DOCUMENT 1

Les corpus de linguistique romane en pays germanophones. Bilan et perspectives

Claus D. Pusch

Pub. linguistiques | *Revue française de linguistique appliquée*

2007/1 - Vol. XII
pages 111 à 124

ISSN 1386-1204

Rev. franç. de linguistique appliquée, 2007, XII-1 (111-124)

Document 1

Les corpus de linguistique romane en pays germanophones. Bilan et perspectives

Claus D. Pusch (Albert-Ludwigs-Universität Freiburg im Breisgau)

1. Introduction

Sous le titre « Cent ans de linguistique sur corpus », C. Marchello-Nizia (2005, 12) explique que cette linguistique sur corpus est loin d'être une innovation de la fin du XXe siècle :

> For over a century now, a considerable number of descriptive studies in historical linguistics have been based on the exhaustive analysis of a given text to study, synchronically or diachronically, a phenomenon from an earlier stage of a language. For Old French, the first work of this kind was most likely Jules Le Coultre's doctoral thesis, defended in 1875 in Dresden, Germany, on word order in one of Chrétien de Troyes' *romans*.

C'est peut-être le hasard qui a voulu qu'un des premiers travaux empiriquement basés sur une collection de données langagières structurée (pour reprendre une définition brève mais assez répandue d'un corpus linguistique, cf. Sinclair 2004) et dont Marchello-Nizia souligne l'importance pour l'étude du français, ait été présenté et publié en Allemagne. Si les pays germanophones, lieu d'origine de la discipline appelée « philologie romane », ont joué pendant presque un siècle un rôle de première importance dans l'étude de la langue et de la littérature françaises et de ses langues-sœurs issues du latin, leur engagement dans ce qu'on peut appeler, avec B. Habert (2005), l'approche outillée de la linguistique française et romane a été plus modeste, et est resté, de surcroît, mal connu. Le but du présent article est de dresser un bilan provisoire - sans prétention d'exhaustivité - de la contribution des romanistes germanophones dans le domaine des corpus linguistiques.

Pour ce faire, un certain nombre de réserves s'imposent : d'une part, les définitions actuelles de ce qui peut être considéré comme corpus linguistique au sens moderne présupposent, plus ou moins explicitement, que les données qui constituent le corpus soient accessibles sur support électronique, en ligne ou *off-line*. Un tel trait définitoire, certainement approprié dans d'autres contextes, réduirait considérablement l'éventail des travaux à présenter ici ; il a donc été écarté et j'inclurai aussi des corpus qui ne sont disponibles que sous forme imprimée. Par contre, je ne mentionnerai que les corpus publiquement accessibles et laisserai de côté toute référence à des corpus et bases de données non publiques ou à consulter sur place, chez son ou ses auteur(s). D'autre part, je me focaliserai davantage sur les corpus oraux au détriment des corpus écrits, dans le seul but d'une délimitation plus ciblée de l'objet de cette contribution ; car si, à la suite de C. Marchello-Nizia, on peut considérer comme « corpus linguistique » toute collection de données langagières structurée, quel que soit son statut diamésique (i.e. son origine médiale ; cf. § 4), il devient vite malaisé de déterminer combien de (pré)structuration il faut pour qu'un ensemble de textes devienne corpus. Pour remédier à cela, le support électronique est souvent choisi comme critère décisif : une anthologie de textes littéraires sous forme de volumes imprimés n'est pas, a priori, un corpus linguistique, mais dès que cette même anthologie est disponible sur cédérom ou comme base de données sur internet, elle est plus facilement considérée comme tel. Or, si ce critère définitoire est écarté pour les besoins du présent article, il ne pourra pas être appliqué pour faire la part entre les « vrais » corpus écrits et les « faux ». Ceci dit, la question de savoir où commence un corpus utilisable à des fins d'analyse linguistique, se pose aussi dans le domaine des données orales, car on sait bien que les critères de fidélité dans la transcription varient en fonction du public visé. Un corpus oral destiné aux sociologues ou aux historiens contemporains ne sera pas forcément à la hauteur des attentes des usagers

linguistes. Mais malgré ce flou que partagent les corpus oraux et les corpus écrits, « phénoménologiquement » un corpus oral est plus facilement identifiable qu'un corpus écrit, justement parce qu'il y a eu nécessairement une étape de transcription et qu'une certaine formalisation s'ensuit. Cela est d'ailleurs la dernière réserve que j'aimerais avancer : sont considérés corpus oraux dans cet article uniquement les collections de données langagières orales qui comprennent une mise à l'écrit du parlé, accompagnée ou non de documents audio ou visuels. L'existence des seuls enregistrements ne justifie pas, dans l'optique de mon article, de parler de corpus linguistique[1].

2. L'étude des langues romanes parlées : les corpus allemands pionniers

Comme le souligne O. Baude (2005, 11) dans le récent *Guide des bonnes pratiques [...] des corpus oraux*, « [l]es données de langue parlée collectée avant l'ère de l'informatique ne peuvent pas être comparées à ce qu'on appelle aujourd'hui « corpus de langue parlée ». [...] Aucune [sc. : des collections anciennes] ne pouvait atteindre de très grandes dimensions [...] et, dans ces données, la recherche d'information ne pouvait se faire que manuellement. » Ces inconvénients, si fortement ressentis de nos jours où l'analyse des corpus se fait presque uniquement à l'aide de logiciels concordanciers et où les conclusions à tirer d'une telle analyse passe surtout par la dimension quantitative, étaient moins pressants au moment de la constitution de ces premiers « corpus » oraux, car leur perspective était, dans beaucoup de cas, la dialectologie. Il s'agissait en effet de documenter un patrimoine linguistique considéré en péril, et l'intérêt des dialectologues dans cette démarche concernait surtout le lexique (en relation avec les objets de la culture du quotidien, souvent menacés aussi de disparition) et les formes narratives orales figées des « contes et légendes ». Dans ce contexte, il faut mentionner les travaux du *Deutsches Spracharchiv* ; si sa collection de publications *Lautbibliothek der deutschen Mundarten* (Göttingen, 1958-1964) était consacrée surtout aux dialectes allemands, la série qui lui a fait suite, *Phonai. Lautbibliothek der europäischen Sprachen und Mundarten* (Tübingen, à partir de 1969), contient aussi des échantillons de langues et dialectes romans. Or les documents de *Phonai* sont restés inaperçus dans les études romanes. La même constatation vaut pour les documents édités par le *Phonogrammarchiv* de l'Université de Zurich. Si cette institution s'est intéressée avant tout aux dialectes alémaniques de Suisse, elle a tout de même dans son inventaire quelques documents sonores en langues romanes du plus grand intérêt, notamment en ce qui concerne les dialectes rhéto-romanches des Grisons, les dialectes lombards du Tessin et les variétés franco-provençales et françaises de la Suisse Romande. La plupart de ces enregistrements sont anciens et ont parfois déjà une valeur historique, et un certain nombre d'entre eux ont fait l'objet d'une réédition numérique remastérisée[2].

