

Claus D. Pusch (Freiburg im Breisgau)

A survey of spoken language corpora in Romance

Der Beitrag liefert einen Überblick über sprechsprachliche Korpora in romanischen Sprachen, die in publizierter Form – gedruckt, auf elektronischen Speichermedien oder durch das WWW des Internet – verfügbar sind, und versteht sich als Ergänzung und Erweiterung des von Koch / Oesterreicher (1990) vorgelegten Korpusinventars. Neben den bibliographischen Angaben enthalten die Einträge eine Kurzcharakterisierung der Korpora nach Parametern wie Umfang und Zahl der Teilkorpora, diasystematische Markierung oder Transkriptions- und Notationsprinzipien.

1. Preliminary remarks

Although corpus linguistics in the current sense of the term is a rather recent methodological approach in Romance linguistic research, the use of authentic language data, whether elicited or recorded in naturally occurring communication, is not at all new in our discipline.¹ Romance linguists already have at their disposal a quite impressive corpus of corpora, especially since during the last decades the interest in the specific features and peculiarities of spoken (as opposed to written) language has gained tremendous importance in Romance linguistics as a whole and in the sub-disciplines dedicated to the individual Romance languages.

If the general – and erroneous – impression prevails that there are few spoken language resources in Romance and that sources for authentic oral speech are insufficient, then this is due to the fact that many corpora either remain unpublished or otherwise unavailable. Many known and quoted corpora are in reality inaccessible or at least difficult to access because their authors are unable or unwilling to distribute them in printed form or on other kinds of media. Privacy and copyright restrictions are often invoked as obstacles for the publication of spoken language corpora, but in other cases, corpora are not published as a whole because the data has only been transcribed fragmentarily or no funding was available for a decent publication of the entire data in printed form, a task that even for small- and mid-size corpora means huge books with limited economic interest for the publishers. Alternative ways of publishing, e.g. on CD-ROM or via Internet, are certainly a promising solution as far as these economic restrictions are concerned, but do not solve the more fundamental problems of corpus constitution and availability nor questions of ethics related to this.

The second reason for the unjustified, yet frequently repeated claim that in Romance linguistics only poor spoken language resources were available, is that many published corpora are virtually unknown. For a long time the publication of spoken language corpora, unlike editions of historically significant texts – i.e., written language corpora –, has not been acknowledged as a scientific achievement of its own right and therefore these texts have been relegated to the status of appendixes or confidentially published as non-commercial ‘gray literature.’ Both forms of publishing make the retrieval of the data through bibliographical information channels difficult, if not impossible. Thus, a considerable amount of spoken language corpora remain under-used or even totally unexploited, which is a highly regrettable state of affairs in view of the presumed scarcity of spoken language resources that Romance linguists use to complain about, and even more regrettable when one considers the extremely time-consuming efforts of the corpus editors.

The scope of this article is to contribute to remedy this situation by surveying some currently available spoken language corpora in different Romance languages. Lists of corpora already exist for a couple of languages (cf. Blanche-Benveniste / Jeanjean 1986: 201-209, Boisvert / Laurendeau 1988, Stein 1995 for French; Bilger 1996, Bosco / Bazzanella [on the CD-ROM to this volume], Cresti 2000: 13-15, Hazaël-Massieux 1996, Marcuschi 2000: 51-57 and Voghera 1996 for other Romance languages), but most of these overviews suffer from

1 I would like to express my gratitude to Dinah Callou (Rio de Janeiro), Maria Elias Soares (Fortaleza), Susana Martorell de Laconi (Salta), Amparo Morales (San Juan / Puerto Rico), Johannes Kabatek (Freiburg i. Br.) and Jacyra Mota (Salvador) for providing me with review copies of several Ibero-Romance corpora; to Stefan Barne (Trier), Klaus Bochmann (Leipzig), Elisabetta Bonvino (Rome), Wolfgang Dahmen (Jena) and Rudolf Windisch (Rostock) for their advice and valuable information; and to Susan Flocken (Freiburg i. Br.) for revising this paper and correcting the insufficiencies of my English.

some shortcomings of one or the other kind.² An important list of published (printed) corpora in three Romance languages – French, Spanish, and Italian – is found in Koch / Oesterreicher (1990: 25-49). Their principle of presenting the corpora by applying a unified list of descriptive parameters will be adopted in this contribution, although in a somehow simplified form.³

For the sake of space, a number of restrictions had to be applied when choosing the corpora to be listed. Only *published* corpora were included which are generally *available in university libraries* or by inter-university book exchange, or – in the case of electronically available corpora – via Internet. Therefore, ‘gray’ literature was included when it proved to be referenced in university catalogues and fairly easy to borrow. On the other hand, this requirement put limits on the inclusion of corpora published in PhD dissertations and comparable works, which are unproblematic to track down when published, e.g., in the US or Germany but far less accessible when written in France or Spain.⁴ Furthermore, only such corpora were taken into consideration that were established for *linguistic purposes* or where the authors mentioned a specific interest in language facts, even if this interest was not primarily linguistically motivated, but rather ethno-linguistic, folkloristic or anthropological, as long as evidence was available that the transcriptions were faithful and not polished up. Non-spontaneous forms of oral literature such as poems and songs were excluded, as well as pieces of oral history published for non-linguistic purposes.

The treatment of *dialectologic corpora* has proved very problematic: these texts, which have been collected since the very beginning of dialectology and therefore constitute, in many cases, the earliest ‘spoken’ language corpora available, sometimes meet the criteria of ‘modern’ oral corpora fairly well. Frequently, however, they contain exclusively semi-spontaneous forms of orality, such as story-telling of the *contes et légendes* type, and they totally lack dialogical features. Dialect corpora have therefore been included only with much caution. They were more readily accepted when a dialogical character of the texts was obvious, whenever the recordings were available with the transcriptions, and also when other corpora, more akin to what corpus linguists without a particular dialectologic interest consider to be a ‘spoken language corpus,’ were scarce. Dialect corpora were also included in the case of languages where dialectology continues to be a central focus of linguistic research, such as in the case of Italian or Catalan. They have been discarded, however, when the text samples were very short.⁵ Needless to emphasize that it is sometimes hard to make a clear distinction between ‘dialect’ corpora and ‘general’ corpora which exhibit diatopically marked variation.⁶

Acquisitional corpora, documenting the L1 learning stages of small children, have been completely left aside in the present article (but see Plunkett [in this volume] for some information on French child corpora).

