord), (2000b) : *Linguistique sur corpus. Etudes*

*et réflexions,* Cahiers de l'université dePerpignan., n°31. Presses Universitaires de Perpignan, pp. 77-93.

BILGER, M. & BLANCHE-BENVENISTE, C., (1999) : « Français parlé-oral spontané. Quelques réflexions », in *L'oral spontané*, *Revue Française de Linguistique Appliquée*, Vol. IV-2, pp. 21-31.

BILGER, M., BLASCO, M., CAPPEAU, P., SABIO, F., &SAVELLI, M.J., (1997) : « Transcription de l'oral et interprétation : illustration de quelques difficultés », in *Recherches sur le français parlé*, n° 14, Publications de l'Université de Provence, pp. 57-86.

BLACHE, P., (2000) : « A quoi sert l'annotation syntaxique de corpus ? », in BILGER, M. (éd.) (2000a) : *Corpus. Méthodologie et Applications linguistiques*, Champion, Paris, pp. 82-93.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C. & VAN DEN EYNDE, K., (1990) : *Le français parlé, études grammaticales, éd. du CNRS*, coll. Sciences du langage, Paris.

BLANCHE-BENVENISTE, C., (1994) : « Quelques caractéristiques grammaticales des sujets employés dans le français parlé des conversations » in M. Yaguello (éd.), *Subjecthood and subjectivity. The status of the Subject in linguistic theory*. Paris: Ophrys, pp. 77-108.

BLANCHE-BENVENISTE, C., (1997) : « De l'utilité du corpus linguistique », in *Corpus. De leur constitution à leur exploitation, Revue Française de Linguistique Appliquée*, vol. I – fasc.2, pp. 25-42.

BLANCHE-BENVENISTE, C., (1999) : *Approches de la langue parlée en français.* Paris, Ophrys.

BLANCHE-BENVENISTE, C., (2000a) : « Analyse de deux types de passif dans les productions de français parlé », in L. SCHOESLER (éd.), *Le Passif,* Copenhague : Musuem Tusculanum Press

BLANCHE-BENVENISTE, C., (2000b) : « Corpus de français parlé », in BILGER, Mireille (éd.), (2000a) *Corpus. Méthodologie et Applications linguistiques*, Champion, Paris, pp.15-25.

BLANCHE-BENVENISTE, C., (2000c) : « Convergences de matériel grammatical permettant d'établir des typologies textuelles » in BILGER, M., (coord), (2000b) : *Linguistique sur corpus. Etudes et réflexions,* Cahiers de l'université de Perpignan., n°31. Presses Universitaires de Perpignan, pp. 103-116.

BLANCHE-BENVENISTE, C., (2002) : « Phrase et construction verbale », in M. Charolles, Le Goffic et M.A. Morel (coord) : *Y a-t-il une syntaxe au-delà de la phrase ?* Verbum, Tome XXIV, P.U.N, Nancy.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C. et VAN DEN EYNDE, K, (1990) : *Le Français parlé : études grammaticales.* Paris, éditions du CNRS.

BLOOMFIELD, L. (1927) : « Literate and illiterate speech », *American Speech* 2-10, 432-439; réédité dans C. F. HOCKET, A (1970) : *Leonard Bloomfield Anthology. Bloomington*: Indiana University Press.

CAPPEAU, P.,

COMBETTES B., (1999) : « Thématisation, topicalisation : leur rôle respectif dans l'évolution du français », in *La thématisation dans les langues*, Textes réunis par C. Guimier,P. Lang. pp. 133-159.

DEBAISIEUX, J. M., (2001) : « Contraintes syntaxiques et discursives des emplois de *quant à* et *en ce qui concerne* dans les corpus oraux », in *Cahiers de Praxématique*, 37, Praxiling, Université Paul Valéry, Montpellier,pp. 125-146.

DEBAISIEUX, J. M., (2002) : « Le fonctionnement de *parce que* en français contemporain : étude quantitative », in PUSCH C.D. & RAIBLE W. (Hrsg.) *Romanistische Korpuslinguistik – Romance Corpus linguistics*, Gunter Narr Verlag, Tübingen, pp. 349-362.

