

Une distinction mesurable: corpus oraux et écrits sur le continuum de la deixis*

JEAN-MARC DEWAELE

Birkbeck College, University of London

ABSTRACT

L'opposition oral/écrit est analysée à travers la variation dans les proportions des catégories grammaticales dans vingt-et-un corpus oraux et écrits italiens, néerlandais, français et d'interlangue française. Une analyse factorielle permet d'extraire une dimension unique sur laquelle s'opposent deux groupes de catégories grammaticales. Les substantifs, les articles, les adjectifs et les prépositions se situent près du pôle négatif tandis que les pronoms, les verbes, les adverbes, les interjections et les conjonctions se situent près du pôle positif de cette dimension qui reflète le continuum de la deixis. Les corpus oraux se situent près du pôle déictique/implicite et s'opposent aux corpus écrits qui se rapprochent du pôle explicite du continuum. Les énoncés oraux sont davantage ancrés dans le contexte spatio-temporel du locuteur tandis que le discours écrit est généralement plus indépendant du contexte spatio-temporel, ce qui explique son caractère plus nominal et sa proportion plus importante de mots essentiellement non-déictiques ou explicites.

I INTRODUCTION

L'oral et l'écrit sont 'à la fois des abstractions et des catégories difficiles à isoler (puisque les faits de société nous obligent à les considérer comme en continuum); mais ce sont des ordres qu'il serait fâcheux de ne pas distinguer dans la description' constate Gadet dans son étude intitulée très justement *Une distinction bien fragile: oral / écrit* (1996b: 26).

On pourrait donc se demander, premièrement, s'il est possible de distinguer l'oral et l'écrit à l'aide de critères linguistiques rigoureux, et, deuxièmement, s'il existe des mesures objectives permettant de placer quelconque extrait de discours sur le continuum oral/écrit. Une mesure reflétant la distinction essentielle entre l'oral et l'écrit permettrait de dépasser les classements relativement subjectifs des registres (1) utilisés à l'oral et à l'écrit (voir par exemple

* Une version antérieure de la présente étude a été présentée sous le même titre au Colloque Annuel de l'Association for French Studies: Français oral, français écrit à l'ère des nouvelles technologies, University of East Anglia, septembre 1998. Nous tenons à remercier Ian Short et les lecteurs anonymes de JFLS pour leurs excellentes suggestions.

Jean-Marc Dewaele

Peytard, 1970; 1977). Le placement d'un extrait de discours sur le continuum oral/écrit pourrait se faire indépendamment de la situation de communication dans laquelle il a été produit et ne dépendrait plus de critères relativement superficiels comme la présence de mots familiers ou l'omission du 'ne' dans la négation. Il permettrait finalement d'analyser la variation à l'intérieur de l'oral et de l'écrit et de découvrir les chevauchements éventuels (Marty, 1991) entre les deux modalités: l'écrit dans l'oral et l'oral dans l'écrit. L'abondante littérature sur ce sujet (voir Gadet, 1996c) pourrait suggérer que la difficulté de définir et d'isoler ces catégories a été affrontée et résolue.

Notre analyse approfondie de quelques travaux importants en la matière révèle cependant des lacunes, des problèmes d'interprétation liés à des choix méthodologiques et finalement l'absence d'une mesure valide qui reflète l'essence de l'oral et de l'écrit. Gadet (1996a) regrette l'absence d'une définition empirique des registres. Ceux-ci sont définis par les situations dans lesquelles ils ont été produits. Elle constate en outre que les termes censés représenter un même niveau ne sont jamais synonymes à cause de la variété de gammes sur lesquelles ils jouent.

Nous tenterons dans le présent travail d'éviter les nombreux écueils méthodologiques et théoriques, et nous proposerons une mesure que nous estimons théoriquement et empiriquement plus solide, et capable de distinguer des extraits de discours oraux et écrits appartenant à des langues différentes, indépendamment de la situation dans laquelle ils ont été produits.

2 APPROCHES ET MÉTHODOLOGIES

Les recherches sur l'opposition oral/écrit se classent en deux grandes catégories: d'une part l'approche non-quantitative où l'on essaye de saisir l'essence des deux types de discours en dépassant les phénomènes de surface (Gumperz; Kaltman et O'Connor, 1984); d'autre part l'approche quantitative, où l'on tente de répertorier les éléments caractéristiques des deux modalités en se basant sur leur fréquence dans le discours. Les études quantitatives ont souvent abouti à des résultats contradictoires, ce qui n'est guère étonnant lorsque l'on considère les méthodologies parfois fort divergentes qui ont été utilisées. Beaman (1984), par exemple, a analysé les résultats contradictoires d'études quantitatives portant sur le discours écrit et parlé et a conclu que trop de chercheurs négligent les facteurs tels que les différences de registre, de formalité et de temps d'exécution. Ajoutons à cela l'absence d'analyse de la variation individuelle, la taille réduite des corpus – surtout des corpus oraux – et le flou dans la définition des variables linguistiques considérées. Les recherches quantitatives sur les modalités écrit/oral se subdivisent à leur tour en plusieurs courants. Certains se concentrent sur la syntaxe ou sur le lexique, d'autres considèrent des éléments syntaxiques avec des éléments lexicaux ou purement grammaticaux. Au sein de chaque domaine on fait à nouveau une distinction entre les études où l'aspect quantitatif n'est que secondaire et celles

Une distinction mesurable

où l'aspect quantitatif revêt une importance telle que les conclusions ne sont que le résumé de la description statistique des éléments traités.

Les recherches quantitatives sur les modalités écrit/oral que nous proposons d'analyser apportent toutes un élément ou un aspect qui nous semble vital pour la définition de l'oral et de l'écrit. Il s'agit essentiellement de travaux sur l'anglais mais les principes généraux qu'ils dégagent nous paraissent également applicables à d'autres langues.

Le débat sur les différences pragmatiques entre discours oral et écrit s'est surtout concentré sur le degré d'implication du locuteur/écrivain. Bresson (1976) et Olson (1977), par exemple, arguent que l'oral est implicite et hautement contextualisé alors que l'écrit est explicite et décontextualisé. Selon eux, l'information transmise à l'oral dépend davantage du contexte non-linguistique et du gestuel, le discours oral est plus elliptique et déictique que le discours écrit. D'autres, comme Tannen (1982), contestent cette interprétation et soutiennent au contraire que ce n'est pas la modalité qui détermine la dépendance du contexte d'un discours, mais plutôt le genre, le registre et les intentions communicatives des participants. Grâce à ce débat les chercheurs ont pris conscience du fait que la modalité ne peut pas être analysée sans tenir compte d'autres variables indépendantes.

Redeker (1984) par exemple, limite l'influence des variables qui co-varient avec la modalité (écrit/oral), notamment le sujet, le temps de préparation, la formalité de la situation et l'audience, et ne manipule que les deux premières variables (modalité et sujet). Elle utilise les dimensions 'implication versus détachement' et 'fragmentation versus intégration' de Chafe (1982), avec la collaboration de ce dernier, pour distinguer le discours oral du discours écrit chez huit étudiantes anglophones. Elle leur a demandé de lui raconter une brève histoire personnelle et de lui fournir une explication sur le système d'éducation américain. Elle leur a ensuite demandé de lui écrire une lettre. L'auteur spécifia qu'elles devaient écrire ce qui leur passait par la tête sans trop se soucier de la forme. Trois semaines plus tard, les participantes furent priées d'écrire un texte portant sur les mêmes sujets que ceux traités lors des interviews.