To qualify for the survey, corpora had to be organized in coherent units larger than sentences or paragraphs, i.e. they had to look like *texts*, for which, in the case of printed corpora, a *minimum length* of ten – generally consecutive – printed pages was required. Therefore, oral data quoted in the body of a linguistic study and lists of isolated sentences were discarded although the data sometimes met the aforementioned quantitative requirement. And – last but not least – the language data had to be available in *transcribed form*, so that oral corpora that are published e.g. only as audio CDs or tapes but without a transcription volume or booklet were not considered, even if a linguistic perspective was beyond doubt.

The following list does by no means claim to be exhaustive. Corrections and addenda are welcome and will be published on the companion web site to this volume.⁷

2 For example, Blanche-Benveniste / Jeanjean (1986) and Boisvert / Laurendeau (1988) list mostly unpublished or only locally accessible corpora and even corpus projects that were never entirely carried out.

3 Koch / Oesterreicher’s list goes beyond the scope of this article as it also includes non-linguistically-oriented sources of spoken language data, namely for Italian.

4 It is for this reason that, e.g., the fascinating urban Acadian French corpus by Marie-Eve Perrot, quoted in the paper of Wiesmath (in this volume), has not been included in this survey.

5 This explains why neither the *Profilo dei dialetti italiani* series (Pisa: Pacini, 1974-1988) – despite the fact that the volumes come with a vinyl record including the recordings – nor the *Textos andaluces en transcripción fonética* published by M. Alvar and P. García Mouton (Madrid: Gredos, 1995) are found in this survey. Also the Gascony Occitan texts published by J. Wüest and A. Kristol (*Aqueras montanhas. Etudes de linguistique occitane : Le Couserans (Gascogne pyrénéenne)*, Tübingen / Basel: Francke, 1993) and which I have described elsewhere (cf. Pusch 2001: 81) are not included due to the shortness of the transcribed samples.

6 For example, the Spanish *Habla culta* corpora, presented in section 5, despite their express focus on the speech of educated speakers, exhibit features of variation that could classify as ‘dialectologic.’

7 Cf. <<http://www.corpora-romanica.net>>.

2. Descriptive parameters

Corpora in the following list are ordered according to languages and, within the languages, alphabetically. The entries start with a complete set of bibliographical data, including – wherever available – the ISBN (International Standard Book Number). This is followed by 6 descriptive parameters which, for the sake of brevity, are not repeated with each entry but have simply been numbered. They concern:

1 the type of media the corpus is published on (printed, electronic-offline, i.e. CD-ROM, floppy disk etc., or electronic-online, i.e. via Internet),⁸ the size of the corpus and the number of texts that the corpus is made up of. The corpus size, if not quantified by the corpus editors, has been calculated by counting the words of two selected pages and by extrapolating from this result, which obviously is not a very reliable, but easily applicable method as long as texts are homogeneous. Otherwise, I abstained from giving a number of words count;

2 the availability of meta-information, i.e. whether a biographic and sociolinguistic profile of the informants is provided, or whether hints on the area (region, town) where the recordings had been made, or the date when the field work was carried out, are included;

3 the diasystematic features of the corpus text in the sense of Flydal / Coseriu (cf. Coseriu 1988), i.e. whether they display specific diatopical (“dialects”), diastratical (“sociolects”) or diaphatic (styles and registers) features;

4 the transcription and notation systems, i.e. if the transcription is phonetic / phonological, orthographic or mixed and if, for example, turns are ordered on subsequent individual transcription lines (linear notation) or if they are ordered in a more complex multi-level frame with time-iconic vertical alignment, as suggested by the HIAT scheme (Ehlich 1993) which is particularly popular in German corpora. This parameter field also includes information on further categories that the corpus is tagged for, and whether this is done by in-line or interlinear tags;


5 the availability of an (interlinear, parallel or separate) translation into another language; the significance of this parameter should not be under-estimated as soon as non-standard varieties and “lesser used” languages are concerned; and, finally,

6 the accessibility or availability of the recordings (i.e., in almost all cases, the audio recordings) that the transcription is based on.

An evaluating comment or some remarks on specific features of the corpus under review may close the entry. For the sake of brevity, corpora already described in Koch / Oesterreicher’s list are only referenced bibliographically, but not presented in detail; readers are referred to Koch / Oesterreicher (1990) instead.


3. Corpora of spoken French

3.1. European French





 Beeching, Kate s.a.: *Un corpus d’entretiens spontanés*. Bristol: University of the West of England. Downloadable at <<http://www.uwe.ac.uk/facults/les/staff/kb/main.html>>.

1 Electronic corpus (PDF file). Approx. 150.000 words. 95 texts of variable length. **2** Basic meta-information available. All texts have been recorded in Northern (Brittany, Paris) and Southern France (Minervois, Lot) between 1980 and 1990. **3** Some texts slightly marked diatopically or diastratically. Most texts are (semi-guided or unguided) interviews. **4** Orthographic transcription. **5** No translation available. **6** Recordings available on request from the author.

Comment: This varied corpus contains the data analyzed in Beeching (to appear).

 Biggs, Patricia *et al.* 1973: *Enquête socio-linguistique sur Orléans. Les transcriptions du corpus d’Orléans*. Colchester: University of Essex. No ISBN; typescript.

1 Printed corpus. 3.820 pages, approx. 500.000 words. 85 texts available in separate fascicles.⁹ **2** Detailed meta-information available. All recordings made in 1969 in and around Orléans. **3** Only interviews. Almost no dia-

8 This is also marked through a symbol at the beginning of the bibliographical reference, with  meaning ‘published in printed form,’  ‘published on off-line electronic media (floppy disk, CD-ROM) or downloadable via Internet,’  ‘available on-line only, without download option,’ and  indicating ‘published on audio CD, cassette or tape’.

9 Referenced fascicles are numbered 1 to 169 but with gaps. I was unable to find out with certainty if the fascicles with the missing numbers have been published.

topic but some diastatic features. **4** Orthographic transcription in a simplified score notation. **5** No translation available. **6** Audio tapes are said to have been distributed with the fascicles but this could not be confirmed.

Comment: This is the completed part of the huge *Corpus d'Orléans* (315 hours, 4,5 Mio words; cf. Blanc / Biggs 1971; Blanche-Benveniste / Jeanjean 1987: 206; Bergounioux / Baraduc / Dumont 1992 and Durand / Laks / Lyche [in this volume; chap. 2]) with 42 hours of recordings transcribed. The typescript edition is accessible in few university libraries only, but there exist more readily accessible versions (see the subsequent Biggs / Dalwood [1976] corpus and the ELILAP [De Kock *et al.* 1980-] corpus).