DEBAISIEUX, J.M. & DEULOFEU, J.(à paraître) : « Fonctionnement microsyntaxique de modifieur et fonctionnement macrosyntaxique en parataxe des constructions introduites par *que* et *parce que* en français parlé, avec extension au cas de *perché* et *che* en italien parlé », Actes du colloque *Il Parlato Italiano,* Napoli, 13-15 février 2003.

DEBAISIEUX, J.M. & DEULOFEU, J.(à paraître) : « Etude comparative des fonctionnements non subordonnés des morphèmes *que* et *parce que* en français et *que* et *porque* en castillan », *III International Contrastive Linguistics Conference*, Santiago de Compostela, 23-26 Septembre 2003.

DELORMIER, D., et MOREL, M., (1986) : « Caractéristiques intonatives et syntaxiques des incises », in *DRLAV*, 34-35, pp. 141-60.

DEULOFEU, J., (1989) : « Les couplages de constructions verbales en français parlé : effet de cohésion ou syntaxe de l'énoncé », in *Recherches sur le Français Parlé*, n°9, pp. 111-141.

DEULOFEU, J., (2000) : L'innovation en syntaxe en français contemporain", in ROUSSEAU, J., DEMARTY J., (éd) : *Français de l'avenir et avenir du français*, C.I.E.P., Didier, Paris, pp. 43,57.

DEULOFEU, J., (2003) : «L'approche macrosyntaxique en syntaxe : un nouveau modèle de rasoir d'occam contre les notions inutiles ? » in *SCOLIA* ; n°16, Publications de l'université Marc Bloch, Strasbourg, pp. 77-95.

FONSECA-GREBER, B., & WAUGH, L. R. ,(2003) : « On the Radical Difference between the Subject Personal Pronouns in Written and Spoken European French», in LEISTYNA,P., & MEYER,C.F. (éd), (2003) *Corpus Analysis. Language structure and language Use,* Language and Computers, Studies in Practical Linguistics, n°46, Amsterdam-New-York, Rodopi B.V., pp. 225-241.

GADET, F., (1999) : « La variation diaphasique en syntaxe », in Barbéris, J.M., (éd.) : « Le français parlé. Variétés et discours ». *Praxilingue,* (Université Montpellier-3),pp. 211-228.

GADET, F. (2000a) : « Des corpus pour (ne)… pas », in BILGER, M. (éd.), (2000a) : *Corpus. Méthodologie et Applications linguistiques*, Champion, Paris, pp. 156-168.

GADET, F. (2000b), "Derrière les problèmes méthodologiques du recueil des données", in BILGER, M. (coord), (2000b) : *Linguistique sur corpus. Etudes et réflexions,* Cahiers de l'université de Perpignan., n°31. Presses Universitaires de Perpignan, pp. 59-77.

GALA, N., (à paraître) : « Des indices sur la fiabilité des sorties ou comment un analyseur robuste pourrait s'auto-évaluer », in *Evans : Méthodes et outils pour l'évaluation des analyseurs syntaxiques*, Journées d'Etudes de l'ATALA- Paris, mai 2004.

HABERT, B., NAZARENKO, A., SALEM, A., (1997) : *Les linguistiques de corpus*, Paris, Colin.

HABERT, B., (2000) : « Des corpus représentatifs : de quoi, pour quoi, comment ? » in BILGER, M. (coord), (2000b) : *Linguistique sur corpus. Etudes et réflexions,* Cahiers de l'université de Perpignan., n°31. Presses Universitaires de Perpignan, pp. 11-58.

LE GOFFIC, P. (1993) : *Grammaire de la phrase française*, Hachette Supérieur, Paris.

MILLER, J., WEINERT, R., (1998) : *Spontaneous Spoken Language. Syntax ans Discourse*, Clarendon Press, Oxford.

MOREL, M.A. (1996) : *La concession en français,* Ophrys, Coll. L'essentiel, Gap, Paris.

MULLER

*Recherches sur le français parlé*, Groupe Aixois de Recherches en Syntaxe, Publications de l'Université de Provence.

PREVOST S., ( 2003) : « Quant à : Analyse pragmatique de l'évolution diachronique (14ème – 16ème siècle », in Combettes, B., Schnedecker, C., et Theissen, A., (éds), *Ordre et distinction dans la langue et le discours*, Paris, Champion, pp. 39-52.

RIEGEL, M., PELLAT, J. C., RIOUL, R., (1994) *Grammaire méthodique du français*, P.U.F, Paris.

