Claus D. Pusch, Wolfgang Raible (Freiburg im Breisgau)

Romance corpus linguistics and spoken language studies – an introduction to the present volume

One may seriously question the assumption that there has ever been such a thing as Romance linguistics without corpora. It appears indeed absurd to attempt to study the synchronic structure and the diachronic development of (a group of) individual historical languages without falling back upon 'authentic' language facts extracted from 'real life' communicational events. Scholars of Romance linguistics find themselves in a particularly favorable situation: not only are the individual Neo-Latin languages well documented throughout their period of existence, almost from the very beginning; there is also a large array of textual evidence of the ancestral language out of which the Romance tongues evolved.

However, during the last decades, the notion of 'corpus' has acquired a more specific, technical reading, and the eclectic use of citations from poetry, narrative literature or charters, or isolated examples picked up randomly from newspapers or oral conversation is not what corpus-based methodology in the modern sense is about. Nowadays, corpus linguistics means the systematic collection of language data and its organization and storage in a somehow structured and unified way, either for the purpose of mere documentation or for that of subsequent analysis, and the analysis itself.

Obviously, this approach is not really new, not even in Romance language studies which have been considered until recently as a kind of *parent pauvre* among the modern philological disciplines as far as the availability of corpus linguistic resources and the actual use of corpus linguistic methodology (in the current sense) was concerned. It is especially thanks to the ever increasing interest in spoken language – a central matter of Romance linguistics since at least the early 20th century – that linguists have abandoned relying on normative grammars and introspection and have undertaken efforts to systematically document language use and language facts as they occur naturally. If the impression prevails that corpus linguistics in the sense sketched above is a comparatively new approach, then this is certainly due to the technical improvements of recording, storage and analyzing devices during recent times.

The impression that corpus linguistics as a methodological approach to language studies is *en vogue*, is confirmed by the mushrooming of manuals, monographic and collective volumes, and specialized journals devoted to this subject. A specific textbook on Romance corpus linguistics does not (yet) exist; however, numerous introductory manuals which either focus on English or on individual Romance languages, or take a more general perspective, are currently available (McEnery / Wilson ^{2}2001 [1996]; Habert / Nazarenko / Salem

1997; Biber / Conrad / Reppen 1998; Kennedy 1998; Caravedo 1999; Hockey 2000; Spina 2001; Meyer 2002, among others). Other recently published work deals with technical questions related to corpus studies, such as transcription and (an)notation methods (Blanche-Benveniste / Jeanjean 1987; Edwards / Lampert 1993; Garside / Leech / McEnery 1997; Habert / Fabre / Isaac 1998; Van Halteren 1999) or statistical treatment of corpus data (Oakes 1998; Manning / Schütze ^{2}2000), which leads to the mass of literature available on predominantly technical issues related to applied purposes of natural language processing, language engineering, and to computer linguistics (Bunt / Nijholt 2000; Llisterri / Garrido Almiñana 2000; Pierrel 2000; Véronis 2000) or language teaching (Botley *et al*. 2000; Rossini Favretti 2000; cf. Lenz 2000 for further references). Other volumes, like this edition closer to Romance linguistics, put together papers on either corpus-methodological questions and corpus projects or on case studies on linguistic phenomena in Romance (and other) languages which rely on corpus data (AFLA 1996; Blecua *et al*. 1999; Bilger 2000 and 2001; Cresti 2000; De Kock 2001).

The present volume has a bipartite structure: The first part of papers is devoted to methodological and technical issues of corpus linguistics, centered – as the title of the volume suggests – on corpora of contemporary spoken Romance languages but slightly extending beyond this scope, with historical corpora and Germanic corpus linguistics also being touched upon. This first section is opened by Hildegard Klöden's general considerations on the challenges, achievements and limits of corpus-based studies on Romance vernaculars. Cristina Bosco and Carla Bazzanella also focus on limits and shortcomings of currently available corpus resources in Romance, their point being the encoding of contextual information in corpora, necessary to make the communicative event fully understandable and retrievable in corpus analyses. Uli Reich's article follows the same argument, focussing on the ever-recurring problem of representativity of corpora, namely in discourse-variational perspective.