📖 Biggs, Patricia / Dalwood, Mary 1976: *Les Orléanais ont la parole. Teaching guide and tapescript*. London: Longman. ISBN 0-582-33122-6.

Comment: See Koch / Oesterreicher (1990: 31s). Small portion of the *Corpus d'Orléans* published within a language-teaching manual. With audio cassette containing the recordings.

📖 Blanche-Benveniste, Claire / Rouget, Christine / Sabio, Frédéric (eds.) 2002: *Choix de textes de français parlé. 36 extraits* (= Les français parlés; 5). Paris: Champion. ISBN 2-7453-0553-0.

Comment: This corpus was about to be published when the present article went into print. Therefore, no detailed description can be given at this time.

📖 Cosnier, Jacques / Kerbrat-Orecchioni, Cathérine (eds.) 1987: *Décrire la conversation*. Lyon: Presses Universitaires de Lyon. ISBN 2-7297-0307-1. Corpus: 377-390.

Comment: See Koch / Oesterreicher (1990: 32) and Bruxelles / Traverso (in this volume).

📖 De Kock, Josse *et al.* (eds.) 1980-: ELILAP – *Etude linguistique de la langue parlée*. Leuven: Leuven University / Linguistics Department. Accessible at <<http://bach.arts.kuleuven.ac.be/elicop/>>.

1 On-line corpus. Approx. 1,1 Mio words and 800 texts. **2** Detailed meta-information available. **3** Marked on all diastatic levels, but to a different degree depending on the text. Also many generally unmarked texts. **4** Orthographic transcription; a small proportion has also been transcribed phonetically. **5** No translation available. **6** No recordings available.

Comment: This highly useful on-line corpus is made up of three sub-corpora, all of them recorded by British linguists between 1968 and 1976 for language-teaching purposes: an enlarged portion of the *Corpus d'Orléans* (see further up, Biggs *et al.* [1973] corpus; ELILAP contains 80 hours / 904.000 words), a part of the *Livre parlé de Tours* corpus (4 hours / 37.000 words) and the *Voix d'Auvergne* corpus (recorded in Clermont-Ferrand and surroundings; 17 hours / 187.000 words). See ELILAP's web-site for a detailed description of the sub-corpora.

The web interface allows simple and advanced searches for word frequencies and concordance tables. However and with only a few exceptions, the interface neither allows access to the corpus texts as such nor includes a download option for offline analysis.

📖 Eschmann, Jürgen 1984: *Texte aus dem 'français parlé'* (= Tübinger Beiträge zur Linguistik; 257). Tübingen: Narr. ISBN 3-87808-857-4.

Comment: See Koch / Oesterreicher (1990: 32s).

The following four corpora are presented together due to their common conception and identical presentation:

📖 Fossat, Jean-Louis / Valière, Michel 1977: *Histoire de la vie rurale en Poitou. Récits d'un étalonnier*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 95 pages, approx. 28.000 words. 1 text. **2** Detailed meta-information available. Entire recording made with only one and the same informant, a 60 year old owner of breeding stallions living in Magne (Dép. Vienne). **3** Diastatic and moderate diatopic features.

📖 Icart-Séguy, Hélène 1976: *Dialogues de femmes*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 90 pages, approx. 25.000 words. 5 texts. **2** Only basic meta-information available. Recordings document elicited conversation in a feminists' discussion group. **3** Generally unmarked speech.

📖 Lafage, E. 1976: *Le français des enfants de l'Isle-en-Dodon (31)*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 80 pages, approx. 30.000 words. 1 text. **2** Complete meta-information available. Recording made in late 1975 with 12-15 year old school-children in L'Isle-en-Dodon (Dép. Haute-Garonne). **3** Slightly marked on diatopic level.

📖 Rivenc-Chiclet, Marie-Madeleine 1976: *Eleveurs et négociants de la Haute-Saône. Dialogues de transaction*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 87 pages, approx. 25.000 words. 3 texts. 2 Complete meta-information available. Cattle dealers' negotiation talks, recorded in November 1972 in the Vesoul (Dép. Haute-Saône) region. 3 Clearly diatopically marked, with moderate diatopic features.

The following characteristics are shared by all of these four corpora:

4 Orthographic transcription with slight phonetic adjustments. 5 No translation available, but many lexical explanations given in footnotes. 6 No recordings available.

Comment: The French part of the numerous lexicographically and socio-linguistically oriented – and little known – corpora published in the 70's by the *Institut d'Etudes Méridionales* conducted by J.-L. Fossat (the other part of the series consists of Occitan corpora; see section 4.2). Another corpus of the same collection, but which was not available for review for this survey, is the following:

📖 Desprats, Béatrice 1976: *Le vocabulaire technique de la mégisserie à Graulhet. Textes des travailleurs des métiers du cuir*. Toulouse: Université Toulouse II-Le Mirail. No ISBN.

📖 François, Denise 1974: *Français parlé. Analyse des unités phoniques et significatives d'un corpus recueilli dans la région parisienne*. Paris: S.E.L.A.F. No ISBN. Corpus: 761-838.

Comment: See Koch / Oesterreicher (1990: 33s). Frequently referred to as *Corpus d'Argenteuil*.

📖 Gougenheim, Georges et al. 1964: *L'élaboration du français fondamental (1er degré). Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier. No ISBN. Corpus: 239-253.

1 Printed corpus. 14 pages. 9 texts. 2 Only very basic meta-information available. Recordings made in 1952/53 in Paris and the Savoie region. 3 Generally unmarked on all diasystematic levels. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

Comment: See Gülich (1970) and Zeidler (1982) for details on this corpus.

📖 Gülich, Elisabeth 1970: *Makrosyntax der Gliederungssignale im gesprochenen Französisch (= Structura; 2)*. München: Fink. No ISBN. Corpus: 20-24, 35-37, 57-62, 68-70, 141s, A1-A77.

1 Printed corpus. 84 pages, approx. 45.000 words. 13 texts. 2 Only very basic meta-information available. Some recordings made in the early 1950's in Paris and Savoie (France), while others were made in 1966 in Belgium. 3 Different types of texts, but most of them almost monological. Belgian texts only slightly diatopically marked. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

Comment: This somehow heterogeneous corpus consists partly of recordings made for the "Français fondamental" project (cf. Gougenheim et al. [1964] corpus). Belgian texts are radio recordings made by Gülich.

📖 Hölker, Klaus 1988: *Zur Analyse von Markern (= Beihefte zur ZFSL, N.F.; 15)*. Stuttgart: Steiner. ISBN 3-515-05246-1. Corpus: 183-307.