Une des premières études allemandes basée systématiquement sur corpus oraux et sans doute un travail pionnier à maints égards, est le livre d'E. Gülich, *Makrosyntax der Gliederungssignale im gesprochenen Französisch* (1970). Issue de la thèse de l'auteure, cette étude a pour objet les marqueurs de discours – un sujet qui sera amplement developpé plus tard (D. Schiffrin 1987) mais qui est magistralement traité par Gülich avec deux décennies d'avance. Pour ce faire, elle a choisi un certain nombre de textes provenant du corpus du *Français fondamental* (Gougenheim & *al*. 1964), dont certains non publiés ailleurs, en y ajoutant des transcriptions propres d'émissions de radio enregistrées en Belgique francophone[3]. A la même époque, H. Stammerjohann (1970) publie un long article,

[1] L'inventaire des corpus oraux mentionnés dans cet article est forcément très sélectif. Pour une liste plus complète cf. Pusch (2002).
[2] Pour plus d'information sur ces enregistrements, cf. <www.phonogrammarchiv.uzh.ch>.
[3] D'autres corpus de taille réduite élaborés dans le cadre d'études de pragmatique et publiés en annexe à

Strukturen der Rede, dans le bulletin de l'Académie de la Crusca de Florence, qui comprend quelque 35 pages de transcription d'italien parlé enregistré dans des situations relativement informelles à Florence et ses alentours. Si Stammerjohann ne peut pas totalement échapper au marquage diatopique inévitable et omniprésent dans les corpus d'italien, langue où les dialectes jouissent d'une très grande vitalité, la variation diatopique n'est pas plus que chez Gülich l'objectif du corpus. Les transcriptions suivent un système de notation déjà très développé, combinant la transcription linéaire dominante avec des éléments de transcription en partition (cf. § 6). Or les transcriptions dans Stammerjohann (1970) ne constituaient qu'un échantillon très réduit des enregistrements effectués par l'auteur au milieu des années 1960 à Florence, qui totalisent presque 47 heures. Une partie de ces documents sonores sont actuellement en voie de numérisation et (re-)transcription par l'équipe du laboratoire LABLITA de l'Université de Florence ; ce « nouveau » corpus Stammerjohann doit être publié sous forme de livre avec cédérom (cf. Scarano & Signorini 2005 pour des détails).

3. Les années 1980 : des corpus originaux mais peu exploités…

Après les corpus oraux pionniers des années 1970, dont les enregistrements remontent à la 2de moitié des années 1960, c'est au cours des années 1980 qu'on assiste à un certain engouement pour ce genre de projet dans les études romanes en pays germanophones. Cela est peut-être dû à la publication d'un ouvrage fondamental sur l'approche théorique de l'oralité, *Gesprochenes und geschriebenes Französisch* de L. Söll (1974), dont on reparlera. Le livre de Söll traite surtout de la divergence entre langage parlé et langage écrit en français, et c'est aussi au français que se réfèrent les corpus les plus originaux publiés dans les années 1980.

Le corpus le plus varié de cette époque fructueuse, en ce qui concerne les genres textuels, les constellations discursives et les thématiques abordées, est probablement le recueil de J. Eschmann (1984), *Texte aus dem 'français parlé'*, qui inclut aussi une certaine variation diatopique avec des échantillons en français régional auvergnat et méridional. Moins varié, mais plus ambitieux était le projet de R. Meyer-Hermann, entamé à partir de 1982 avec un premier corpus consacré à l'espagnol (du Mexique) et dont le but était de doter la linguistique romane de corpus dans un grand choix de langues, y compris les moins enseignées et les moins étudiées comme, p. ex., le rhéto-romanche, le titre choisi pour cette collection étant *Bielefelder Text-corpora romanischer Sprachen*. A ce premier volume (intitulé *Spanisch II* parce que la publication d'un volume *Spanisch I* contenant de l'espagnol européen était déjà envisagé) faisaient suite un corpus de catalan, rédigé par R. Guàrdia (1985), et un volume consacré au français élaboré par G. Schmale & E. Schmale-Buton (1984). Ces trois volumes ont été publiés de façon artisanale, sous forme dactylographiée, et de manière plutôt confidentielle et n'ont donc pas connu une large diffusion. Une version imprimée commerciale n'a d'ailleurs jamais vu le jour, pas plus que les corpus ultérieurs annoncés déjà dans le volume inaugural de 1982. Aucune exploitation systématique de ces ressources ne semble avoir eu lieu. Si les corpus d'espagnol mexicain et de catalan étaient constitués d'enregistrements de la radio, relativement faciles à obtenir, le volume *Französisch I. Conversations téléphoniques* de 1984 est plus original, car il contient, comme laisse supposer son titre, des échanges privés et semi-privés par téléphone. Si les corpus issus d'enquêtes par téléphone sont aujourd'hui assez répandus, notamment en ingénierie du langage, le corpus *Französisch I* de Schmale-Buton & Schmale constituait une innovation remarquable pour son temps de production.

---

des thèses de doctorat se trouvent dans Honningfort (1993) et Stark (1997).

VERBALE UND PARAVERBALE EREIGNISSE

EV	(54)	(55)	(56)	53
A	mais attendez ⌋ ah oui oui⌋ c'est c'est⌈//	oui⌈	ah oui je vois c'est de l'aut' côté⌈	
B				
C	là c'est le⌋	le trocadéro⌋		
D				

EV	(57) (58) (59) (60) (61)	(62)	(63)
A	ah non⌈		ah oui non mais là⌋//
B			
C	voilà le trocadéro est là⌈	v'voyez⌈	après la tour eiffel⌋ et alors c'est par
D			

EV	(64) 65/66	65	(67)	66	(68)	69	(70)
A						ah parce qu'on m'avait dit⌈ euh plutôt par	
B							
C	là⌈ qu'on va près des portes de versailles⌈						
D							

EV	69 (71)	(72)	
A	la convention⌋ par là⌋		où est-ce la convention⌋
B			
C		ah la convention c'est pas aut' chose⌈	
D			