Les pôles des dimensions de Redeker sont caractérisés par certains indices de fréquence sur 1000 mots. Le pôle *implication*, sur la première dimension, est caractérisé, entre autres, par un grand nombre d'auto-références; de présents historiques; d'expressions populaires; de mots marquant l'emphase; de signes d'autosurveillance, d'expressions vagues. L'autre pôle de la dimension illustre le *détachement* et est caractérisé par un grand nombre de passifs, de formes verbales au plus-que-parfait, le discours indirect et les expressions littéraires. Le pôle *fragmentation*, sur la deuxième dimension, est caractérisé par un grand nombre de propositions indépendantes et d'autres propositions coordonnées tandis que le pôle *d'intégration* est caractérisé par la fréquence d'adjectifs attributifs, de participes présents, de nominalisations, de verbes, d'adjectifs et de noms en cascade et de propositions subordonnées.

Jean-Marc Dewaele

L'auteur applique ensuite une analyse de variance pour déterminer l'effet de la modalité et des genres.¹ Redeker constate que le discours oral est plus impliqué que l'écrit ($F(1,7) = 13,8$; $p < 0,01$) et moins détaché ($F(1,7) = 6,7$; $p < 0,01$). L'oral s'avère également plus fragmenté ($F(1,7) = 88,0$; $p < 0,001$) et moins intégré ($F(1,7) = 61,3$; $p < 0,001$) que l'écrit. Ces dimensions semblent donc fort bien différencier l'oral de l'écrit. Il y a cependant un problème psychométrique dont l'auteur est conscient: les indices qui constituent les dimensions ne sont pas de nature identique, ce qui risque de provoquer des distorsions lors de l'analyse statistique. Certains indices sont en outre beaucoup plus fréquents que d'autres; or, l'auteur constitue son échelle en se basant sur les valeurs absolues et non sur les valeurs relatives, d'où un risque supplémentaire d'interprétation erronée. Redeker trouve une corrélation significative négative entre les pôles *intégration* et *fragmentation* de sa deuxième dimension ($r = -0,89$), ce qui suggère qu'il s'agit véritablement d'une même dimension; la corrélation cependant entre les pôles *implication* et *détachement* de la première dimension est non-significative ($r = -0,05$) ce qui suggère un manque de consistance interne. Redeker admet qu'il y a un problème: 'the unidimensionality of the scales seems questionable' (1984: 52-53).

Les oppositions que Chafé et Redeker découvrent entre discours oraux et écrits sont intéressantes mais le choix d'indices extrêmement hétérogènes limite la portée des résultats.

Toujours dans le cadre du débat sur la question de savoir si l'écrit est plus explicite que l'oral, Mazzie (1987) confirme partiellement une hypothèse de Tannen (1982), notamment que le contenu influence plus la forme du discours que la modalité. Le mérite de l'étude de Mazzie (1987) est de s'être basée sur un corpus relativement important d'échantillons oraux et écrits de treize sujets anglophones qui ont été testés dans six situations différentes. Dans une première tâche orale les sujets étaient priés de décrire au chercheur ce qu'ils voyaient sur un nombre de dessins présentant des opérations arithmétiques. La deuxième tâche orale consistait à expliquer les mêmes dessins au chercheur. Ces deux tâches sont dénommées 'narration' et 'résumé' devant une véritable audience (le chercheur). Le chercheur répète alors l'expérience en priant les sujets de faire semblant de se trouver face à une audience imaginaire très nombreuse. Les tâches 'narration' et 'résumé' devant des audiences 'réelles' et 'imaginaires' sont alors répétées pour l'écrit. Après la transcription des données, l'auteur a entrepris la répartition de tous les syntagmes nominaux (SN) référentiels en cinq catégories. Une première catégorie ('evoked') regroupe les SN qui font allusion à quelque chose qui a déjà été mentionné dans le texte ou qui est présent dans le contexte. Les catégories deux et trois ('plain inferrable' et 'set inferrable') contiennent les

¹ Les termes 'registre' et 'style' sont utilisés comme des synonymes et se réfèrent à des variétés de langage déterminées par la situation (Biber, 1994: 51).

Une distinction mesurable

NP dont la relation avec un élément mentionné plus tôt peut être inférée. La quatrième catégorie regroupe les SN introduisant des entités nouvelles dans le discours ('brand new'). Dans la cinquième catégorie Mazzie rassemble les SN qui se réfèrent à une entité qui n'a pas encore été mentionnée dans le discours, et qui est connue par l'interlocuteur. Mazzie remarque qu'un texte présentant une proportion élevée de SN inférables doit être considéré comme moins explicite. Elle constate cependant que les textes oraux ne contiennent globalement pas une proportion plus élevée d'information inférée que les textes écrits. L'auteur constate que la proportion des catégories de SN inférables varie plutôt en fonction du contenu.

On est en droit de se demander si Mazzie a choisi les meilleures variables dépendantes pour mesurer l'explicité relative d'un texte. La distinction entre les catégories des SN inférés, évoqués et connus, ne semble pas très claire: l'information contenue dans les SN de ces trois catégories dépend du contexte spatio-temporel. Seuls les SN de la catégorie quatre (SN nouveaux) sont relativement plus indépendants du contexte et nous semblent donc plus appropriés comme indicateurs de l'explicité. Une rapide analyse des données de Mazzie révèle une variation systématique de cette dernière catégorie suivant la modalité. Les textes écrits produits dans la tâche de la narration contiennent en moyenne 27% de SN 'nouveaux' versus 22% pour les extraits oraux dans la même tâche. La même variation apparaît dans la tâche du résumé où la proportion de SN 'nouveaux' est en moyenne 26,5% pour l'écrit contre 21,5% pour l'oral (Mazzie, 1987: 41). Le discours oral serait dès lors nettement moins explicite que le discours écrit. Une autre observation de Mazzie permet de renforcer l'hypothèse du caractère plus explicite du discours écrit: elle constate que la fréquence relative des substantifs était nettement plus élevée dans les extraits écrits (59,5% dans la tâche de la narration, 92,5% dans la tâche du résumé) que dans les extraits oraux (55,5% dans la tâche de la narration, 79,5% dans la tâche du résumé). L'auteur découvre une variation inverse entre les modalités au niveau de la proportion de pronoms. Les extraits oraux dans les deux tâches contiennent beaucoup plus de pronoms (36% et 19% respectivement) que les extraits écrits (33,5% et 5,5% respectivement). Il semble par conséquent qu'en plus de l'effet du contenu sur l'explicité dans le matériel recueilli par Mazzie, l'effet de la modalité sur l'explicité, quelque peu négligé par l'auteur, soit tout aussi réel.

Les travaux impressionnants de Biber sur l'anglais (1988, 1995), de Biber *et al.* (1994, 1998) et de Biber et Finegan (1991, 1994) sur la variation de 17 registres dans 481 textes tirés du corpus Lancaster-Oslo/Bergen (960.000 mots) et 6 registres oraux tirés de 148 extraits oraux du corpus London-Lund anglais, s'inscrivent dans la lignée des travaux de Chafe (1982), Tannen (1982) et Redeker (1984). Comme Redeker, Biber (1988) analyse des phénomènes de co-occurrence entre 67 variables linguistiques, dont l'occurrence est calculée sur 1000 mots, et analyse ensuite la variation entre différents registres oraux et écrits. Biber et Finegan (1991) admettent que le choix du niveau de regroupe-

Jean-Marc Dewaele

ment des variables représentait un véritable dilemme. Ils expliquent qu'ils ont choisi le niveau le plus bas possible (pronoms personnels 1^e personne, 2^e personne . . .), utilisant le critère de la différenciation fonctionnelle, parce qu'ils pouvaient ainsi combiner certaines variables par la suite si cela s'avérait nécessaire. Biber (1995: 135) ajoute que les analyses factorielles peuvent de toute façon distinguer des regroupements fonctionnels parmi les variables de base. Celles-ci sont de nature syntaxique (constructions passives, types de phrases coordonnées, subordonnées et relatives), lexicale (présence de certains mots ou expressions, indice de longueur de mots, de richesse lexicale) et morphologique (présence d'articles, de prépositions . . .).