Comment: See Koch / Oesterreicher (1990: 34). Corpus contains exclusively doctor-patient interaction.

📖 Honnigfort, Eva 1993: *Der segmentierte Satz. Syntaktische und pragmatische Untersuchungen zum gesprochenen Französisch der Gegenwart (= Münstersche Beiträge zur Romanischen Philologie; 8)*. Münster: Nodus. ISBN 3-89323-558-2. Corpus: 271-355.

1 Printed corpus. 84 pages. 9 texts. 2 Detailed meta-information available. Recordings made between 1979 and 1987 in the French-speaking part of Belgium. 3 Moderately diatopically marked. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

📖📻 Lindqvist, Christina 2001: *Corpus transcrit de quelques journaux télévisés français*. Uppsala: Uppsala Universitet. ISBN 91-506-1532-7.

1 Printed corpus. 258 pages. 19 texts. 2 Basic meta-information available. All texts are transcribed from videotaped television news programs recorded in 1993. Speakers are well-known broadcast professionals. 3 Diasystematically neuter. 4 Orthographic transcription in elaborated linear notation, including time-line, some prosodic features and annotations on three subjects that the author of the corpus was particularly interested in (phonetic realization of schwa, *liaison*, and negation). 5 No translation available. 6 Full recordings available as MP3 files on a CD-ROM enclosed with the volume.

📖 Ludwig, Ralph 1988: *Korpus. Texte des gesprochenen Französisch. Materialien I (= ScriptOralia; 8)*. Tübingen: Narr. ISBN 3-87808-691-1.

Comment: See Koch / Oesterreicher (1990: 34s). One of the first Romance corpora that present different oral text types ranging from casual to formal oral speech according to the model of orality developed by Koch / Oesterreicher (1985).

📖 Marconot, Jean-Marie 1985: *L'analyse de la conversation. Le livre de Vauvert*. S.I. [Mussidan]: FEDEROP / MARPOC. ISBN 2-85792-042-3.

1 Printed corpus. 92 pages. 96 texts, some of them very short. 2 Basic meta-information available. Recordings made in Vauvert (near Nîmes) between 1976 and 1980. 3 Diatopically and diastratically marked. One text (13 pages) is in Occitan (*languedocien*). 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

📖 Martins-Baltar, Michel *et al.* (eds.) 1989: *Entretiens. Transcription d'un corpus oral* (= Cahiers du Français des Années Quatre-vingts, Hors Série; 1). Saint-Cloud / Paris: E.N.S. de Fontenay-St Cloud – CREDIF / Didier. ISBN 2-905769-15-7.

1 Printed corpus. 260 pages, approx. 270.000 words. 25 texts. 2 For most texts, detailed meta-information is given. Recordings made in different places throughout France in the early 1980's. 3 All texts are interviews. Partly moderate diatopic features. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

📖 Scherer, Hans-Siegfried 1984: *Sprechen im situativen Kontext. Theorie und Praxis der Analyse spontanen Sprachgebrauchs* (= Romanica et Comparatistica; 3). Tübingen: Stauffenberg. ISBN 3-923721-53-6. Corpus: 202-377.

Comment: See Koch / Oesterreicher (1990: 35). A remarkable corpus, based on “Candid Camera”-style video recordings, that encodes systematically para-linguistic activities of the speakers.

📖 Schmale-Buton, Elisabeth / Schmale, Günter 1984: *Französisch I. Conversations téléphoniques* (= Bielefelder Text-Corpora romanischer Sprachen; 1). Bielefeld: Universität Bielefeld / Fakultät für Linguistik und Literaturwissenschaft. No ISBN; typescript.

1 Printed corpus. 314 pages, approx. 50.000 words. 71 texts. 2 Detailed meta-information available. Recordings made at different places of France in the early 1980's. 3 Slightly marked diastratically, but almost unmarked as to diatopic features. 4 Orthographic transcription in a simplified score notation. Intonation patterns marked in-line. 5 No translation available. 6 No recordings available.

Comment: This interesting corpus comprises telephone calls on three different levels of conceptual formality (“communications privées”, “communication professionnelles”, “communications institutionnelles”).

📖 Stark, Elisabeth 1997: *Vorstellungsstrukturen und „topic“-Markierung im Französischen. Mit einem Ausblick auf das Italienische* (= Romanica Monacensia; 51). Tübingen: Narr. ISBN 3-8233-4791-8. Corpus: 308-356.

1 Printed corpus. 46 pages, 14.000 words. 9 texts. 2 Detailed meta-information available. All recordings (except for two, which were video-taped TV discussions) made in 1995 with urban upper middle-class informants in Paris. 3 Generally unmarked speech. 4 Orthographic transcription in linear notation. 5 No translation available. Passages skipped in the transcription are summarized in German. 6 Recordings may be available on request from the author.

📖 Stempel, Wolf-Dieter 1987: Die Alltagserzählung als Kunst-Stück; in: Erzgräber, Willy / Goetsch, Paul (eds.): *Mündliches Erzählen im Alltag, fingiertes mündliches Erzählen in der Literatur* (= ScriptOra; 1). Tübingen: Narr, 105-135. ISBN 3-87808-741-1. Corpus: 122-135.

Comment: See Koch / Oesterreicher (1990: 34).

📖 Waugh, Linda / Lawson, Aaron s.a.: *French corpus*. S.I.: Cornell University. Accessible and downloadable at <http://www.people.cornell.edu/pages/adl6/french_corpus.htm> (HTML files).

1 Electronic corpus. Approx. 119.000 words. 27 texts. 2 Momentarily almost no meta-information available. Transcriptions include several spontaneous dialogues among youngsters and students, and some university lectures. 3 Generally unmarked speech. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

3.2. Overseas French

📖 Beauchemin, Normand / Martel, Pierre / Théorêt, Michel 1973-81: *Echantillon de textes libres*. Sherbrooke: Université de Sherbrooke. No ISBN.

Comment: See Koch / Oesterreicher (1990: 33), Blanche-Benveniste / Jeanjean (1987: 202), Boisvert / Laurendeau (1988: 247-49). 270.000 words corpus of Québec French, also referred to as *Sherbrooke Corpus* or *Corpus*

de l'Estrée. A short sample in electronic form (ASCII file) is downloadable from the *Oxford Text Archive* website; cf. <<http://ota.ahds.ac.uk>>. An electronic version of the entire corpus may be obtained on request from the authors.

📖 Highfield, Arnold R. 1979: *The French dialect of St. Thomas, U.S. Virgin Islands. A descriptive grammar with texts and glossary*. Ann Arbor: Karoma. ISBN 0-89720-026-8. Corpus: 148-227.