DESKRIPTION DER POTENTIELL KOMMUNIKATIVEN ODER INFORMATIVEN NONVERBALEN EREIGNISSE

54) A wendet sich wieder C zu, dreht dabei den Kopf weiter in die angegebene Richtung.
55) C macht mit der flach hochgestellten Rechten eine ruckartige Abwärtsbewegung.
56) A wendet sich ganz links um und weist mit dem Zeigefinger der Linken in eine auf der anderen Seite des Weges gelegene Richtung.
57) C blickt kurz zu B.
58) A blickt zurück zu C.
59) A zieht die ausgestreckte Linke zum Körper zurück; dadurch weist der Zeigefinger nach oben.
60) In die frühere Richtung deutend, wendet sich C wieder A zu.
61) A neigt, weiterhin C anblickend, den Kopf etwas nach links.
62) C dreht sich wieder nach rechts.
63) C macht eine wurfartige Bewegung mit dem Arm nach rechts.
64) C wendet sich erneut A zu und deutet mit der Rechten nach rechts in die Wegerichtung.
65) C holt mit der Rechten weit nach hinten aus und macht, mit den Knien ruckend, eine bogenförmige Bewegung quer nach oben in Richtung A.
66) B stellt die Koffer wieder ab und legt den Kopf (mit gequältem Gesichtsausdruck) leicht nach links.
67) C führt die Rechte zum Gesicht und reibt sich unterhalb des rechten Auges.
68) A läßt den Arm mit dem nach oben weisenden Zeigefinger sinken.
69) B hustet, hält sich dabei die Linke vor den Mund.
70) A weist mit dem Daumen der Linken leicht seitlich nach hinten.
71) C hebt mit weit geöffnetem Mund den Kopf.
72) C nickt mit dem Kopf.

Figure 1. *Extrait du corpus Scherer* (1984).

Avec le recueil d'Eschmann et la collection fondée par Meyer-Hermann, la linguistique romane allemande met à la disposition des chercheurs les premiers volumes entièrement composés de transcriptions de langage parlé. Or, pour beaucoup de temps encore, les transcriptions publiées en annexe dans des volumes dont le titre ne laisse pas nécessairement supposer qu'ils contiennent de telles données, restent la règle plutôt que l'exception. Inévitablement, ces corpus « camouflés » risquent davantage de passer inaperçus et de ne jamais être exploités en dehors des études ciblées pour lesquelles ils ont été rassemblés. C'est le cas notamment du corpus de K. Hölker (1988), inclus dans un livre au titre peu explicite *Zur Analyse von Markern* (« A propos de l'analyse de marqueurs »). L'auteur s'y intéresse aux marqueurs discursifs et notamment aux marqueurs de clôture du français. Pour étudier leur distribution et leurs fonctions dans des interactions verbales authentiques, il se sert d'enregistrements de conversations entre un médecin et ses patients, enregistrements qui ont été effectués dans le cabinet du médecin. La parole en contexte médical constitue, comme d'ailleurs la parole pathologique (aphasique ou autre), un domaine extrêmement délicat pour la linguistique sur corpus et est désormais inaccessible, pour des raisons éthiques et juridiques, dans beaucoup de pays, en tout cas en ce qui concerne des corpus de diffusion large et facile ; c'est pour ces restrictions actuelles que les données de Hölker (1988) méritent particulièrement d'être signalées.

Un corpus tout à fait exceptionnel et très en avance pour son temps est décrit et partiellement publié dans l'ouvrage de H.S. Scherer (1984), *Sprechen im situativen Kontext. Theorie und Praxis der Analyse spontanen Sprachgebrauchs*. Il s'agit là peut-être du premier corpus multimédia élaboré en linguistique romane. Pour analyser l'usage spontané de la langue (française), l'étude se base sur des transcriptions multimodales d'enregistrements d'émissions télévisées de la série *Caméra invisible* des années 1960. Les transcriptions des événements verbaux, présentées sous forme de partition (cf. chap. 6 et fig. 1), sont assorties de notes descriptives détaillées des événements non-verbaux concomitants « potentiellement communicatifs ou informatifs », comme dit l'auteur. Si on se rend compte que les techniques de notation multimodale, englobant les composantes prosodique, gestuelle et cinétique de l'événement communicatif, ont relativement peu évolué depuis et que, dans ce domaine, on est encore loin d'un standard, on ne peut que reconnaître un grand mérite à ce corpus de « caméra invisible ». Même avec le progrès technique survenu entre-temps, les corpus vidéo sont restés peu nombreux dans la linguistique romane des pays germanophones, et aucune publication d'importance n'est à signaler.

4. Les nouveaux corpus et le continuum de l'oralité

Les linguistes romanistes des pays de langue allemande étaient sensibles depuis la première moitié du XXe siècle au rôle de la variation dans la description de la langue ou des langues historiques spécifiques. Si, dans le dernier tiers du siècle, cette sensibilité au fait variationnel n'a pas mené à des études poussées de quantification variationnelle (l'école de la sociolinguistique variationnelle labovienne n'a pas trouvé d'adhérents dans la romanistique d'expression allemande), elle a impulsé des recherches fructueuses dans le domaine de la linguistique textuelle, car on attribuait une part importante du potentiel variationnel à des facteurs découlant du genre textuel dans lequel l'échange communicatif avait lieu. C'est l'« Ecole de Tübingen » notamment, avec E. Coseriu et ensuite B. Schlieben-Lange, qui a contribué au développement de cette ligne de recherche focalisée sur les traditions discursives et textuelles, approche nettement fonctionnelle et pragmatique qui, ces derniers temps, a donné naissance à une pragmatique historique néo-philologique centrée, elle aussi, sur le texte dans sa position dans les réseaux des « traditions de parler » (*Traditionen des Sprechens*), pour reprendre le terme de Schlieben-Lange (1983). Le livre de Söll (1974), déjà mentionné,