Les analyses factorielles permettent en effet à Biber de dégager six facteurs en anglais, dont nous retenons les trois principales. Biber (1998: 149) admet que l'interprétation du pôle positif de la première dimension, susceptible d'expliquer la proportion la plus importante de variance, n'a pas été aisée. On y retrouve les formes verbales et pronominales de la première et deuxième personne, les contractions, les usages emphatiques, les différents types d'adverbes et d'interjections. Ces variables représentent un nombre important de fonctions qui reflètent l'interaction directe, les attitudes et sentiments personnels, la fragmentation et un contenu général. Biber reprend le terme ('involvement') de Chafe (1982) et Tannen (1982) pour caractériser ce pôle positif (Biber, 1995: 143). Le pôle négatif regroupe les substantifs, prépositions, adjectifs attributifs, la longueur de mots et la richesse lexicale. Ce pôle caractériserait une haute densité d'information et est nommé *informationnel* ('informational'). Biber observe que sur cette première dimension les registres oraux se situent près du pôle *implication* alors que les registres écrits se rapprochent du pôle *informationnel*.

Les verbes aux passé, les pronoms de la troisième personne et l'aspect perfectif obtiennent des coefficients de saturation élevés sur la deuxième dimension (*orientation narrative* versus *orientation non-narrative*) tandis que les clauses relatives s'opposent aux adverbes de temps et de lieu sur la troisième dimension baptisée *référence explicite* ('elaborated') versus *référence ancrée dans la situation* ('situation-dependent').

Biber (1988) conclut qu'il n'existe pas de dimension unique permettant d'expliquer toute la variance entre les registres. Il constate aussi qu'aucune dimension ne permet de saisir toutes les différences entre l'oral et l'écrit mais que, malgré tout, les conversations spontanées apparaissent comme l'oral stéréotypique alors que la prose académique représente l'écrit stéréotypique. Un doute apparaît quant à la validité des dimensions quand on constate que sur la première dimension, le corpus de conversations téléphoniques est plus proche du pôle *implication* que le corpus de conversations en face à face (Biber, 1995: 146). La même différence apparaît sur la troisième dimension: les conversations téléphoniques y sont plus proches du pôle *référence ancrée dans la situation* que les conversations en face à face (1995: 157). Ce résultat est surprenant parce que l'étude de Moscovici et Plon (1966) a démontré que des

Une distinction mesurable

interlocuteurs qui se voient emploient considérablement moins de substantifs et de ‘mots de liaison’ que lorsqu’ils ne se voient pas. Ils ont également constaté que les conversations téléphoniques sont beaucoup plus nominales que les conversations en face à face. Celles-ci sont en effet davantage ancrées dans le contexte spatial, les rendant ainsi plus ‘impliquées’ que les conversations téléphoniques. Le problème de validité des dimensions de Biber découle selon nous du niveau et du grand nombre de variables linguistiques très hétérogènes. Sa première et sa troisième dimension semblent en outre refléter des concepts très similaires: que peut donc distinguer le pôle ‘impliqué’ de sa première dimension, du pôle ‘référence ancrée dans la situation’ de sa troisième dimension? Ne s’agit-il pas du même phénomène?

Le choix des termes ‘informationnel’ vs. ‘impliqué’ nous semble malheureux, car il suggère qu’un discours plus impliqué contiendrait moins d’information, ayant un contenu plus général et plus ambigu (Biber, 1995: 144), ce qui est discutable.² Il convient cependant de reconnaître que Biber a dépassé les dichotomies traditionnelles dans la description des registres oraux et écrits en les plaçant sur des continuums à l’aide de techniques statistiques rigoureuses. Il combine d’ailleurs ses analyses quantitatives avec des analyses qualitatives. Biber *et al.* (1995) ont en outre appliqué l’analyse multidimensionnelle à des corpus oraux et écrits d’autres langues. Il est ainsi apparu que la dimension 1 de l’anglais présente des similarités avec la dimension 3 en coréen et la dimension 1 en somalien.

Nous nous pencherons à présent sur quelques études qui se distinguent des précédentes par leur absence de traitement statistique de données tirées de corpus fermés de taille importante. Ces approches ont cependant permis de dégager des variations entre les modalités au niveau des proportions des catégories grammaticales. Les analyses se concentrent le plus souvent sur les ‘mots pleins’ ou formes ‘lexicales’. L’intérêt pour les mots ‘grammaticaux’ aussi appelés ‘mots-outils’ ou ‘mots vides’ y est très limité.

Les psychologues français Moscovici et Humbert (1960), par exemple, calculent les proportions variables des adverbes, des verbes et des substantifs pour distinguer le français oral du français écrit. Ils remarquent que les adverbes sont presque deux fois plus nombreux à l’oral qu’à l’écrit. Ce sont surtout les verbes et les substantifs qui retiennent leur attention. Ces deux classes accusent en effet la variation la plus prononcée en français. Le ‘quotient verbes/substantifs’ (QVS) semble être systématiquement plus élevé dans le langage oral que dans le langage écrit. Quelques années plus tard Moscovici et Plon (1966), déjà mentionnés précédemment, approfondiront ces recherches sur la variation du QVS. Leur conclusion confirme les résultats antérieurs: le discours oral se distingue par une proportion plus importante de verbes.

² Biber et Finegan (1994: 320) proposent une meilleure interprétation de cette dimension une qui ne sera cependant plus mentionnée dans les publications ultérieures. Les pôles refléteraient, selon les auteurs, l’élaboration et l’économie en termes greciens.

Jean-Marc Dewaele

Müller (1985) entreprend également ce type de comparaison de catégories grammaticales. Dans son analyse de quelques échantillons de textes du 'Monde', il découvre qu'il y a, au niveau de la fréquence absolue des occurrences, 26,7% de substantifs, 8,1% d'adjectifs, 16,8% de verbes, 5% d'adverbes et 43,5% de 'mots grammaticaux'. Il ajoute que la fréquence des mots grammaticaux peut même approcher 50% en français parlé. Biber *et al.* (1998) présentent également une brève analyse de certaines catégories grammaticales. Ils calculent le quotient substantifs-verbes dans plusieurs registres oraux et écrits et constatent qu'il est beaucoup plus élevé dans la prose académique que dans les conversations, ce qui refléterait selon les auteurs le caractère plus statique de l'écrit et plus dynamique de l'oral (1998: 69). Après avoir constaté une proportion supérieure de pronoms à l'oral, les auteurs approfondissent leur analyse et découvrent que dans les conversations en face à face, plus de la moitié des pronoms sont de nature exophorique. La prose académique contient quelques pronoms anaphoriques mais la référence s'y fait davantage avec des substantifs anaphoriques. Les auteurs ne comparent malheureusement pas les conversations en face à face avec les conversations téléphoniques où on imagine qu'il y aurait moins de référence exophorique.

Aucun des chercheurs cités ne fournit cependant une hypothèse convaincante pour expliquer la variation observée.

L'oeuvre de référence dans cette tradition est sans aucun doute le livre de M. A. K. Halliday *Spoken and written language* (1989). Son influence a été, et est encore, fondamentale (Blanche-Benveniste, 1994; 1995).

Halliday (1989) développe quatre critères lexicaux et syntaxiques pour distinguer le discours écrit du discours oral. Deux critères, ceux de la densité lexicale et du nombre de mots lexicaux par énoncé, sont basés sur la traditionnelle dichotomie entre mots lexicaux et mots grammaticaux. Les deux critères syntaxiques considèrent la complexité de la structure de la phrase et le nombre d'énoncés. Le discours oral se distinguerait par des valeurs élevées au niveau syntaxique, alors que le discours écrit serait plus complexe au niveau lexical. La densité lexicale se calcule en considérant la proportion relative de mots lexicaux 'pleins' dans un échantillon de discours (1989: 62). La catégorie des substantifs est le membre principal de l'ensemble des mots lexicaux. A la classe des mots non-lexicaux appartiennent selon lui les déterminants – dont les articles –, les pronoms, la plupart des prépositions, les conjonctions, certains adverbes et les verbes conjugués ('finite verbs'). Halliday estime que seule la proportion relative des mots lexicaux est importante: 'The relative frequency of grammatical items can be ignored, since all of them fall into the relatively frequent bracket (1989: 64). Certains mots lexicaux comme *thing, people, way, do, get, have, go, good, many*' (1989: 64) se situent à la frontière des mots grammaticaux à cause de leur fréquence élevée dans le discours. Leur présence répétée risque de diluer la densité lexicale d'un discours. L'idéal serait selon Halliday que le linguiste engagé dans le calcul de la densité lexicale d'un échantillon distingue les mots lexicaux à haute fréquence de ceux qui sont

Une distinction mesurable

moins fréquents. Le degré de densité lexicale, tel qu'il est conçu par Halliday, se baserait donc sur la proportion de mots lexicaux à faible fréquence.