These more general papers are followed by contributions which present currently on-going or recently accomplished corpus projects and developments of analysis tools for corpus processing. These papers are arranged according to languages. The first five contributions deal with French corpora. Mireille Bilger gives an overview of the corpora and corpus design principles used by the Aix-en-Provence-based research group GARS / DELIC (*Groupe Aixois de Recherche en Syntaxe*), whereas Sylvie Bruxelles and Véronique Traverso present the corpora developed by the Lyon-based GRIC (*Groupe de Recherche sur les Interactions Communicatives*), and Michel Francard, Geneviève Geron and Régine Wilmet draw up a description of the Belgian French corpora elaborated by the VALIBEL research group. Patrice Brasseur describes the fieldwork for and the lexicographic utilization of a corpus of Newfoundland French, whereas Jacques Durand, Bernard Laks and Chantal Lyche present a corpus-based project that aims at documenting the phonological peculiarities of French in different parts of the francophone world.

These papers are followed by three others which also examine French but from different aspects. Valérie Kervio-Berthou and Joseph Reisdoerfer describe corpora that are located on the borderline between French and German, the first being a parallel corpus established for the development of lexicographic resources for special purposes and the second concerning contemporary Luxemburgish and, among others, the impact exerted on it by French. The corpus presented in Sophie Prevost's and Serge Heiden's article documents Medieval French, and the authors point out the difficulties created by poorly standardized and polymorphic historical texts for corpus annotation and parsing.

The subsequent paper by Helena Britto, Marcelo Finger and Charlotte Galves also presents a diachronic corpus, in this case on Portuguese as documented in texts dating from the 16th to the 19th century. Leonel de Alencar describes a graphic interface tool which aims at improving the retrieval of information in currently available machine-readable corpora of Portuguese by facilitating the formulation of complex queries.

With the following four papers, we remain in the realm of Ibero-Romance but turn to Catalan, a language for which despite its sociolinguistic minority status an impressive number of corpus resources are under development. Núria Alturo, Emili Boix and M. Pilar Perea give an overview of the Catalan reference corpus CUB developed at the University of Barcelona. Lídia Pons and Mar Massanell demonstrate the corpus-linguistic utility of language data collected in a traditional dialectologic perspective. Jaume Corbera gives some samples of another dialectologically oriented corpus but which is being taped on video, documenting the Catalan varieties of the Balearic Islands. Oriol Camps, Lluís de Yzaguirre and Anna Matamala describe a speech-correction tool developed for enhancing the pronunciation of broadcast speakers, but emphasize the uses that could be made of this tool in research and foreign language teaching.

The sub-section on corpus projects is closed by Stefan Schneider's contribution on a database adaptation of the LIP corpus of spoken Italian and the development of a graphic query interface. Also the following article, which is in fact the synthesis of three separate contributions, deals with the user-friendly multi-media linking-up of transcriptions and audio/video data and describes the tools developed for this purpose, and applied mainly to German corpora, at the IDS (*Institut für Deutsche Sprache*, Mannheim). Claus D. Pusch's article provides bibliographical information on publicly available spoken language corpora in various Romance languages, published either in printed form or on electronic storage media.

Whereas the first part of the present volume concentrates on corpora as such and on related technical points, the second part contains 17 papers which use spoken language corpora as empirical basis for the analysis of specific linguistic phenomena. This does not necessarily preclude them from discussing – extensively, in some cases – questions of corpus design, transcription and notation principles, or the representativity of the data they are based upon. Julie Auger, for example, deals with the surfacing of epenthesis in Picard, a

langue d'oïl spoken in Northern France, and evaluates the reliability of written dialect corpora as compared to oral corpora. Carsten Sinner and Flora Klein-Andreu are both concerned with European Spanish in a variationist perspective, with Sinner discussing the importance of various parameters in corpora made up of sociolinguistic interviews and Klein-Andreu using interview and conversational data for describing pronominal change such as *leísmo* and *loísmo* phenomena. Also Konstanze Jungbluth treats Spanish pronouns – devoting herself to demonstratives – and stresses, again as Bosco / Bazzanella and Reich in part 1 of the volume, the necessity of encoding situational and other contextual information in spoken language corpora.