constituait une étape décisive dans ce courant d'étude ; car s'il était acquis depuis bien longtemps que la langue orale divergeait de la langue écrite – et en français, plus que dans toute autre langue romane –, la distinction « écrit » *vs* « oral » avait été traitée, en bonne tradition structuraliste, de manière binaire et assez statique ; or, l'apport nouveau de l'étude de Söll, repris et considérablement développé à partir de 1985 par P. Koch & W. Oesterreicher (1990), consistait dans l'introduction de la distinction entre une oralité / scripturalité médiale et une oralité / scripturalité conceptuelle. Si, dans cette conception, le côté du médium continuait à être structuré de façon binaire – un événement communicatif était produit ou bien dans le code graphique et donc (médialement) écrit, ou bien dans le code phonique et donc (médialement) oral –, la conception de cette production pouvait s'échelonner entre deux pôles, entre une oralité (conceptuelle) prototypique et une scripturalité (conceptuelle) prototypique, et faire appel à des traits soit du code parlé (de la norme de l'oral) soit du code écrit (la norme de l'écrit). Ce modèle permet un reclassement, plus adéquat des genres textuels ; ainsi, si une conversation libre entre amis se déroule (médialement) oralement et correspond (conceptuellement) à la norme de l'oral, un sermon ou un discours télévisé, médialement oral aussi, est beaucoup moins « spontané » et « authentique », donc plus loin d'une conception orale et plus proche d'un conception écrite ; tandis qu'une interview ou un cours universitaire occuperaient une position conceptuelle intermédiaire dans ce modèle. Pour éviter toute ambiguïté terminologique entre oralité / scripturalité dans leur acception de médium et celle de conception, Koch & Oesterreicher ont proposé de parler, au niveau conceptuel, de *Nähesprache* vs *Distanzsprache* (« langue de proximité » *vs* « langue de distance »). Leur modèle a servi de base théorique à un grand nombre de travaux pluridisciplinaires dans le cadre du centre de recherche collaborative (*Sonderforschungsbereich*) 221 de l'Université de Fribourg-en-Brisgau entre 1985 et 1996 (cf. Raible 1998 pour un résumé des travaux).

Cette nouvelle vision d'une transition sous forme de continuum entre oralité et scripturalité, entre langue de proximité et langue de distance, a eu des répercussions sur les corpus oraux produits par la romanistique allemande. En effet, Koch & Oesterreicher (1990, 26s.) exigent que la variation conceptuelle, désignée aussi comme niveau variationnel diamésique, soit prise en compte, voire reflétée dans les corpus, et soulignent que cette prise en compte du continuum de l'oralité (et de la scripturalité) constitue un critère de première importance au moment d'évaluer la représentativité du corpus. Le premier volume de transcriptions qui correspond à ce modèle et satisfait les exigences de Koch & Oesterreicher est le livre *Korpus: Texte des gesprochenen Französisch. Materialien I* de R. Ludwig (1988[4]). Une fois de plus, le corpus est de taille réduite mais d'une grande qualité technique et méthodologique : Ludwig inclut des situations communicatives sur trois niveaux de formalité, la conversation libre en famille (langage de proximité accentuée), la discussion télévisée (langage à mi-chemin entre proximité et distance) et le cours magistral universitaire (langage oral de distance). La notation que l'auteur utilise est une forme simplifiée du système HIAT (cf. § 6), qui par là fait son entrée dans les corpus romans. Cette même notation est utilisée par d'autres auteurs de corpus qui, eux, tiennent également compte du modèle de « proximité *vs* distance linguistique » de Koch & Oesterreicher : ainsi, R. Wiesmath (2006 ; les enregistrements datent de la fin des années 1990) nous livre un corpus établi selon ces principes pour le français acadien parlé dans la région de Moncton au Nouveau-Brunswick (Canada), le plus grand recueil linguistique de données orales de cette variété d'Outre-Mer actuellement disponible sous forme publiée et qui comprend des interviews libres et semi-guidées, des extraits d'émissions radiophoniques et des conférences publiques ; tandis que C. Pusch (2001) présente un corpus de confection semblable mais de taille plus réduite du parler occitan de

[4] Un second volume, annoncé implicitement par le titre de ce livre, n'a jamais vu le jour.

Gascogne, avec des conversations libres, des interviews, des discussions de radio, des extraits de conférences et des discours publics. Dans ces deux corpus, le système HIAT est appliqué de manière assez rigoureuse, et il y a un premier pas vers la multimédialité dans la présentation des données : les transcriptions ne sont plus présentées sous forme imprimée mais sur cédérom (permettant par là une recherche d'occurrences par ordinateur, limitée, il est vrai, par les contraintes du format de fichier PDF) qui inclut aussi quelques échantillons sonores.

Si les applications systématiques aux corpus oraux du modèle « langue de distance *vs* langue de proximité » de Koch & Oesterreicher sont restées peu nombreuses jusqu'ici, force est de constater que ce modèle a fortement influencé la vision de l'oralité des linguistes romanistes allemands et les a incités à utiliser les données orales avec d'autant plus de précaution.

5. La parole venue d'Outre-Mer : données créoles et autres

Depuis les années 1980, les études romanes dans les pays de langue allemande vivent une période d'intérêt grandissant pour les variétés d'Outre-Mer. Sans vouloir entamer ici une explication exhaustive de cet engouement pour les parlers lointains, on ne peut pas passer sous silence le rôle de l'espagnol qui gagne du terrain dans la romanistique (au détriment surtout du français) et qui inévitablement tend à orienter le regard des linguistes vers les Amériques, entraînant dans la foulée les études portugaises vers une perspective extra-européenne semblable. Cependant, les corpus ibéro-romans d'origine allemande, à part des volumes des *Bielefelder Text-corpora* déjà mentionnés, sont restés très limités en nombre, les linguistes spécialistes dans la discipline ayant préféré se joindre à des projets internationaux de plus d'ampleur[5]. Par contre, plusieurs ouvrages et collections de textes consacrés aux variétés françaises des Amériques et aux langues créoles – notamment celles à base lexicale française – des Caraïbes et de l'Océan Indien ont vu le jour. A part le corpus canado-acadien de Wiesmath, déjà mentionné, c'est la Louisiane romane qui a attiré l'intérêt des chercheur(e)s : ainsi, I. Neumann (1985) nous livre une étude exhaustive assortie de textes oraux (mais pas uniquement) du créole français louisianais, tandis que C. Stäbler (1995) consacre un volume entier aux transcriptions de ses données orales du français cadien, parler acadien de la Louisiane, qui lui ont servi de base empirique à sa thèse doctorale publiée la même année. Ce corpus, en format HIAT, a la particularité d'ajouter à la transcription proprement dite une traduction interlinéaire à l'allemand[6] et une notation graphique des contours prosodiques au-dessus des paroles transcrites, trait rare dans les corpus imprimés de langue orale. Pour les créoles à base lexicale française insulaires, l'Océan Indien et le Pacifique sont, pour l'instant, mieux documentés que la Caraïbe : si certains ouvrages ne contiennent que des échantillons très réduits de transcriptions de l'oral, publiés en annexe (tel Ehrhart 1993 pour le créole de Saint-Louis en Nouvelle-Calédonie, Michaelis 1994 pour le créole seychellois, et Kriegel 1996 pour le mauricien), le livre d'A. Bollée et M. Rosalie (1994), *Parol ek memwar. Récits de vie des Seychelles*, constitue le premier volume entièrement voué à la documentation d'un créole français parlé et d'accès facile grâce à la publication dans une collection bien établie – la *Kreolische Bibliothek* – chez un éditeur avec

[5] Mentionnons néanmoins les corpus de l'espagnol européen de C. Sinner, publiés sur cédérom dans Pusch & Raible (2002) et composés d'interviews sociolinguistiques, et le volume de S. Barme (2002) comprenant des textes oraux du portugais européen et brésilien.