Deux observations s'imposent d'ores et déjà. Il semble en premier lieu que Halliday base son critère de densité lexicale sur des observations théoriques plutôt qu'empiriques. Les échantillons de discours que l'auteur propose comme illustration de sa théorie sont trop courts pour servir de fondement empirique à ses affirmations. Dans la colonne de gauche, il présente la phrase en style 'écrit', par exemple: '*Slavish imitation of models is nowhere implied.*' Dans la colonne de droite on trouve l'équivalent oral de la même phrase: '*It is not implied anywhere that there are models which should be slavishly imitated*' (1989: 62). Considérant ces deux exemples, Halliday conclut que le discours écrit est lexicalement plus dense.

Deuxièmement, Halliday ne spécifie pas à partir de quel moment un mot lexical devient un mot à 'haute fréquence'. La mise en pratique du critère de densité lexicale paraît par conséquent difficile. Les mêmes remarques valent également pour son critère de 'mots lexicaux par énoncé'.

Halliday constate ensuite que le contenu lexical se concentre dans des groupes nominaux: '*All the meat of the message is in the nominals*' (1989: 66). Il explique cette forte tendance à encoder le contenu lexical dans des structures nominales par l'élasticité de celles-ci (1989: 75). Il nous semble que l'élasticité n'est pas l'unique raison motivant l'utilisation de structures nominales dans le discours. Nous développerons des explications alternatives plus loin dans la présente étude. Halliday se rend bien compte qu'il est difficilement acceptable de définir le discours oral uniquement par des caractéristiques négatives, c'est-à-dire par l'absence relative ou le taux limité de la densité de l'information (1989: 66). Halliday constate que les phénomènes sont représentés comme des produits dans le discours écrit, alors que dans le discours oral ils sont représentés comme des configurations de processus reliés les uns aux autres de façons différentes. La conséquence de cette différence est, selon lui, que le discours oral est plus complexe syntaxiquement (1989: 86). Sa conclusion est que le discours écrit est statique et dense tandis que le discours oral est dynamique et syntaxiquement complexe. Le discours oral possède une structure 'chorégraphique' (1989: 87).

En focalisant son attention sur les mots lexicaux, Halliday néglige cependant l'interprétation de la variation des catégories grammaticales des verbes conjugués, des adverbes et des pronoms. Le fait qu'un pronom n'ait pas de contenu lexical n'implique pas qu'il est un piètre porteur d'information comme le suggère Halliday. Dans un contexte spatio-temporel donné, l'énoncé *Il le leur a dit* est aussi dense en information que le pendant nominalisé *X a dit Y à Z*. Il nous semble au contraire que les mots non-lexicaux permettent davantage de condenser l'information.

Cette faiblesse dans sa théorie peut s'expliquer par une approche purement théorique de l'opposition oral-écrit et par une reprise de concepts traditionnels non-opérationnels. Comment peut-on imaginer que la proportion de sub-

Jean-Marc Dewaele

stantifs puisse varier indépendamment de celle des déterminants qui l'accompagnent ? La vision classique, reprise par Halliday, selon laquelle le discours écrit refléterait davantage des phénomènes statiques, considérés comme produits, alors que le discours oral rendrait plutôt des phénomènes dynamiques, repose sur l'opposition aristotélicienne entre état et action. Les substantifs rendraient les états tandis que les verbes refléteraient l'action. Un discours riche en substantifs serait par conséquent plus 'statique'. L'opposition substantif-statique versus verbe-dynamique et son dérivé écrit-statique versus oral-dynamique peut paraître valable à première vue, mais une analyse critique révèle les insuffisances d'une telle distinction. Halliday, qui fait du substantif le membre principal des 'mots lexicaux', ne prétend d'ailleurs pas mesurer le caractère statique d'un discours mais sa densité lexicale, c'est-à-dire sa densité d'information. L'instrument qu'il développe ne mesure donc pas ce qu'il prétend être la différence fondamentale entre discours écrit et oral, notamment l'opposition 'phénomène considéré comme produit' ou 'phénomène considéré comme processus'.

Il convient, aux termes de ces analyses, de dresser un bilan provisoire. Le mérite des travaux de Redeker (1984), de Mazzie (1987) et de Biber (1988; 1995), de Biber et Finegan (1991, 1994) et de Biber *et al.* (1994, 1998) est d'avoir travaillé sur des corpus fermés et d'avoir utilisé des techniques statistiques rigoureuses. Il ressort de ces études que le discours oral est globalement plus impliqué ou implicite, davantage ancré dans le contexte non-linguistique que le discours écrit. Ces chercheurs constatent que d'autres variables co-varient avec la modalité et qu'il faut en tenir compte. Le défaut de ces analyses réside, selon nous, dans l'hétérogénéité des variables linguistiques qui réduit la valeur psychométrique des dimensions. Ce type de défaut est absent dans la deuxième série d'études (Moscovici et Humbert, 1960; Moscovici et Plon, 1966; Müller, 1985; Biber *et al.*, 1995; Halliday, 1989) puisque les analyses portent uniquement sur des catégories grammaticales. Les analyses sont malheureusement incomplètes à cause de la prédilection pour les mots lexicaux et l'absence d'intérêt pour les mots grammaticaux. Les corpus utilisés sont souvent limités et l'absence de traitement statistique des données limite la portée des conclusions.

Nous proposons donc d'analyser de façon rigoureuse la variation de toutes les catégories grammaticales dans des corpus oraux et écrit, avant de présenter une mesure unique pouvant refléter la position d'un extrait de discours sur le continuum oral/écrit.

3 ANALYSE DES DONNÉES

3.1 *La proportion de catégories grammaticales comme variable dépendante*

Jolivet (1982) se demande si le linguiste un peu consciencieux peut reprendre la taxinomie traditionnelle des catégories grammaticales sans aucune critique.

Une distinction mesurable

Il conclut finalement que c'est le caractère probabiliste de l'attribution d'un 'mot' à une partie du discours, énoncé explicitement par Zipf, 'qui rend "malgré tout" utilisables les renseignements épars et non homogènes que l'on peut trouver dans la littérature, si toutefois les valeurs citées ont été observées sur des échantillons de taille suffisante et si les classes retenues ne sont pas trop fines' (1982: 258).

Nous considérerons la variation dans la proportion des catégories grammaticales du discours écrit et oral dans six corpus oraux et quinze corpus écrits. Comme il a été suggéré que les différences entre l'écrit et l'oral seraient d'ordre universel (Fielding et Fraser, 1978: 223), nous analyserons des corpus de langues typologiquement assez proches, notamment l'italien, le néerlandais, le français ainsi que l'interlangue française de locuteurs néerlandophones.

3.2 Méthodologie

Les cinq corpus italiens, comptant chacun 100.000 mots, sont tirés du dictionnaire de fréquence de Bortolini *et al.* (1971). Il comprend un corpus de journaux, de littérature technique, de romans, de théâtre et de scénarios de cinéma. Ces deux derniers corpus permettront éventuellement de retrouver des traces de l'oral dans l'écrit.