The two following articles deal with the expression of modality and politeness, an important and fertile area of spoken language studies: Hanne Andersen is interested in the use of adverbials such as *un peu*, *presque* and *peut-être* to avoid face-threatening acts, as documented in spoken French corpora. Cristina Bertoli Sand examines the distribution of negation raising in Italian subordinate structures on the basis of the LIP and other corpora and concludes that this is also a pragmatically driven device to avoid face-threatening acts in conversation.

The four subsequent articles all examine the discourse functions of conjunctive elements in French and their pragmatic implications: Jeanne-Marie Debaisieux and Monique Frei study the non-canonical uses of – originally causal – *parce que* on the basis of material from France and French-speaking Switzerland, respectively. José Deulofeu and Jean Véronis, on the one hand, and Raphaële Wiesmath, on the other, are concerned with non-canonical uses of the mere conjunctive *que*, but whereas Deulofeu / Véronis undertake a comparative analysis of these uses on the basis of spoken language data from different parts of France – emphasizing the discrepancy between Northern and Southern varieties –, Wiesmath compares the occurrences and the absence of conjunctive *que* in a trans-Atlantic perspective, confronting a corpus of Canadian Acadian French with other material from overseas and from European varieties of this language.

The contributions by Katja Ploog and Michael Schreiber deal with context-induced (and context-resolved) ellipsis in spoken language. Ploog compares syntactic 'structural holes' as they occur in European French spoken in France, and in the French urban vernacular of Abidjan / Ivory Coast. Schreiber, on the other hand, compares French and German under this angle, making also partly use of parallel translation corpora. Finally, Bernadette Plunkett is concerned with ellipsis in child French, with her paper focussing on the occurrence of zero-marked subjects in a corpus that documents three children from different French-speaking areas (France, Belgium and Canada) and discussing the methodological shortcomings of other acquisitional corpora of French.

The next two papers use French, Spanish and German data in order to compare styles of speaking in very specific communicational situations: Wolfgang Kesselheim and Britta Thörle investigate language use in industrial companies, whereas Isabel Zollna treats the (prosodic and other) specificities of routinized and publicly performed speech events such as joint prayers, train

announcements, etc. The final two contributions take a very different approach to oral data: While Ioana Vasilescu, Jean-Marie Hombert and François Pellegrino describe experimental research which aims at determining the factors that facilitate the recognition of spoken Romance languages by hearers unfamiliar with them, Stefan Pfänder provides us with an ethnological reading of corpus texts in French-based creoles, which had originally been collected for linguistic purposes in the francophone Caribbean and French Guyana.

The CD-ROM available with this volume contains additional material such as audio and video files, corpus extracts and analyzing tools, referring to the articles in the book. The CD-ROM also includes an additional contribution by Bernadette Plunkett and Cécile de Cat and the full version of Pusch's paper.

The vast majority of the contributions collected in this volume are revised papers originally read during the *1st Freiburg Workshop on Romance Corpus Linguistics*, which had been held at Albert-Ludwig University of Freiburg im Breisgau in early October 2000. The editors wish to express their gratitude to the German Research Council DFG (*Deutsche Forschungsgemeinschaft*, Bonn) and to the Ministry of Science of the *Land* of Baden-Württemberg (*Wissenschaftsministerium Baden-Württemberg*, Stuttgart) for their generous support which enabled us to carry out the workshop; to our publisher Gunter Narr who willingly agreed to publish this volume along with the CD-ROM in the *ScriptOralia* series; and to Susan Flocken (Freiburg i. Br.) for her linguistic advice on this and some other contributions written in English by non-native speakers.[1]

References

AFLA (ed.) 1996: *Corpus : de leur constitution à leur exploitation* (= Revue Française de Linguistique Appliquée; 1:2). Paris: Association Française de Linguistique Appliquée.

Biber, Douglas / Conrad, Susan / Reppen, Randi 1998: *Corpus linguistics. Investigating language structure and use*. Cambridge: CUP.