[6] Suivant le modèle de Stäbler (1995), des traductions intercalées ont été incluses aussi dans les corpus de Pusch (2001) et Wiesmath (2006), mentionnés plus haut. Malheureusement, la traduction interlinéaire ou en parallèle avec une autre langue (comme l'anglais dans le cas de Wiesmath, cf. fig. 3), utile pour l'utilisation des corpus en dehors de la linguistique romane et notamment en linguistique générale, n'est pas proposée par beaucoup de corpus à ce jour.

une bonne diffusion. Dans cette même collection parait en 2001 un volume collectif élaboré par Ludwig, Telchid & Bruneau-Ludwig et intitulé *Corpus créole*, qui contient un bon choix de transcriptions tirées des créoles antillais (guadeloupéen, dominicain et haïtien, avec en plus des textes du créole de la Guyane) et des créoles de l'Océan Indien (mauricien et seychellois). On déplore certainement l'absence de textes réunionnais mais ce recueil a l'avantage de présenter tous les enregistrements transcrits sur deux CD audio attachés au livre, enregistrements dont la qualité sonore n'est malheureusement pas tout à fait convaincante.

Il est à noter qu'une bonne partie des corpus des variétés romanes lointaines ici mentionnés ont été élaborés par des chercheurs liés d'une manière plus ou moins directe au centre de recherche collaborative fribourgeois SFB 221 sur la relation « scripturalité *vs* oralité » dont il a été question au paragraphe précédent. Le fait que les langues créoles soient des idiomes en train de conquérir leur espace scriptural explique l'attachement de ce groupe de linguistes à ces langues de genèse (relativement) récente.

A part cette contribution sans doute très importante de la linguistique romane de provenance allemande pour la connaissance des langues et variétés romanes d'Outre-Mer, il faut aussi mentionner son apport important à la connaissance de l'usage oral d'une langue romane moins éloignée géographiquement mais néanmoins mal connue, à savoir le roumain. En effet, une équipe de recherche animée par K. Bochmann, de l'Université de Leipzig, est à l'origine des premiers corpus oraux de confection rigoureuse et soignée établis en dehors de la Roumanie. Curieusement (mais les auteurs le justifient), les premiers volumes présentent des données de communautés roumanophones situées partiellement ou entièrement en dehors de l'Etat roumain, c'est-à-dire en Moldavie (Bochmann & Dumbrava 2001) et en Ukraine (Bochmann 2004), et l'absence de traductions limite l'accessibilité de ces corpus aux non-spécialistes en romanistique balkanique.

6. Systèmes de transcription et de notation

Depuis le corpus de banlieue parisienne de D. François (1974), connu aussi sous le nom de *Corpus d'Argenteuil*, peu d'éditions de textes tirés de l'oral ont choisi l'usage d'une transcription phonétique ou phonologique, considérée - à juste titre - comme coûteuse et de lecture pénible. Les corpus d'origine allemande présentés jusqu'ici ont tous privilégié la transcription orthographique, limitant l'usage de la transcription phonétique - généralement en signes API - à quelques mots isolés où la divergence entre la prononciation attestée par rapport à la prononciation standard valait, aux yeux du transcripteur, d'être signalée. En plus, on fait un usage parfois généreux de ce que C. Blanche-Benveniste et C. Jeanjean (1985) ont appelé des « trucages orthographiques », où la prononciation divergente est indiquée non pas par des signes phonétiques conventionnels mais par une orthographe « phonologisante » (en signes alphabétiques standard) approximative.

Plus intéressante que la question des systèmes de transcription est celle qui concerne les systèmes de notation. La solution la plus facile est certainement celle d'utiliser une présentation des mouvements communicatifs transcrits comme dans une pièce de théâtre imprimée, ce qui donne une notation « stage play » ou notation linéaire, caractérisée par un retour à la ligne pour chaque changement de locuteur. Une telle notation, techniquement simple à réaliser, se prête bien à une analyse automatique moyennant logiciels de concordance et à un enrichissement par l'introduction de balises structurantes, p. ex. selon les conventions de la Text Encoding Initiative (cf. Habert 2005, 100s.). En revanche, la notation linéaire est mal adaptée à des phénomènes typiques de l'interaction orale entre plusieurs locuteurs, à savoir les chevauchements et l'intervention simultanée, phénomènes qui requièrent le recours aux crochets ou à d'autres marquages spécifiques souvent pas très clairs.

Pour remédier à cela et permettre une représentation fidèle de l'interaction entre plusieurs

interlocuteurs, le système de notation HIAT (acronyme quelque peu forcé de « Halbinterpretative Arbeitstranscription », « transcription de travail semi-interprétative »), développé dans les années 1970 par les linguistes germanistes K. Ehlich et J. Rehbein (cf. Ehlich & Rehbein 1976, Rehbein & *al*. 2004, et pour une présentation succincte et très complète, Ehlich 1993), a été introduit dans les corpus de langues romanes allemands et utilisé, avec plus ou moins de libertés, p.ex. dans Ludwig (1988), Stäbler (1995), Pusch (2001) et Wiesmath (2006). Ce système de notation attribue à chaque locuteur une ligne pour la transcription de sa production orale ; ces lignes, dont le nombre varie en fonction du nombre des intervenants, sont superposées dans ce que les auteurs appellent une « surface », qui s'apparente à une partition musicale ; d'où la désignation de ce type de notation comme « notation en partition » (*Partiturnotation*) ou « score notation » en anglais.

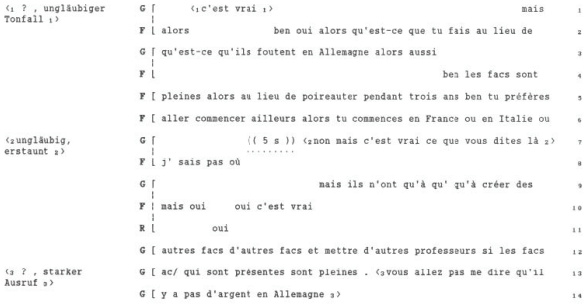

Figure 2. *Notation HIAT simplifiée* (extrait du corpus Ludwig 1988).