Les neuf corpus néerlandais sont tirés des dictionnaires de fréquence de Uit den Boogaart (1975) et De Jong (1979). Uit den Boogaart (1975) contient cinq corpus écrits, comptant chacun 120.000 mots (journaux, textes techniques, revues, magazines et romans) et deux corpus oraux (comptant chacun 60.000 mots) constitués à partir d'interviews avec des locuteurs ayant un niveau d'éducation académique et non-académique. Le dictionnaire de fréquence de De Jong (1979) est basé sur des conversations avec 80 locuteurs (40 hommes et 40 femmes) dans une situation informelle et une situation formelle. Dans le premier cas les locuteurs parlaient de sujets familiers avec des amis, dans le deuxième cas ils parlaient de sujets plus généraux avec les chercheurs. Les locuteurs constituaient un échantillon représentatif de la population en termes d'âge et de statut social. De Jong a retenu un extrait de 750 mots de chaque locuteur dans chacune des situations. Le corpus informel (constitué de 8 subcorpus) sera comparé avec le corpus formel (également constitué de 8 subcorpus).

Comme aucun dictionnaire de fréquence français ne fournit des données fiables et complètes sur la composition de son lexique, nous avons constitué deux corpus écrits: un premier de 825 mots basé sur un extrait d'un éditorial de 1991 intitulé 'Fin de partie dans la Corne' de G. Salamé dans 'Libération' et un deuxième de 815 mots basé sur un extrait d'un discours de F. Mitterrand paru au même moment dans 'Le Monde') et deux corpus de français parlé (un extrait de 800 mots d'une interview télévisée avec une call-girl parisienne et un extrait d'une interview de 805 mots avec F. Mitterrand dans l'émission 'Sept sur sept' avec A. Sinclair).

Jean-Marc Dewaele

Nous avons également utilisé trois corpus d'interlangue française avancée produits par 27 étudiants néerlandophones (8 femmes, 19 hommes). Un premier corpus de 16.273 mots est basé sur des conversations avec le chercheur locuteur natif dans une situation formelle (un examen oral). Le deuxième corpus de 31.900 mots est basé sur des conversations avec les mêmes locuteurs dans une situation informelle. Les sujets traités dans les deux situations couvraient les études, les loisirs, les idéaux, les convictions politiques et religieuses. Ces entretiens d'une durée de dix à vingt minutes ont été immédiatement transcrits en français orthographique (Blanche-Benveniste et Jeanjean, 1987). Le corpus compte environ 15 heures de conversation, ce qui représente 51.103 mots d'apprenants. Le troisième corpus est basé sur les dissertations d'examen des mêmes étudiants et contient 8.349 mots. Le choix des corpus d'interlangue avancée est motivé par les possibilités qu'ils offrent d'analyser non seulement l'effet de la modalité, mais aussi l'effet de la formalité de la situation sur le discours oral d'un même groupe d'apprenants.

3.3 Composition des lexiques écrits et oraux

Afin d'analyser la variation entre discours oral et discours écrit au niveau de la composition lexicale, nous avons calculé les proportions de mots appartenant aux différentes classes grammaticales dans les 21 corpus (voir tableau 1).

L'effet de la modalité sur la proportion de chaque catégorie grammaticale a été analysé avec un test 't' de student. Les résultats (voir tableau 2) suggèrent, qu'à l'exception des verbes, la proportion de toutes les catégories grammaticales varie de façon significative entre l'écrit et l'oral. Les corpus écrits, au niveau des occurrences, sont plus riches en substantifs, articles, adjectifs, prépositions et conjonctions. Les corpus oraux par contre sont plus riches en pronoms, adverbes et interjections. La proportion de verbes y est marginalement plus élevée. Le sens de la variation est identique pour les différentes langues et interlangues.

L'hypothèse de Halliday, fondée sur la dichotomie 'mots grammaticaux / mots lexicaux', semble être partiellement en contradiction avec les faits: effectivement, la proportion des mots lexicaux tels que les substantifs et les adjectifs croît de façon significative dans le discours écrit, mais l'augmentation d'articles et de prépositions, qui sont pourtant des mots grammaticaux, est tout aussi marquée. Prétendre que la densité lexicale est plus élevée dans la langue écrite paraît par conséquent relever d'une généralisation abusive.

Avant de proposer une interprétation théorique des résultats, il convient de raffiner davantage notre analyse de données. Il faudrait idéalement une valeur unique résumant l'ensemble de la variation observée entre les 21 corpus qui nous permettrait de situer chaque corpus sur le continuum oral/écrit. L'analyse factorielle permet d'obtenir un indice statistique fiable (Biber, 1988, 1998; Dewaele, 1995). Celui-ci se calcule en soumettant la proportion des classes grammaticales de chaque corpus à une analyse factorielle. Le but de l'analyse

Une distinction mesurable

Tableau 1. Composition (en %) des corpus français (F), italiens (I), néerlandais (N) et d'interlangue française (IL) écrits (ECR) et oraux (OR)

Corpus	subs- tantifs	articles	adjec- tifs	pro- noms	verbes	adver- bes	prépo- sitions	conjon- ctions	inter- jections
ECR journal F	27.7	17.6	16.4	3.6	10.7	4.2	15.0	4.3	0.0
ECR technique I	23.0	18.4	13.1	5.2	16.2	4.2	14.9	5.0	0.0
ECR journal I	21.9	15.9	12.8	5.1	17.2	6.1	16.1	5.0	0.1
ECR journal N	26.0	14.7	10.6	5.6	16.7	7.2	14.5	4.7	0.0
ECR technique N	23.1	15.0	10.8	6.7	16.6	8.0	13.8	6.0	0.0
ECR revue N	24.2	11.6	10.9	8.6	17.7	8.7	13.9	4.3	0.1
ECR roman I	18.7	13.4	11.1	8.6	20.2	8.2	13.5	6.2	0.2
ECR magazine N	21.8	9.8	11.1	10.1	18.7	9.7	12.2	6.4	0.1
ECR discours F	20.2	14.7	7.9	11.0	17.7	9.3	11.6	7.2	0.0
ECR rédaction IL	20.8	13.4	8.8	13.7	18.9	8.1	9.5	6.4	0.1
ECR roman N	18.5	10.5	10.0	13.3	20.6	10.5	10.3	6.1	0.3
ECR théâtre I	16.3	10.0	10.9	11.8	24.2	11.2	9.1	5.4	1.1
ECR cinéma I	14.7	8.1	10.2	12.0	26.7	12.7	8.6	5.9	1.1
OR formel F	17.1	11.1	9.5	12.7	18.9	10.8	9.9	7.8	1.8
OR formel IL	18.3	13.8	7.6	8.5	19.6	10.3	9.0	11.9	0.9
OR informel F	16.2	10.6	7.6	18.6	20.0	11.5	7.8	5.1	2.5
OR académique N	13.8	10.0	7.5	14.6	18.6	18.7	8.3	7.5	1.2
OR informel IL	13.5	10.5	6.0	17.9	18.7	16.6	6.7	8.2	1.7
OR n/académique N	13.3	8.9	7.0	16.7	19.6	20.1	6.6	6.6	1.3
OR formel N	11.2	8.3	8.2	16.2	18.9	16.4	6.5	7.8	6.6
OR informel N	10.8	6.8	7.5	16.7	18.9	17.5	6.1	6.8	9.0

Tableau 2. Résultats d'un test 't' sur l'effet de la modalité sur les proportions des catégories grammaticales

	Valeur de 't'	Valeur de 'p'	Moyenne Ecrit	Moyenne Oral
substantifs	4.7	0.000	21.3%	14.3%
articles	2.6	0.017	13.3%	10.0%
adjectifs	4.4	0.000	11.1%	7.6%
pronoms	-4.2	0.000	8.9%	15.2%
verbes	-0.4	0.713	18.6%	19.1%
adverbes	-5.0	0.000	8.3%	15.2%
prépositions	5.1	0.000	12.5%	7.6%
conjonctions	-3.4	0.003	5.6%	7.7%
interjections	-3.5	0.002	0.2%	3.1%

factorielle est de réduire une masse de variables originales à un petit nombre de 'super-variables' ou facteurs dérivés. Comme les éléments soumis à l'analyse sont de nature identique (catégorie grammaticale), la valeur psychométrique du facteur devrait être élevée.