Bilger, Mireille (ed.) 2000: *Corpus : Méthodologie et applications linguistiques* (= Les français parlés; 3). Paris: Champion.

— (ed.) 2001: *Linguistique sur corpus : études et réflexions* (= Cahiers de l'Université de Perpignan; 31). Perpignan: Presses Universitaires de Perpignan.

Blanche-Benveniste, Claire / Jeanjean, Colette 1987: *Le français parlé. Transcription et édition*. Paris: Didier.

Blecua, José Manuel *et al.* (eds.) 1999: *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Bellaterra: Universitat Autònoma de Barcelona / Ed. Milenio.

1 A companion web page to this volume, where updates to articles may be posted, is available on the <http://www.corpora-romanica.net> web site.

Botley, Simon Philip *et al*. (eds.) 2000: *Multilingual Corpora in Teaching and Research* (= Language and Computers; 22). Amsterdam / Atlanta: Rodopi.

Bunt, Harry / Nijholt, Anton (eds.) 2000: *Advances in Probabilistic and Other Parsing Technologies* (= Text, Speech and Language Technology; 16). Dordrecht / Boston / London: Kluwer.

Caravedo, Rocío 1999: *Gramática española: enseñanza e investigación. Vol. 6: Lingüística del corpus: cuestiones teórico-metodológicas aplicadas al español*. Salamanca: Editorial de la Universidad de Salamanca.

Cresti, Emanuela 2000: *Corpus di italiano parlato. Volume I: Introduzione*. Firenze: Accademia della Crusca.

De Kock, Josse *et al*. 2001: *Gramática española: enseñanza e investigación. Vol. 7: Lingüística con corpus: catorce aplicaciones sobre el español*. Salamanca: Editorial de la Universidad de Salamanca.

Edwards, Jane A. / Lampert, Martin D. (eds.) 1993: *Talking data. Transcription and coding in discourse research*. Hillsdale: Erlbaum.

Garside, Roger / Leech, Geoffrey / McEnery, Tony (eds.) 1997: *Corpus annotation. Linguistic information from computer text corpora*. London / New York: Longman.

Habert, Benoît / Nazarenko, Adeline / Salem, André 1997: *Les linguistiques de corpus*. Paris: Armand Colin.

— / Fabre, Cécile / Isaac, Fabrice 1998: *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*. Paris: InterEditions / Masson.

Hockey, Susan 2000: *Electronic Texts in the Humanities*. Oxford: OUP.

Kennedy, Graeme D. 1998: *An Introduction to Corpus Linguistics*. London / New York: Longman.

Lenz, Susanne 2000: *Korpuslinguistik* (= Studienbibliographien Sprachwissenschaft; 32). Tübingen: Groos.

Llisterri, Joaquim / Garrido Almiñana, Juan M. 2000: La ingeniería lingüística en España. Madrid: Instituto Cervantes (<http://cvc.cervantes.es/obref/anuario/parte2/cap3/indice.htm>).

MacEnery, Tony / Wilson, Andrew ²2001: *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Manning, Christopher D. / Schütze, Hinrich ²2000: *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Meyer, Charles F. 2002: *English Corpus Linguistics. An Introduction* (= Studies in English Language). Cambridge: CUP.

Oakes, Michael P. 1998: *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Pierrel, Jean-Marie (ed.) 2000: *Ingénierie des langues*. Paris / Oxford: Hermes.

Rossini Favretti, Rema (ed.) 2000: *Linguistica e informatica.* Corpora*, multimedialità e percorsi di apprendimento*. Rome: Bulzoni.

Spina, Stefania 2001: *Fare i conti con le parole. Introduzione alla linguistica dei corpora*. Perugia: Guerra.

Van Halteren, Hans (ed.) 1999: *Syntactic Wordclass Tagging* (= Text, Speech and Language Technology; 9). Dordrecht / Boston / London: Kluwer.

Véronis, Jean (ed.) 2000: *Parallel Text Processing. Alignment and use of translation corpora* (= Text, Speech and Language Technology; 13). Dordrecht / Boston / London: Kluwer.