Dans l'espace bidimensionnel de la « surface », les interventions des locuteurs - les « événements » communicatifs - sont positionnées les unes par rapport aux autres en suivant le principe de l'iconicité temps > espace, et un axe chronométrique placé en bas de la surface, proposés dans des versions très élaborées de HIAT, permettrait même de positionner les événements transcrits en durée et en position chronologique absolues, option cependant peu utilisée et totalement absente des corpus romans établis selon ce système. Si les premiers corpus qui appliquent cette notation le font encore d'une manière sommaire et graphiquement peu ambitieuse (cf. fig. 2, extraite du corpus Ludwig 1988), en accord avec les moyens qu'offraient les logiciels de traitement de texte de l'époque, les applications ultérieures de HIAT dans les corpus d'origine allemande se caractérisent par un formalisme plus poussé, intégrant des délimitations graphiques des surfaces, des transcriptions morphologiques analytiques ou des traductions idiomatiques interlinéaires et le marquage d'éléments para-verbaux (cf. fig. 3, extraite du corpus Wiesmath 2006).

Si le système HIAT satisfait non seulement les exigences d'une reproduction fidèle des polylogues et, en même temps, certaines exigences esthétiques, force est d'admettre que ce système de notation est de facture complexe, notamment si l'on se sert d'un logiciel de traitement de texte pour créer les surfaces de transcription, et ne permet pas une analyse

assistée par ordinateur efficace car le formalisme graphique est illisible pour les concordanciers.

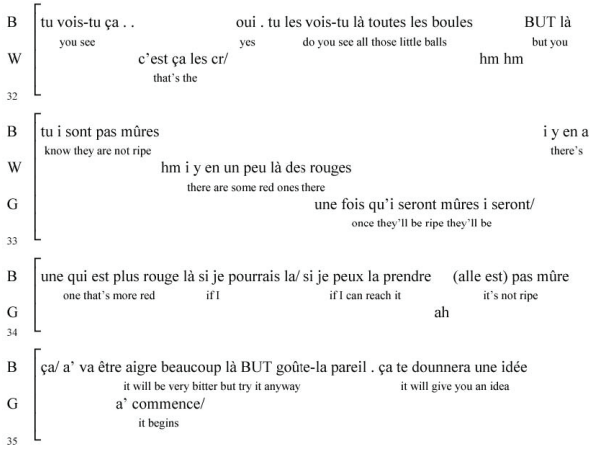

Figure 3. *Notation HIAT plus élaborée* (extrait du corpus Wiesmath 2006).

En plus, comme la linéarité verticale s'ajoute à la linéarité horizontale, la lecture des surfaces HIAT est parfois compliquée et malaisée. La tâche de générer des représentations HIAT et d'autres notations moins élaborées à partir des mêmes données a été considérablement simplifiée avec le logiciel d'édition EXMARaLDA, développé par T. Schmidt dans le cadre du centre de recherche collaborative sur le multilinguisme SFB 538 basé à Hambourg (cf. Schmidt 2002) et distribué gratuitement[7].

Le second système de notation très répandu en Allemagne, et de plus en plus utilisé en linguistique romane, s'appelle GAT, acronyme de *Gesprächsanalytisches Transkriptionssystem* (« système de transcription pour l'analyse du discours »). Egalement issu de la germanistique et décrit dans l'article de Selting & *al*. 1998 (disponible aussi sur le web dans des versions mises à jour), ce système suit le principe de la notation linéaire mais attribue à chaque tour de parole (délimité selon des critères intonatifs / prosodiques) un retour à la ligne même si c'est le même locuteur qui poursuit. Comme le système HIAT, la notation GAT prévoit plusieurs niveaux de complexité de la transcription : le niveau de base comprend la transcription séquentielle des événements communicatifs assortie d'un nombre limité de traits segmentaux et prosodiques (cf. fig. 4, issue du corpus encore inédit de Pfänder, Skrovec & Stäbler, en prép.), qui peuvent être raffinés dans une transcription fine ultérieure. Au niveau

[7] Pour des renseignements détaillés et le téléchargement du logiciel cf. <www.rrz.uni-hamburg.de /exmaralda>.

basique, GAT se contente de signes disponibles sur n'importe quel clavier d'ordinateur usuel, afin que l'introduction des transcriptions dans un logiciel d'édition ou de traitement de texte ne soit pas ralentie par l'appel de caractères ou symboles spéciaux. Dans la transcription fine, les conventions de GAT prévoient même - à l'instar de HIAT - des niveauxd'information interlinéaires pour inclure des éléments paraverbaux et explicatifs, mais, étant donné le caractère linéaire du système, de telles versions « riches » de transcriptions GAT risquent vite de devenir difficiles à lire et à comprendre.

```
290  Saskia :     <<acc> (j'ai) fait mais tu=as fait exPREs ;>
                  non mais il manquê de bon sENs -
     Caroline :   [((éclat de rires))]
     Saskia :     [c'est pas possIble]
                  je craque non je t'assure
295               je crAquê là hein , (---)
                  et c'est pas tOUt (-)
                  le PIre (.)
                  c'est que moi je=suis à effectif plein=
                  =je suis à trENte-deux élèves (-)
300               et lui le mardi il a quINze élèves (.)
                  les parents ne lui font pAs confiance,
                  il ne veulent plUs mettrê
                  les gamins en classe ,
     Caroline :   <<erstaunt> oh ,>
305  Saskia :     et: ils sont vênus me voir lundi dernier
                  en me disant=écoutEZ:
                  madame quarello ,
                  parce=que moi=c'est madame quarello hein ;
                  mAda:me=hein
310               en zep on:: -
```

Figure 4. *Notation linéaire GAT*
(extrait du corpus inédit de Pfänder, Skrovec & Stäbler, en prép.)