1 Printed corpus. 78 pages, approx. 22.000 words. 36 texts. 2 Basic meta-information available in the descriptive part of the book. Fieldwork carried out in the French-speaking community of Carenage (St. Thomas) in 1974/75. 3 Strongly marked on the diatopic level. 4 Phonetically-oriented transcription, using a system originally developed for transcribing Haitian Creole and not conform to IPA. Linear notation, with many texts being monologue-like. 5 No translation available, but the entire vocabulary is explained in a French-English glossary that follows the corpus. 6 No recordings available.

Comment: A very valuable corpus of the French patois of the Caribbean island of St. Barthélemy, as spoken by St. Bartian emigrants.

📖 Markey, Patricia Ann Pease 1977: *Tahitian French. A study in tense and aspect*. S.l.: University of Michigan. PhD dissertation, DAI no. 7804764. Corpus: 162-191.

1 Printed corpus. 28 pages, approx. 5.500 words. 8 texts, some very short. 2 Only rudimentary meta-information available. 3 Although intended to be a corpus of Tahitian regional French, the data is only moderately diatopically marked. Most texts are interviews. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

🎧📖 Maury, Nicole / Tessier, Jules 1991: *A l'écoute des francophones d'Amérique. Textes des enregistrements*. Montréal: Centre Educatif et Culturel. ISBN 2-7617-0956-X.

1 Audio corpus with printed transcription booklet. 22 pages. 23 texts. 2 Basic meta-information available. 3 Strongly marked on the diatopic level, as the documentation of the different varieties of North American French (for language teaching purposes) is the main goal of this corpus. 4 Orthographic transcription in linear notation. 5 No translation available. 6 Recordings available on an audio-cassette, to which the transcription booklet actually constitutes a mere annex.

Comment: Short audio samples from different French-speaking communities in Canada and the U.S.

📖 Smith, Jane S. 1994: *A morphosyntactic analysis of the verb group in Cajun French*. S.l.: University of Washington. PhD dissertation, DAI no. 9434353. Corpus: 170-232.

1 Printed corpus. 62 pages, approx. 22.000 words. 3 texts. 2 Only rudimentary meta-information available. Elicited conversation between students. 3 Diatopically marked (Acadian French as spoken in Louisiana), with interferences of English. 4 Orthographic transcription with slight adjustments, in linear notation. 5 No translation available. 6 No recordings available.

📖 Stähler, Cynthia K. 1995: *La vie dans le temps et aeteur. Ein Korpus von Gesprächen mit Cadiens in Louisiana* (= ScriptOra; 79). Tübingen: Narr. ISBN 3-8233-4569-9.

1 Printed corpus. 245 pages. 10 texts. 2 Complete meta-information available. Recordings made in Southwestern Louisiana in 1988/89. 3 Diatopically marked (Acadian French as spoken in Louisiana), with moderate interferences from English. 4 Orthographic transcription with very few phonetic adjustments. Simplified HIAT-oriented score notation. Intonation patterns marked in-line. 5 Interlinear translation into German. 6 Recordings may be requested from the author.

📖 Thomas, Gerald 1983: *Les deux traditions. Le conte populaire chez les Franco-Terreneuviens*. Montréal: Bellarmin. ISBN 2-89007-519-2. Corpus: 201-442.

1 Printed corpus. 59 pages, approx. 35.000 words. 14 French texts (plus 1 mixed English-French text and 18 all-English texts; these have not been taken into account here). 2 Detailed meta-information available. All recordings made between 1972 and 1981 with 3 informants from French-speaking communities in Western Newfoundland. 3 Strongly marked on the diatopic (Newfoundland Acadian and Laurentidian French) and on the diaphatic level (narrative texts of the *conte* type). 4 Orthographic transcription with phonetic adjustments in linear transcription; many texts are monologues. 5 No translation available (only English texts are translated into French). 6 No recordings available.

Comment: Linguistically usable corpus with an ethnographic-folkloristic scope.

4. Corpora of other Gallo-Romance languages

4.1. Franco-Provençal

📖 Charpigny, Florence / Grenouiller, Anne-Marie / Martin, Jean-Baptiste 1986: *Marius Champailler, paysan de Pélussin*. Aix-en-Provence: Edisud / Ed. du CNRS. ISBN 2-85744-237-8. Corpus: 30-231.

1 Printed corpus. 81 pages, approx. 40.000 words. 35 texts, arranged in 6 thematic groups. 2 Detailed meta-information available. All recordings made with one informant, an 80 year old farmer and viticulturist from Pélussin (Dép. Loire). 3 Diatopically and diaphatically marked. All texts are narrative in character, some being of the *contes et légendes* type, but most are auto-biographic. 4 Hybrid phonetically-oriented transcription, not conform to IPA. 5 Parallel translation into French. 6 No recordings available.¹⁰

4.2. Occitan

The following three corpora are presented together due to their common conception and identical presentation. They constitute the Occitan part of the corpora collection published in the 1970's by the *Institut d'Etudes Méridionales* of Toulouse; see sect. 3.1 for the French part.

📖 Besch-Commengue, Bruno 1977: *Le savoir des bergers de Casabède. Volume I: Textes gascons pastoraux du Haut Salat*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 105 pages, approx. 20.000 words. 37 texts. 2 Complete meta-information available. Most recordings are dialogues between the researcher and a single informant, an elderly shepherd from Sentenac d'Oust in the Upper Couserans (Pyrenees). 3 Diatopically (Gascony Occitan) and diastratically marked (conversations about traditional breeding activities).

📖 Gonzalez, Daniel 1976: *L'occitan parlat jos tèrra. Los carbonièrs de Carmaus. Tome 1: Tèxtes sul trabalh del cròs*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN.

1 Printed corpus. 122 pages, approx. 32.000 words. 2 texts. 2 Complete meta-information available. Both recordings were made in late 1975 with six retired miners who had worked in the coal mines of Carmaux (north of Toulouse). 3 Diatopically (*occitan languedocien*) and diastratically marked (workmen's colloquial speech).

📖 Jagueneau, Liliane / Valière, Michel 1976: *L'occitan parlat à Lespignan (Hérault). La langue des viticulteurs*. Toulouse: Université de Toulouse II-Le Mirail. No ISBN. Corpus: 8-57.

1 Printed corpus. 24 pages, approx. 6.000 words. 1 text. 2 Very detailed meta-information given. The recording was made in 1973 with a 62-year old wine-grower. 3 Diatopically marked (Languedoc Occitan).