Un seul facteur (ou dimension) émerge de l'analyse factorielle, expliquant

Jean-Marc Dewaele

Tableau 3. Coefficients de saturation des catégories grammaticales sur la dimension 'Explicité'

Catégorie	Coefficients de saturation
substantifs	-.96
articles	-.88
adjectifs	-.86
prépositions	-.97
pronoms	.92
verbes	.59
adverbes	.92
conjonctions	.52
interjections	.69

69% de la variance dans les données. Les coefficients de saturation permettent d'identifier les éléments caractérisant les deux pôles (voir tableau 3). Ce facteur a été baptisé 'explicité/implicite'. Il corrèle de façon hautement significative avec d'une part, les pronoms, les verbes, les adverbes, les interjections et de d'autre les noms, les articles, les adjectifs et les propositions. Le coefficient pour les conjonctions est non-significatif. Il est possible de placer chaque corpus sur ce continuum grâce aux notes individuelles (voir figure 1). Les corpus écrits se situent du côté du pôle explicite du continuum alors que les corpus oraux se situent près du pôle implicite. Une analyse de variance (ANOVA), avec la modalité comme effet principal, confirme que la différence en degré d'explicité des corpus est statistiquement significative: $F(1, 19) = 24,7$; $p < 0,0001$. La variable 'langue' n'a pas d'effet significatif: $F(1,2) = 0,774$; $p < 0,501$.

Une observation de l'ordre des corpus à l'intérieur de chaque modalité révèle à nouveau une grande similitude entre les corpus des différentes langues. Les corpus journalistiques, techniques sont plus explicites que les romans, les discours. L'écrit 'oral' (Gregory et Carol, 1981), comme les textes de théâtre ou de scénarios de cinéma, se rapprochent davantage du pôle implicite. Il y a à cet endroit même un petit chevauchement entre corpus écrits et oraux. Il est intéressant aussi de constater que le corpus d'interlangue écrite se situe vers le milieu du continuum. Cela suggère que le style écrit des non-natifs est encore assez proche de l'oral. Les corpus de néerlandais et d'interlangue qui opposent les styles des mêmes individus recueillis en situation 'formelle' et 'informelle' montrent que le discours en situation formelle devient plus explicite. Dewaele (1999b) a constaté à l'aide d'une analyse de variance multiple (MANOVA) que l'effet de la situation sur les 16 subcorpus néerlandais de De Jong (1979) est statistiquement significatif: $F(1, 4) = 24,4$; $p < 0,0002$). La variation est similaire, et tout aussi significative, entre les deux composante du corpus d'interlangue orale: $F(1, 52) = 35,5$; $p < 0,0000$ (Dewaele, 2000). Nous avons soutenu qu'un locuteur opte pour un style plus

Une distinction mesurable

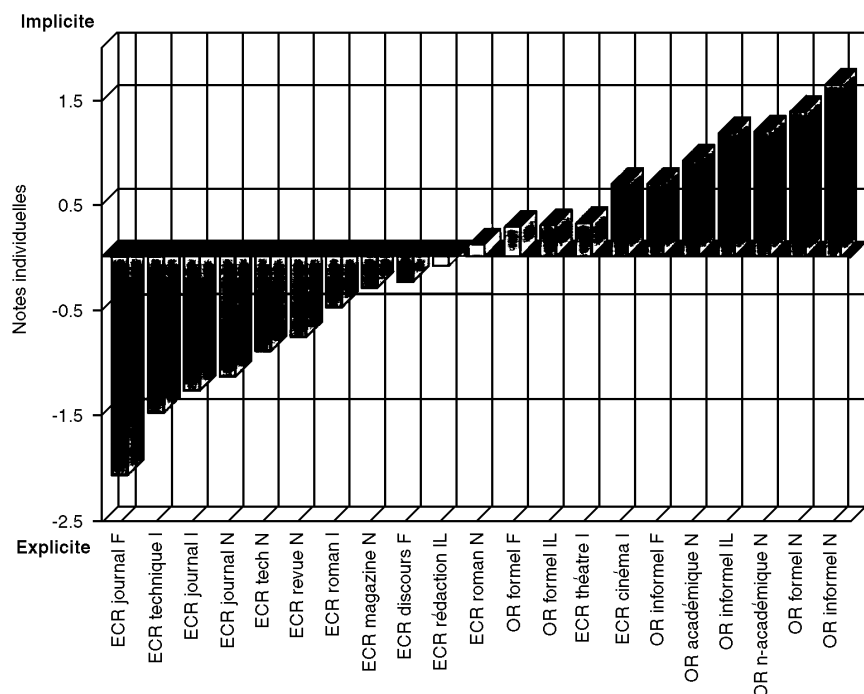


Figure 1. Situation des corpus oraux et écrits sur la dimension de la deixis

implicite parce que celui-ci est plus économique du point de vue cognitif (Dewaele, 1995; 1996b). Si toutefois le locuteur craint, pour une raison probablement inconsciente, que son intention communicationnelle soit mal inférée, il optera pour un style plus explicite et plus lourd. Il applique ainsi la ‘maxime de quantité’ de Grice dont l’interprétation varie en fonction de la situation. Nous avons démontré que la personnalité du locuteur (Dewaele, 1999a), son statut social (2000), son sexe (1998), déterminent le choix du degré d’explicité du style oral.

4 DISCUSSION

Notre analyse empirique des variations de proportions des catégories grammaticales a ouvert la voie à un nombre d’observations devant nous permettre de déduire le principe sous-jacent qui distingue le groupe substantif, articles, adjectifs, prépositions/discours écrit versus verbes, adverbes, pronoms, interjections/discours oral. La différence réside dans ce que Levelt (1989) appelle ‘le degré d’ancrage des énoncés dans le contexte spatio-temporel’ et nous permet ainsi de renouer avec les observations de Olson (1977), de Redeker (1984) de Biber (1988, 1994, 1995) et de Biber *et al.* (1998). De ce fait, on peut

Jean-Marc Dewaele

transcender l'opposition manichéenne de Halliday (1989) pour qui l'oral est lexicalement moins dense (ce qui serait un trait négatif) mais syntaxiquement plus complexe (ce qui serait un trait positif).

Le choix de la modalité par l'émetteur détermine clairement la variation. Cette variation entre les deux modalités découle des choix que le locuteur fait à chaque instant au cours de sa production langagière. S'il opte pour le discours écrit, l'émetteur utilise automatiquement un discours dont les énoncés sont moins ancrés dans le contexte spatio-temporel, ce qui se traduit par un nombre accru d'occurrences de substantifs, d'articles, d'adjectifs et de prépositions. Leckie-Tarrie (1995) constate que lorsque le contexte de la situation est absent, le discours devra être plus explicite. Il faudrait idéalement, selon elle, calculer le degré d'explicité/implicité de chaque mot individuel (1995: 123). Un tel calcul semble difficile mais il est possible de retrouver les mots implicites dans les classes grammaticales qui contiennent une majorité de mots pouvant servir à installer un des quatre types de deixis énumérés par Levelt (1989).

- 1) La deixis de personne qui s'exprime à l'aide de pronoms déictiques qui changent de référent chaque fois qu'un autre locuteur prend la parole.
- 2) La deixis de lieu qui peut s'exprimer à l'aide d'adverbes comme 'ici' et qui se réfère aux environs immédiats du locuteur.
- 3) La deixis temporelle qui s'établit à travers des mots comme 'avant', 'demain'. Ces mots n'ont de sens que par rapport au 'moi-ici-maintenant' du locuteur et aux connaissances partagées des interlocuteurs. Un autre type de deixis temporelle s'installe avec les temps verbaux et l'aspect des verbes conjugués. L'usage déictique des temps peut se faire en se référant à l'actualité du locuteur ou à une actualité différente, indépendante de celui-ci. Des phrases comme *Je viendrai vendredi* ou *Je venais vendredi* n'ont de sens pour l'interlocuteur que s'il connaît le contexte temporel dans lequel le locuteur a ancré ses énoncés.
- 4) La deixis du discours qui apparaît lorsqu'un interlocuteur se réfère à un élément précédent ou suivant dans la conversation (anaphores ou cataphores).