7. La langue de proximité en diachronie

Si au paragraphe 1 on a justifié notre de choix de considérer, dans le présent article, uniquement les corpus oraux à cause, entre autres, de la délimitation parfois épineuse, dans les données issues de l'écrit, de ce qui compte comme corpus et de ce qui ne compte pas comme tel, il vaut la peine néanmoins de jeter un coup d'œil sur deux corpus qui cherchent à documenter la langue (conceptuellement) orale, donc la langue de proximité dans des époques reculées de la langue française. Il est bien connu que certains genres textuels écrits, p. ex. la correspondance privée ou les journaux intimes, sont proches du langage informel. G. Ernst, de l'université de Ratisbonne, est à l'origine de deux corpus remarquables qui tirent profit de ces genres médialement écrits mais conceptuellement de proximité linguistique : d'une part, il a publié, sous le titre *Gesprochenes Französisch zu Beginn des 17. Jahrhunderts* (« Le français parlé au début du XVIIe siècle », 1985), le *Journal d'Héroard* (nom sous lequel ce corpus est le plus souvent cité), notes de la plume de Jean Héroard (1551-1628), médecin du jeune dauphin, futur roi Louis XIII. Ce journal livre un témoignage assez fiable de la production orale du prince, et les concordances de ces données sont publiées sur des microfiches attachées au livre. D'autre part, Ernst, en collaboration avec B. Wolf, vient de publier (2005) sous le titre *Textes français privés des XVIIe et XVIIIe siècles* un cédérom avec des transcriptions

diplomatiques de journaux intimes et de correspondance privée de personnes cultivées de différentes régions de France, autre ressource qui s'annonce précieuse pour l'étude en diachronie de la langue française de proximité.

8. Projets en cours et perspectives pour l'avenir

Etant donné l'importance grandissante accordée aux corpus dans les études de linguistique (diachroniques et synchroniques), au point que, selon C. Marchello-Nizia (2005, 14), les linguistes qui ne basent pas leurs analyses sur des données de corpus se voient désormais obligés de justifier leur choix, il n'est pas surprenant qu'il y ait un grand nombre de projets de corpus oraux (ou qui incluent des données orales) soit en cours soit dans leur stade initial. La série de colloques *Rencontre Fribourgeoise de la Linguistique de Corpus Appliquée aux Langues Romanes* qui réunissent tous les trois ans des spécialistes dans la matière en provenance des pays germanophones mais aussi de beaucoup d'autres pays, fait office de forum pour ces projets, documentés dans les actes correspondants (cf. Pusch & Raible 2002 ; Pusch, Kabatek & Raible 2005). Parmi les projets en cours, on peut mentionner la base de données BRATOLI, qui fait partie des ressources électroniques élaborées dans le cadre du centre de recherche collaborative SFB 441 *Linguistische Datenstrukturen* (« Structures de données linguistiques ») basé à l'université de Tübingen et qui comprend des transcriptions de l'espagnol et du portugais sud-américains. Dans le même contexte géographique, le projet LatinUS, réalisé sous la direction de G. Knauer à l'université Humboldt de Berlin, prévoit une présentation multimodale alignée de documents oraux et écrits qui ont rapport avec et documentent l'usage public de l'espagnol aux Etats-Unis. Un autre projet – encore à l'état d'ébauche mais dont les préparatifs vont bon train – a été lancé par S. Pfänder, W. Raible & C. Pusch sous l'acronyme CIEL_F. Son but est la constitution d'un corpus (oral et écrit) du français dans le monde ; évidemment, un tel projet « mondial », qui prend comme modèle le corpus ICE (« International Corpus of English », cf. p. ex. Greenbaum 1996) sans pour autant avoir les mêmes ambitions quantitatives, ne pourra pas rester une entreprise « allemande », et un réseau international de coopération, à l'instar de celui créé pour le projet PFC (« Phonologie du français contemporain », cf. Habert 2005, 89), est en train de se mettre en place. En relation avec ce projet CIEL_F, surtout pour permettre la mise à disposition pour le public intéressé (qui ne comprend pas seulement la communauté des linguistes) de certains corpus partiels du projet avant que celui-ci soit terminé, S. Pfänder vient de lancer une nouvelle collection de livres intitulée éloquemment *Transcriptions / Transkriptionen*. Les volumes de cette collection, dont le premier vient de sortir (Chaban, Kriegel & Pfänder 2007) et le second est dans un état de préparation avancé (Pfänder, Skrovec & Stäbler en prép.), incluront systématiquement des cédéroms qui permettront l'accès aux données sonores.

Les corpus en général et les corpus oraux, demandant des phases d'enquêtes sur le terrain, de transcription, de validation, de mise en forme et de publication sur support adéquat (qui, de nos jours, est presque obligatoirement un support électronique multimodal), sont des projets d'envergure et de « longue haleine », et la situation économique actuelle des recherches en sciences humaines et en lettres n'est pas des plus propices pour ce genre d'entreprises. Cependant, la linguistique sur corpus a prouvé qu'elle était indispensable non seulement pour la connaissance de la langue et des langues, mais aussi pour leur didactique et leur enseignement, sans oublier les applications techniques en TAL et ailleurs. Espérons qu'elle pourra poursuivre le même développement dynamique observé durant ces dernières années.

Dr. Claus Dieter Pusch
Romanisches Seminar / Albert-Ludwigs-Universität Freiburg im Breisgau
Werthmannplatz 3 - D-79085 Freiburg im Breisgau
Tel. : +49 761 203 3172 ; <claus.pusch@romanistik.uni-freiburg.de>

Références

Barme, S. (2002). *Corpus des phonisch-nähesprachlichen Brasilianisch und europäischen Portugiesisch*. Germersheim, Johannes-Gutenberg-Universität Mainz.

Baude, O. & *al*. (2005). *Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux*. Paris / Orléans, CNRS / PU Orléans.

Blanche-Benveniste, C. & Jeanjean, C. (1987). L*e français parlé. Transcription et édition*. Paris, Didier.

Bochmann, K. (2004). *Gesprochenes Rumänisch in der Ukraine. Soziolinguistische Verhältnisse und linguistische Strukturen*. Leipzig, Leipziger Universitätsverlag.

Bochmann, K. & Dumbrava, V. (eds.) (2000). *Limba Română vorbită în Moldova istorică*. Vol. 2. Texte. Leipzig, Leipziger Universitätsverlag.

Bollée, A. & Rosalie, M. (eds.) (1994). *Parol ek memwar. Récits de vie des Seychelles*. Hambourg, Buske.

Caban, M.-C., Kriegel, S. & Pfänder, S. (2007). *L'Europe de voies en voix. Témoignages franco-allemands de la migration européenne / Europa - Wege zur Vielstimmigkeit.* Berlin, BWV Berliner Wissenschafts-Verlag.

Dubois, J. & *al*. (1991). *Discourse transcription*. Londres, Longman.

Ehlich, K. (1993). HIAT: A transcription system for discourse data. In Edwards, J.A., Lampert, M. D. (eds.), *Talking data. Transcription and coding in discourse research*, Hillsdale, Erlbaum, 123-148.