The following characteristics are shared by all of these corpora:

4 Orthographic transcription with slight phonetic adjustments. 5 After the transcripts, a translation into French is provided, except for the Gonzalez (1976) corpus which has no translation. 6 No recordings available.

📖 Mariotti, Martine 1990: *Marie Nicolas conteuse en Champsaur*. Aix-en-Provence: Edisud / Ed. du CNRS. ISBN 2-85744-486-9.

1 Printed corpus. 38 pages. 5 texts. 2 Detailed meta-information available. All texts recorded with the same informant, an elderly lady from the Champsaur valley near Gap (Dép. Hautes-Alpes). 3 Diatopically and diaphatically marked; texts consist of *contes et légendes*-type narrations in monologue form. 4 Orthographic transcription in linear notation. 5 Parallel translation into French. 6 No recordings available.

📖 Pusch, Claus D. 2001: *Morphosyntax, Informationsstruktur und Pragmatik. Präverbale Marker im gaskognischen Okzitanisch und in anderen Sprachen* (= ScriptOralia; 124). Tübingen: Narr. ISBN 3-8233-5434-5. Corpus on a CD-ROM attached to the book.

1 Electronic corpus (PDF files). 370 pages. 13 texts. 2 Depending on the text, basic or complete meta-information is available. Recordings made in different places of Gascony (Southwestern France) in 1996; some audio-taped radio programs. 3 Strongly marked on the diatopic level (Gascony Occitan), with interferences from French. Texts are arranged according to their diaphatic position on the informal-to-formal scale as developed by

¹⁰ There exist early recordings and the corresponding transcriptions of Swiss Franco-Provençal dialects made by the *Phonogrammarchiv* of Zurich University, in cooperation with the *Phonogrammarchiv* in Vienna and the *Institut für Lautforschung* in Berlin for some of them, as well as prewar transcriptions of French dialects spoken in the Swiss *Jura Bernois*, all of them with a dialectologic scope; for detailed references cf. *Phonogrammarchiv Zürich* (2001).

Koch / Oesterreicher (1985). **4** Orthographic transcription in HIAT score notation, with interlinear morphemic tags. **5** Interlinear translation into German. **6** Selected partial recordings included in the CD-ROM. Full recordings may be obtained upon request from the author.

☞ Ravier, Xavier 1986: *Le récit mythologique en Haute-Bigorre*. Aix-en-Provence: Edisud / Ed. du CNRS. ISBN 2-85744-236-X. Corpus: 29-115.

1 Printed corpus. 86 pages. 20 texts (12 entirely in Gascony Occitan, 2 mixed French-Occitan, 8 entirely in French). **2** Basic meta-information provided, but no date of recordings indicated. **3** Diatopically and diastratically marked; as the title of the book indicates, all texts are narratives of local *contes et légendes*. **4** Slightly inconsistent orthographic transcription in linear notation. **5** Occitan texts are provided with French translation following the transcript. **6** No recordings available.

Comment: This corpus was collected when the author did fieldwork for the “Atlas Linguistique de la Gascogne,” therefore the linguistic orientation is beyond doubt; however, the transcriptions do not seem entirely faithful.

5. Corpora of spoken Spanish

5.1. European Spanish

☞ Azorín Fernández, Dolores / Jiménez Ruiz, Juan Luis (eds.) 1996 [1999]: *Corpus oral de la variedad juvenil universitaria del español hablado en Alicante*. Alacant: Instituto de Cultura “Juan Gil-Albert”. ISBN 84-7784-287-6.

1 Printed corpus. 357 pages, 178.000 words. 13 texts. **2** Basic meta-information given in the headers that introduce the texts. Recordings made in late 1996 mostly on the campus of Alacant / Alicante University in Eastern Spain. **3** Strongly marked on the diastratic level, but only very slightly marked diatopically. All texts are semi-guided conversations between peers. **4** Orthographic transcription in a linear notation following TEI recommendations for text-descriptive mark-up and including also information on extra-linguistic events. **5** No translation available. **6** No recordings available.

Comment: Interesting corpus documenting exclusively students’ language. An electronic version on CD-ROM may be obtained on request from the authors.

☞ Briz, Antonio (ed.) 1995: *La conversación coloquial (Materiales para su estudio)* (= Cuadernos de Filología; Anejo 16). València: Universitat de València / Departamento de Filología Española. ISBN 84-370-2071-9. Corpus: 55-251.

1 Printed corpus. 152 pages, approx. 55.000 words. 9 texts. **2** Highly detailed meta-information is provided for all the texts, which were recorded in and around the Eastern Spanish town of València between 1989 and 1994. **3** All texts represent casual conversations and are slightly marked diatopically (some interferences from Catalan). **4** Orthographic transcription with phonetic adjustments in a score-like notation. Prosodic features annotated in-line. **5** No translation available. **6** No recordings available.

☞ Criado del Val, Manuel 1980: *Estructura general del coloquio*. Madrid: C.S.I.C. ISBN 84-7143-200-5. Corpus: 75-139.

Comment: See Koch / Oesterreicher (1990: 44s).

☞ Esgueva, Manuel / Cantarero, Margarita 1981: *El habla de la ciudad de Madrid. Materiales para su estudio* (= La norma lingüística culta de la lengua española hablada en Madrid; 1). Madrid: C.S.I.C. ISBN 84-00-04990-X.

Comment: See Koch / Oesterreicher (1990: 47). 14 of the Madrid texts are available in electronic form on the Samper Padilla *et al.* (1998) *Macrocorpus* CD-ROM described in section 5.3.

☞ García Cid, Aranzazu 1995: *Parenthesen, Einschübe und Kommentare: Zur Klassifikation von Nebenprädikationen in gesprochenen spanischen Texten* (= Arbeitspapier N.F.; 25). Köln: Universität / Institut für Sprachwissenschaft. No ISBN. Corpus: 63-110.

1 Printed corpus. 43 pages, 6.300 words. 8 texts. **2** Basic meta-information provided. Recordings made with informants from different parts of Spain but who have been living in Germany for some time. **3** Generally unmarked speech. **4** Orthographic transcription in linear notation. All texts are monologues. **5** Parallel translation into German. **6** No recordings available.

📖 Gómez Molina, José Ramón (ed.) 2001: *El español hablado de Valencia. Materiales para su estudio. I. Nivel sociocultural alto* (= Cuadernos de Filología; Anejo 46). València: Universitat de València / Facultat de Filologia. ISBN 84-370-5250-5.