L'on objectera avec raison que les adjectifs démonstratifs sont des exemples typiques de déictiques et que malgré tout la classe des adjectifs ne semble pas contribuer grandement à ancrer les énoncés dans le contexte spatio-temporel. Il convient ici de remarquer la proportion modeste des adjectifs démonstratifs dans la classe des adjectifs (moins de 8% de l'ensemble des adjectifs dans notre corpus d'interlangue orale). Les adjectifs qualificatifs qui constituent plus de la moitié des occurrences d'adjectifs ne peuvent pas fonctionner comme déictiques. La classe des verbes n'est pas tout à fait homogène non plus en termes de deixis: elle contient une majorité de formes déictiques (les verbes conjugués représentent 67% de l'ensemble des occurrences dans notre corpus d'inter-

Une distinction mesurable

langue orale) et une minorité (33%) de formes non-déictiques comme les infinitifs et les participes.

Nos résultats offrent un support empirique à la thèse de Kleiber (1991: 294) qui estime que les articles, et plus particulièrement l'article défini, n'ont pas de valeur déictique. L'article défini ne renvoie pas directement au contexte spatio-temporel et diffère donc fondamentalement de l'adjectif démonstratif. Crystal (1991: 96) semble également estimer que les articles n'ont pas de valeur déictique. Aucun article ne figure parmi les exemples cités par l'auteur pour illustrer les différents types de déictiques. L'appartenance des articles au groupe des classes grammaticales non-déictiques dans nos corpus (tableaux 2 et 3) n'est par conséquent guère surprenante.

Les interjections illustrent le 'backchanneling' constant dans la conversation. Elles permettent au locuteur de vérifier si l'ancrage spatio-temporel des énoncés est correct et soulignent les intentions communicationnelles.

L'émetteur-écrivain qui désire donner à son discours un 'parfum d'oralité' atteint son objectif en ancrant davantage les énoncés dans un contexte spatio-temporel spécifique, ce qui se traduit par un rapprochement des deux structures, la structure lexicale du discours écrit et la structure du discours oral. A cela s'ajoutent alors éventuellement des choix lexicaux particuliers et d'autres phénomènes de surface. Les énoncés prononcés par des acteurs de cinéma sont en grande mesure ancrés dans le contexte spatio-temporel. Les indications scéniques permettent d'éviter les malentendus qui pourraient découler d'énoncés du type: *Je le vois*. Il en est de même pour le discours théâtral où les énoncés des acteurs s'ancrent dans un contexte spatio-temporel recréé par l'auteur. Le texte romanesque constitue un exemple de discours globalement moins ancré dans le contexte spatio-temporel de l'écrivain, qui peut cependant, comme nous venons de le voir, introduire des passages solidement ancrés dans ce contexte spatio-temporel. Les textes scientifiques et journalistiques, finalement, présentent les meilleurs exemples de discours pouvant exister en s'ancrant de façon minimale dans le contexte spatio-temporel de l'écrivain. Rappelons toutefois que ces constatations ne valent que pour des ensembles d'énoncés constituant des corpus. Il est évident que des exemples individuels d'énoncés ancrés dans le contexte spatio-temporel peuvent apparaître dans tous les types de corpus.

Contrairement à l'émetteur-écrivain, l'émetteur-locuteur produit un discours davantage ancré dans le contexte spatio-temporel, plus pauvre en constructions nominales; ses substantifs et déterminants sont moins nombreux. Les différents types de deixis qui ancrent les énoncés dans ce contexte spatio-temporel s'établissent à l'aide des catégories grammaticales qui s'y prêtent le mieux, c'est-à-dire les verbes conjugués, la majorité des adverbes et des pronoms. L'importance d'une catégorie particulière dans l'établissement de la deixis varie suivant la langue. En français, par exemple, les verbes servent davantage à établir une deixis de temps que de personne. La création d'une

Jean-Marc Dewaele

deixis de personne revient progressivement aux pronoms. Ce phénomène est déjà présent en français standard où aucune différence phonologique n'apparaît dans les désinences verbales des trois premières personnes du singulier et de la troisième personne du pluriel des verbes en '-ER'. Le rôle du pronom dans l'établissement de la deixis de personne s'accroît donc en français. Il en est autrement dans les langues qui n'utilisent pas de pronoms sujets comme l'espagnol ou l'italien. Les verbes doivent donc établir non seulement la deixis temporelle mais aussi la deixis de personne. Un énoncé contenant un verbe sera par conséquent beaucoup plus dépendant du contexte spatio-temporel qu'un énoncé sans verbe. Nous illustrerons la variation en ancrage des énoncés dans le contexte spatio-temporel par quelques exemples spécifiques. Ceux-ci sont artificiels dans le sens que notre approche de la composition du lexique exige un nombre de mots suffisamment important (500 mots semble un minimum, cf. Jolivet, 1982).

Dans l'énoncé *Ils ont détruit le bâtiment*, l'on distingue une deixis de personne ('ils'), de temps et d'aspect (passé global, cf. Wilmet, 1997), de discours ('le bâtiment dont on sait qu'il était là'). La deixis de personne peut être évacuée, ce qui résulte en un énoncé plus formel et passif: *Le bâtiment a été détruit*. La deixis de temps disparaît lorsqu'on remplace le verbe par son substantif correspondant: *La destruction du bâtiment*. Cet énoncé nominalisé est beaucoup plus statique, détaché et impersonnel. Il peut exprimer une règle générale et abstraite qu'on trouverait dans un manuel: *La destruction de bâtiments est une activité qui exige une bonne préparation*. L'énoncé a perdu sa référence à un événement particulier comme dans la phrase originale.

Il semble que l'ancrage des énoncés dans le contexte spatio-temporel est une caractéristique de tous les discours oraux. Fielding et Fraser (1978) présente des données partielles qui suggèrent que l'oral est plus implicite en anglais, Hudson (1994) observe une variation similaire en grec et en suédois, Biber (1995) et Biber et Finegan (1994) trouvent les mêmes tendances en anglais, en somalien, en koréen et en nukulaelae tuvaluan.

5 CONCLUSION

L'analyse de quelques travaux marquants sur les différences entre l'oral et l'écrit a révélé un nombre de pistes utiles et de constatations récurrentes. En tentant d'éviter les écueils méthodologiques et interprétatifs, nous avons analysé l'opposition oral/écrit à travers la variation dans les proportions des catégories grammaticales dans vingt-et-un corpus oraux et écrits italiens, néerlandais, français et d'interlangue française.

L'assertion de Halliday (1989) selon laquelle le discours écrit est plus riche en mots lexicaux que le discours oral ne s'est révélée que partiellement exacte. Il est vrai que, dans les vingt-et-un corpus, la proportion de substantifs est plus élevée dans le discours écrit, mais celle des mots 'grammaticaux': articles, adjectifs et prépositions l'est également ! Empruntant à Biber (1988) l'usage de

Une distinction mesurable

l'analyse factorielle, nous avons extrait une dimension unique sur laquelle s'opposent deux groupes de catégories grammaticales. Les substantifs, les articles, les adjectifs et les prépositions se situent près du pôle négatif tandis les pronoms, les verbes, les adverbes, les interjections et les conjonctions se situent près du pôle positif. Cette opposition reflète le continuum de la deixis. Nos six corpus oraux se situent près du pôle déictique/implicite du continuum, car les énoncés oraux sont davantage ancrés dans le contexte spatio-temporel du locuteur. Le discours écrit est généralement plus indépendant du contexte spatio-temporel, ce qui explique son caractère plus nominal et sa proportion plus importante de mots essentiellement non-déictiques ou explicites. En limitant aux seules catégories grammaticales les variables linguistiques soumises à l'analyse factorielle, nous avons obtenu une dimension qui partage certaines ressemblances avec les dimensions un et trois de Biber (1988, 1995) mais qui semble plus solide, plus claire et plus généralisable à d'autres langues.