Ehlich, K. & Rehbein, J. (1976). Halbinterpretative Arbeitstranskription (HIAT). *Linguistische Berichte* 45, 21-41.

Ehrhart, S. (1993). *Le créole français de St-Louis (le tayo) en Nouvelle-Calédonie*. Hambourg, Buske.

Ernst, G. (1985). *Gesprochenes Französisch zu Beginn des 17. Jahrhunderts: direkte Rede in Jean Héroards "Histoire particulière de Louis XIII" (1605-1610)*. Tübingen, Niemeyer.

Ernst, G. & Wolf, B. (eds.) (2005). *Textes français privés des XVIIe et XVIIIe siècles*. Tübingen, Niemeyer.

Eschmann, J. (1984). *Texte aus dem 'français parlé'*. Tübingen, Narr.

François, D. (1974). *Français parlé. Analyse des unités phoniques et significatives d'un corpus recueilli dans la région parisienne*. Paris, S.E.L.A.F.

Gougenheim, G. & *al*. (²1964). *L'élaboration du français fondamental (1er degré). Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris, Didier.

Greenbaum, S. (ed.) (1996). *Comparing English Worldwide. The International Corpus of English*. Oxford, Clarendon Press.

Guàrdia, R. (ed.) (1985). *Katalanisch I. Debats radiofònics "Parlem-ne" (Ràdio Quatre)*. Bielefeld, Universität / Fakultät für Linguistik und Literaturwissenschaft.

Gülich, E. (1970). *Makrosyntax der Gliederungssignale im gesprochenen Französisch*. Munich, Fink.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Gap-Paris, Ophrys.

Hölker, K. (1988). *Zur Analyse von Markern*. Stuttgart, Steiner.

Honnigfort, E. (1993). *Der segmentierte Satz. Syntaktische und pragmatische Untersuchungen zum gesprochenen Französisch der Gegenwart*. Münster, Nodus.

Koch, P. & Oesterreicher, W. (1990). *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*. Tübingen, Niemeyer.

Kriegel, S. (1996). *Diathesen im Mauritius- und Seychellenkreol*. Tübingen, Narr.

Ludwig, R. (1988). *Korpus. Texte des gesprochenen Französisch. Materialien I*. Tübingen, Narr.

Ludwig, R., Telchid, S. & Bruneau-Ludwig, F. (eds.) (2001). *Corpus créole. Textes oraux dominicains, guadeloupéens, guyanais, haïtiens, mauriciens et seychellois. Enregistrements, transcriptions et traductions*. Hambourg, Buske.

Meyer-Hermann, R. (éd. 1982). *Spanisch II*. Bielefeld, Universität / Fakultät für Linguistik und Literaturwissenschaft.

Marchello-Nizia, C. (2005). A NLP-driven approach to historical linguistics. In Pusch C., Kabatek J. & Raible W., (eds.), 11-30.

Michaelis, S. (1994). *Komplexe Syntax im Seychellen-Kreol. Verknüpfung von Sachverhaltsdarstellungen zwischen Mündlichkeit und Schriftlichkeit*. Tübingen, Narr.

Neumann, I. (1985). *Le créole de Breaux Bridge, Louisiane. Etude morphosyntaxique, textes, vocabulaire*. Hambourg, Buske.

Pfänder, S., Skrovec, M. & Stäbler, C. (en prép.). *Le français tel qu'il se parle en Provence / Gesprochenes Französisch in der Provence*. Berlin, BWV Berliner Wissenschafts-Verlag.

Pusch, C. (2001). *Morphosyntax, Informationsstruktur und Pragmatik. Präverbale Marker im gaskognischen Okzitanisch und in anderen Sprachen*. Tübingen, Narr.

Pusch, C. (2002). A survey of spoken language corpora in Romance. In Pusch, C. & Raible, W. (eds.), 245-264 (version à jour disponible à <www.corpora-romanica.net>).

Pusch, C., Kabatek, J. & Raible, W. (eds.) (2005). *Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft / Romance Corpus Linguistics II: Corpora and diachronic linguistics*. Tübingen, Narr.

Pusch, C. & Raible, W. (eds.) (2002). *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache / Romance Corpus Linguistics: Corpora and spoken language*. Tübingen, Narr.

Raible, W. (ed.) (1998) : *Medienwechsel. Erträge aus zwölf Jahren Forschung zum Thema 'Mündlichkeit und Schriftlichkeit'*. Tübingen, Narr.

Rehbein, J. & *al*. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT*. Hambourg, Universität / SFB 538.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge, Cambridge University Press.

Scarano, A. & Signorini, S. (2005). Corpus linguistics and diachronic variability. A study on Italian spoken language corpora from the 1960s until nowadays. In Pusch, C., Kabatek, J. & Raible, W. (eds.), 191-202.

Scherer, H.S. (1984). *Sprechen im situativen Kontext. Theorie und Praxis der Analyse spontanen Sprachgebrauchs*. Tübingen, Stauffenberg.

Schlieben-Lange, B. (1983). *Traditionen des Sprechens. Elemente einer pragmatischen Sprachgeschichtsschreibung*. Stuttgart, Kröner.

Schmale-Buton, E. & Schmale, G. (1984). *Französisch I. Conversations téléphoniques*. Bielefeld, Universität / Fakultät für Linguistik und Literaturwissenschaft.

Schmidt, T. (2002). *EXMARaLDA – ein System zur Diskurstranskription auf dem Computer*. Hambourg, Universität / SFB 538.

Selting, M. & *al*. (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91-122.

Sinclair, J. (2004). Corpus and text: basic principles. In Wynne, M. (ed.), *Developing Linguistic Corpora : A Guide to Good Practice*, Oxford, Oxbow Books, 1-16.

Söll, L. (1974). *Gesprochenes und geschriebenes Französisch*. Berlin, Schmidt.

Stäbler, C. (1995). *La vie dans le temps et asteur. Ein Korpus von Gesprächen mit Cadiens in Louisiana*. Tübingen, Narr.

Stammerjohann, H. (1970). Strukturen der Rede *Studi di Filologia Italiana* 28, 295-397.

Stark, E. (1997). *Voranstellungsstrukturen und 'topic'-Markierung im Französischen. Mit einem Ausblick auf das Italienische*. Tübingen, Narr.

Wiesmath, R. (2006). *Le français acadien. Analyse syntaxique d'un corpus oral recueilli au Nouveau-Brunswick, Canada*. Paris, L'Harmattan.