1 Printed corpus. 358 pages, approx. 145.000 words. 24 texts. 2 Detailed meta-information available. Recordings were made between 1996 and 1999. 3 All informants are educated speakers with university degrees; many are bilingual (Spanish / Valencian Catalan), with Catalan being in some cases the L1, but interferences are very moderate. No diaphatic variation, as all texts are semi-guided sociolinguistic interviews. 4 Orthographic transcription with some minor adjustments to represent prosodic features. Mostly linear notation, with score-like passages where necessary. 5 No translation available. 6 No recordings available.

Comment: This is the first published corpus of the “Proyecto para el Estudio Sociolingüístico del Español de España y de América” (PRESEEA) initiated in 1996 and which – on the methodological basis of the *Habla culta* project started in the late 1960’s (cf. section 5.2) – aims at describing the urban speech in selected areas of the Spanish-speaking world on three diastatic levels.

📖 Marcos Marín, Francisco *et al.* 1992: *Corpus oral de referencia del español contemporáneo*. Madrid: Universidad Autónoma de Madrid. Accessible for free download at <http://www.llf.uam.es/corpus_lee.html>.

1 Electronic corpus (ASCII files). 1,1 Mio words (6,3 MB). 500 texts, organized in 10 folders (text-typological subgroups). 2 Basic meta-information available in the transcription files’ headers. Recordings made with informants from Madrid and surroundings in 1991. 3 Some texts – namely the journalistic ones – marked diaphatically, but, apart from this, mostly unmarked speech. 4 Orthographic transcription in linear notation. Files encoded in conformity with TEI guidelines, although full file headers are not generally provided. 5 No translation available. 6 No recordings available.

Comment: Currently, the most important and most accessible electronic orality-only corpus for Spanish and for Romance in general.

The following three corpora, documenting the Spanish as spoken in Sevilla, are arranged together as they resulted from the same research project and were published in the same series along common guidelines and transcription principles. Please note that 14 of these Sevillian texts are included in the Samper Padilla *et al.* (1998) *Macrocorpus* on CD-ROM described in section 5.3.:

📖 Pineda, Miguel Ángel de (ed.) 1983: *Sociolingüística andaluza. Vol. 2: Material de encuestas para el estudio del habla urbana culta de Sevilla*. Sevilla: Servicio de Publicaciones de la Universidad de Sevilla. ISBN 84-7405-274-2.

Comment: See Koch / Oesterreicher (1990: 48s).

📖 Roper, Miguel (ed.) 1987: *Sociolingüística andaluza. Vol. 4: Encuestas de nivel popular*. Sevilla: Servicio de Publicaciones de la Universidad de Sevilla. ISBN 84-7405-369-2.

1 Printed corpus. 458 pages, approx. 85.000 words. 24 texts.

Comment: Concerning the other parameters, this corpus is identical with the aforementioned Pineda (1983) corpus; please refer to Koch / Oesterreicher’s (1990: 48s) description.

📖 Ollero Toribio, Manuel / Pineda Pérez, Miguel Ángel de (eds.) 1992: *Sociolingüística andaluza. Vol. 6: Encuestas del habla urbana de Sevilla. Nivel medio*. Sevilla: Servicio de Publicaciones de la Universidad de Sevilla. ISBN 84-7405-823-6.

1 Printed corpus. 222 pages, approx. 95.000 words. 24 texts.

Comment: Also this corpus is comparable to the Pineda (1983) corpus concerning the remaining parameters; refer to Koch / Oesterreicher (1990).

5.2. Overseas Spanish

An important number of printed corpora documenting overseas Spanish is available, thanks to an ambitious joint research program “Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica,” initiated in the late 1960’s (cf. Quilis 1985; Lope Blanch 1986; Koch / Oesterreicher 1990: 43). Although the 12 corpora published hitherto in the series (some of them in more than one printed volume, others only electronically) are far from being uniform, the corpora in this series which concern varieties of overseas Spanish are arranged together in order to present them (the printed corpora which document European cities have been mentioned before [Esgueva / Cantarero’s 1981 Madrid corpus and the Sevilla corpora]; the Canary Islands’ corpus from Las Palmas appears only in the electronic *Macrocorpus* compilation, see Samper Padilla *et al.* [1998] in section 5.3.). Currently, the PRESEEA project continues the *Habla culta* initiative; cf. the Gómez Molina (2001) corpus mentioned in section 5.1.

📖 Barrenechea, Ana María (ed.) 1987: *El habla culta de la ciudad de Buenos Aires. Materiales para su estudio*. 2 vol. Buenos Aires: Universidad Nacional de Buenos Aires. ISBN 950-29-0020-0.
Comment: See Koch / Oesterreicher (1990: 45s).

📖 Caravedo, Rocío 1989: *El español de Lima. Materiales para el estudio del habla culta*. Lima: Pontificia Universidad Católica del Perú. No ISBN.

1 Printed corpus. 257 pages, approx. 110.000 words. 23 texts.

📖 Lope Blanch, Juan (ed.) 1971: *El habla de la ciudad de México. Materiales para su estudio*. México: Universidad Nacional Autónoma de México. No ISBN.

Comment: See Koch / Oesterreicher (1990: 47s).

📖 Marrone, Nila G. 1992: *El habla de la ciudad de La Paz. Materiales para su estudio*. La Paz: Ediciones Signo. No ISBN.

1 Printed corpus. 356 pages, approx. 150.000 words. 32 texts.

📖 Morales, Amparo / Vaquero, María (eds.) 1990: *El habla culta de San Juan. Materiales para su estudio*. Río Piedras: Editorial de la Universidad de Puerto Rico. ISBN 0-8477-3641-5.

1 Printed corpus. 424 pages, approx. 150.000 words. 23 texts.

📖 Otálora de Fernández, Hilda / González G., Alonso 1986: *El habla de la ciudad de Bogotá. Materiales para su estudio. Selección y transcripción de muestras* (= Publicaciones del Instituto Caro y Cuervo; 75). Bogotá: Instituto Caro y Cuervo. No ISBN.

Comment: See Koch / Oesterreicher (1990: 46). A second, enlarged edition, containing some additional transcripts, appeared in 1990.

📖 Rabanales, Ambrosio / Contreras, Lidia (eds.) 1979: *El habla culta de Santiago de Chile. Materiales para su estudio*, vol. 1 (= Boletín de Filología; Anejo 2). Santiago de Chile: Editorial Universitaria / Departamento de Lingüística y Filología. No ISBN.

Comment: See Koch / Oesterreicher (1990: 48).