Une analyse de la variation entre différents registres oraux et écrits a révélé que les défenseurs de Tannen (1982) et les partisans de Olson (1977) ont tous partiellement raison: la modalité détermine clairement le degré d'implicité du discours, mais une fois que le choix du mode est décidé, d'autres facteurs comme le genre, la situation, les intentions communicatives des participants, leur profil social et psychologique affectent l'ancrage du discours dans le contexte spatio-temporel.

Il faudra entreprendre des recherches supplémentaires pour vérifier si le continuum de la deixis est également la principale dimension entre l'écrit et l'oral dans des langues typologiquement plus lointaines.

On pourrait, au terme de ce travail, reprendre le titre de Gadet (1996b), et admettre que la distinction entre oral/écrit est fragile, tout en ajoutant qu'une caractéristique fondamentale, la position d'un extrait de discours sur le continuum de la deixis, est désormais mesurable.

Author's address:

Jean-Marc Dewaele
Department of French
Birkbeck College, University of London
43 Gordon Square
London WC1H 0PD, UK

RÉFÉRENCES

- Beaman, K. (1984). Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In: D. Tannen (éd.), *Coherence in Spoken and Written Narrative Discourse*. Ablex: Norwood NJ, pp. 45–80.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. (1994). An analytical framework for register studies. In: D. Biber et

Jean-Marc Dewaele

- E. Finegan (eds.), *Sociolinguistic Perspectives on Register Variation*. New York et Oxford: Oxford University Press, pp. 31–56.
- Biber, D. (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.
- Biber, D., Conrad S. et Reppen R. (1994). Corpus-based approaches in applied linguistics. *Applied Linguistics*, 15, 2: 169–85.
- Biber, D. et Finegan, E. (1991). On the exploitation of computerized corpora in variation studies. In: K. Aijmer et B. Altenberg (eds.), *English Corpus Linguistics: Studies in the Honour of Jan Svartvik*. London: Longman, pp. 204–20.
- Biber, D. et Finegan, E. (1994). Register and social dialect variation: An integrated approach. In: D. Biber et E. Finegan (eds.), *Sociolinguistic Perspectives on Register Variation*. New York et Oxford: Oxford University Press, pp. 315–47.
- Biber, D., Conrad, S. et Reppen R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Blanche Benveniste, C. et Jeanjean, C. (1987). *Le Français parlé: transcription et édition*. Paris: Didier.
- Blanche Benveniste, C. et al. (1990). *Le Français parlé: études grammaticales*. Paris: Editions du CNRS.
- Blanche-Benveniste, C. (1994). Quelques caractéristiques grammaticales des ‘sujets’ employés dans le français parlé des conversations. In: *Actes du colloque Subjecthood and Subjectivity*. Paris et Londres: Orphys et Institut français du Royaume Uni, pp. 77–107.
- Bortolini, U., Tagliavini, C. et Zampolli, A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Rome: IBM Italia.
- Bresson, F. (1976). Langue écrite et langue parlée. *Français aujourd’hui*, 43: 67–74.
- Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In: D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Ablex, Norwood, NJ, pp. 35–53.
- Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.
- Dewaele, J.-M. (1995). Style-shifting in oral interlanguage: Quantification and definition. In L. Eubank, L. Selinker et M. Sharwood Smith (eds.), *The Current State of Interlanguage*. Amsterdam-Philadelphia: John Benjamins, pp. 231–38.
- Dewaele, J.-M. (1996a). How to measure formality of speech? A Model of Synchronic Variation. In: K. Sajavaara et C. Fairweather (éds.), *Approaches to Second Language Acquisition (Special issue)*, *Jyväskylä Cross-Language Studies*, 17, pp. 119–33.
- Dewaele, J.-M. (1996b). Variation dans la composition lexicale de styles oraux. *International Review of Applied Linguistics*, 24, 4: 261–82.
- Dewaele, J.-M. (1998). The effect of gender on the choice of speech style. *ITL Review of Applied Linguistics*, 119–20: 9–25.
- Dewaele, J.-M. (1999a). L’effet de l’extraversion sur la production du discours de bilingues. *Acquisition et Interaction en Langue Etrangère. Numéro spécial*, vol. 1, pp. 111–26.
- Dewaele, J.-M. (1999b). *Gender, social and situational variables in the choice of speech style in native Dutch*. Manuscrit inédit.
- Dewaele, J.-M. (2000). Quantifier le style dans la conversation. Une analyse de la variation sociolinguistique. *Le Langage et l’homme. Recherches pluridisciplinaires sur le langage*, 35, 4: 233–249.

Une distinction mesurable

- Fielding, G. et Fraser, C. (1978). Language and interpersonal relations. In: I. Markova (ed.), *The Social Context of Language*. Chichester: J. Wiley, pp. 217–32.
- Gadet, F. (1996a). Niveaux de langue et variation intrinsèque. *Palimpsestes*, 10: 17–40.
- Gadet, F. (1996b). Une distinction bien fragile: oral / écrit. *Travaux neuchâtelois de linguistique (TRANEL)*, 25: 13–27.
- Gadet, F. (1996c). Variabilité, variation, variété: le français d'Europe. *Journal of French Language Studies*, 6, 1: 75–98.
- Gregory, M. et Carroll, S. (1981). *Language and Situation*. London: Routledge et Kegan Paul.
- Gumperz, J. J., Kaltman, K. et O'Connor, M. C. (1984). Cohesion in spoken and written discourse. In: D. Tannen (ed.), *Coherence in Spoken and Written Discourse*. Norwood (NJ): Ablex, pp. 3–20.
- Halliday, M. A. K. (1989). *Spoken and Written Language*. Oxford: Oxford University Press.
- Hudson, R. (1994). About 37% of word-tokens are nouns. *Language*, 70: 331–39.
- Jolivet, R. (1982). *Descriptions quantifiées en syntaxe du français. Approche fonctionnelle*. Genève-Paris: Slatkine.
- Kleiber, G. (1991). Sur les emplois anaphoriques et situationnels de l'article défini et de l'adjectif démonstratif. In: D. Kremer (éd.), *Linguistique théorique et synchronique. Actes du XVIIIe Congrès International de linguistique et de philologie romanes*. Tübingen: Niemeyer, pp. 294–307.
- Leckie-Tarry, H. (1995). *Language and Context. A Functional Linguistic Theory of Register*, edited by David Birch. London–New York: Pinter.
- Levelt, W. (1989). *Speaking. From Intention to Articulation*, Cambridge, Mass – London: ACL-MIT Press.
- Marty, N. (ed.) (1991). L'écrit dans l'oral. *Etudes de linguistique appliquée*, 81.
- Mazzie, C. A. (1987). An experimental investigation of the determinants of implicitness in spoken and written discourse. *Discourse Processes*, 10: 31–42.
- Moscovici, S. et Humbert, C. (1960). Etudes sur le comportement verbal. Langage oral et langage écrit. *Psychologie française*, 3: 175–83.
- Müller, B. (1985). *Le Français d'aujourd'hui*. Paris: Klincksieck.
- Olson, D. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, 47: 257–81.
- Peytard, J. (1970). Oral et scriptural: Deux ordres de situations et de descriptions linguistiques. *Langue française*, 6: 35–47.
- Peytard, J. (1977). Situer l'oral. *Le Français moderne*, 45: 193–203.
- Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, 7: 43–55.
- Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 58: 1–21.
- Wilmet, M. (1997). *Grammaire critique du français*. Paris: Hachette.