ROUGET, C., (2000) : « Les nominalisations sont-elles réservées aux descriptions techniques ? », in BILGER, M. (éd.), (2000a) : *Corpus. Méthodologie et Applications linguistiques*, Champion, Paris, pp.296-306.

SABIO, F., (1995) : « Micro-syntaxe et macro-syntaxe : l'exemple des compléments antéposés », in *Recherches Sur le Français Parlé* n° 13, Publications de l'Université de Provence, Aix en Provence, pp. 111-156.

SCHEUTZ, H., (1998) : « weil-Sätze im gesprochenen Deutsch », in Hutterer, C.J., Pavritsch, G.,
(eds) : *Beitrage zur Dialektologie des ostoberdeutschen Sprachraumes*, Göppingen : Kümmerle Verlag.

VALLI, A. & VERONIS, J., (1999) : « Etiquetage grammatical des corpus de parole : problèmes et perspectives », », *Revue Française de Linguistique Appliquée*, vol IV-2, *L'Oral spontané*, pp. 113-134.

VERONIS, J., (2000) Annotation automatique de corpus : panorama et état de la technique. In J.M. PIERREL (éd.), *Ingénierie des langues*, pp. 111-129.

WILLEMS, D., (1998) : « Données et théories en linguistique. Réflexion sur une relation tumultueuse et changeante », in M. BILGER, F.GADET, et K. van den EYNDE (éds.), *Analyse linguistique et approches de l'oral. Recueil d'étude offerts en hommage à Claire Blanche-Benveniste*. Louvain/Paris : Peeters, pp. 79-87.

## Annexes : les corpus oraux informatisés disponibles pour le français

### En France

**Corpus du Gars** (Delic) Plus d'un million de mots transcrits sur support informatique. Sources sonores consultables sur place – Extraction de requêtes grâce au logiciel d'interrogation CorpAIX ( J.M. Adam).

**Corpus de référence du français parlé** : Equipe Delic
440 000 mots. Plus de 36 heures de parole . Corpus échantillonné en fonction de plusieurs situations de paroles et de niveau d'étude. Corpus aligné ( texte et son) interrogeable par le concordancier Contexte ( J. Veronis) - http://www.up.univ-mrs.fr/veronis/DELIC/

**Corpus du Groupe de Recherches sur les Interactions Communicatives**, GRIC, dirigé par Christian Plantin, avec Catherine Kerbrat-Orecchioni, Université de Lyon II (Analyse de conversation). (600heures d'enregistrements)

**Corpus de Pierre Bange**, Lyon II, Centre de Recherches Linguistiques et Sémiologiques, 40 heures d'enregistrements avec transcription orthographique, "Interactions de consultation" et « Interactions de conciliation".

**Corpus du groupe de Mary-Annick Morel, Université de Paris III** (Prosodie), cassettes et transcriptions disponibles sur place.

**Corpus de Durand et Laks,** Université de Toulouse le Mirail (phonologie).Projet international sous la direction de Jacques Durand (ERSS-UMR5610, Université de Toulouse-Le Mirail), Bernard Laks (Université de Paris X) et Chantal Lyche (Université d'Oslo).

**Corpus de Françoise Gadet**, Université de Nanterre (Sociolinguistique), français « populaire».

**Corpus de Barbéris, Boyer**, Université Paul-Valéry de Montpellier (Analyse de discours).

**Corpus de S. Mellet**, Université de Nice (Variations régionales).

**En Belgique**

**VALIBEL** (Variété Linguistiques du français de Belgique), 360 heures de textes parlés transcrits, 22 corpus, 8300 pages de texte transcrits en orthographe française, bandes accessibles pour consultation, sur demande. Contact: francard@frwa.ucl.ac.be

**ELICOP** (Etude Linguistique de la Communication Parlée): LANCOM (transcriptions de 500 heures de français parlé, y compris *Le Corpus d'Orléans* (1968-1971, 315 heures), *Le Livre parlé de Tours* (1974, 120 heures), *La Voix d'Auvergne* (1976, 52 heures) + ELILAP (1993, 26 heures,

transcriptions de 8 heures de jeux de rôles avec Francophones et Néerlandophones apprenant le français - Site web : http://bach.arts.kuleuven.ac.be/lancom/

Résumé