📖 Rabanales, Ambrosio / Contreras, Lidia (eds.) 1990: *El habla culta de Santiago de Chile. Materiales para su estudio*, vol. 2. Bogotá: Instituto Caro y Cuervo. No ISBN.

1 Printed corpus. 541 pages, approx. 190.000 words. 29 texts.

Comment: This second volume contains 23 highly interactional ‘free dialogues’ and six monological formal speeches (“conferencias”).

📖 Rosenblat, Angel (ed.) 1979: *El habla culta de Caracas. Materiales para su estudio*. Caracas: Universidad Central de Venezuela / Instituto de Filología ‘Andrés Bello’. No ISBN.

Comment: See Koch / Oesterreicher (1990: 46s.).

The following features are shared by all the above-mentioned *Habla culta* corpora:

2 Generally basic, sometimes detailed meta-information available. Texts are arranged according to the informants’ gender and age group, with these two variables being evenly distributed. 3 As one may expect from a project stretching over the entire Spanish-speaking world, there is significant diatopic variation, but other diastematic parameters vary much less. All the *Habla culta* corpora include three text types: Interviews with an informant; semi-spontaneous dialogues between two (or several) informants, and situations of formal (e.g., public) speech. Sometimes (as in the La Paz corpus by Marrone [ed. 1992]), secretly recorded texts are added as further text type. 4 Orthographic transcription with very few adjustments to phonetic or prosodic features (except for the Caravedo [1989] corpus, which furthermore contains three interviews in phonetic transcription); linear notation. 5 No translation available. 6 No recordings available.

📖 Green, Katherine Reese 1997: *Non-Standard Dominican Spanish: Evidence of partial restructuring*. New York: City University of New York. PhD dissertation, DAI no. 9720094. Corpus: 256-283.

1 Printed corpus. 27 pages. 4 texts. 2 Only basic meta-information given. 3 Diatopically marked, as title of the thesis indicates. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

The following three corpora, edited by J. Lope Blanch, do not form part of the “Estudio coordinado” project, for they did not focus on educated speech registers, but these corpora are very similar to the *Habla culta* publications as far as corpus design, transcription principles and presentation are concerned:

📖 Lope Blanch, Juan (ed.) 1976: *El habla popular de la ciudad de México. Materiales para su estudio*. México: Universidad Nacional Autónoma de México. No ISBN.

1 Printed corpus. 430 pages, approx. 150.000 words. 34 texts. 2 Complete meta-information given in the headers which introduce the texts. Recordings date from the early 1970's. 3 Obvious diatopic and diastratic features, in accordance with the goal of this corpus. Text types correspond perfectly to the general scheme applied within the *Habla culta* project.

📖 Lope Blanch, Juan (ed.) 1990: *El español hablado en el suroeste de los Estados Unidos. Materiales para su estudio*. México: Universidad Nacional Autónoma de México. ISBN 968-36-1414-0. Corpus: 101-331.

1 Printed corpus. 208 pages, approx. 65.000 words. 20 texts. 2 Mostly detailed meta-information available. Recordings were made in 1985/1986 in four cities in Texas, New Mexico, Arizona and California with elderly informants speaking the traditional varieties of North American Spanish; immigrants who had arrived only recently in the US were excluded. 3 Texts – guided interviews and semi-guided conversations – show clearly diatopically marked features.

📖 Lope Blanch, Juan (ed.) 1995: *El habla popular de la República Mexicana. Materiales para su estudio*. México: Universidad Nacional Autónoma de México / Colegio de México. ISBN 968-36-4819-3.

1 Printed corpus. 563 pages, approx. 200.000 words. 47 texts. 2 Generally detailed meta-information given in the headers. Texts were recorded in the different States of the Republic of Mexico during the fieldwork for the *Atlas Lingüístico de México*. 3 Texts, which are interviews in most cases, show the diatopic variation within non-urban Mexican Spanish; many of them are diastratically marked.

The following characteristics are shared by all these corpora on North American Spanish published by J. Lope Blanch:

4 Orthographic transcription with very few phonetic adjustments, in linear notation. 5 No translation available. 6 No recordings available.

📖 Martorell de Laconi, Susana (ed.) 2000: *Habla culta de la Ciudad de Salta. Materiales para su estudio (desgrabaciones)*. Salta: Instituto Salteño de Investigaciones Dialectológicas 'Berta Vidal de Battini'. ISBN 987-98498-0-9.

1 Printed corpus. 216 pages, approx. 65.000 words. 100 texts, most of them rather short. 2 Basic meta-information available. Informants are distributed in three age groups. 3 Very moderate diatopic and diastratic marking. Diaphasic variation is higher, with texts ranging from low to fairly high formality. 4 Orthographic transcription with minor phonetic adjustments, in linear notation. 5 No translation available. 6 No recordings available.

Comments: This is a *Habla culta* corpus not included in the "Estudio coordinado" project which nevertheless follows closely its guidelines.

📖 Mendoza, José G. 1991: *El castellano hablado en La Paz. Sintaxis divergente*. La Paz: Universidad Mayor de San Andrés / Facultad de Humanidades y Ciencias de la Educación. No ISBN. Corpus: 231-263.

1 Printed corpus. 32 pages, approx. 7.000 words. 3 texts. 2 Basic meta-information available. Recordings made in 1988. 3 Diastratic variation: educated ("variedad culta"), casual ("variedad popular") and L2 (i.e. primary Aymara bilingual) speakers. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

📖 Meyer-Hermann, Reinhard (ed.) 1982: *Spanisch II* (= Bielefelder Text-Corpora romanischer Sprachen; 4). Bielefeld: Universität / Fakultät für Linguistik und Literaturwissenschaft. No ISBN; typescript.

Comment: This corpus documents Mexican Spanish; see Koch / Oesterreicher (1990: 49). The volume *Spanisch I*, which should have been dedicated to European Spanish, has never appeared.

5.3. Corpora including both European and extra-European Spanish

📖 Real Academia Española (ed.) 1998-: *CREA – Corpus de Referencia del Español actual* (= Banco de datos del español). Madrid: RAE. Accessible at <<http://www.rae.es/nivel1/corpusay.htm>>.

1 On-line corpus. CREA is a mixed corpus embracing written (90%) and oral texts (10%); its total size being currently estimated at 125 Mio words, the share of oral speech is supposed to be approx. 12 Mio words. Number of texts unknown. 2 Basic meta-information available in the files' headers, which may be visualized after having completed a search through the web interface. 3 As the corpus includes texts in both European and Overseas Spanish (50% each), diatopic variation is obvious. Other diasystematic parameters are impossible to check systematically, as corpus texts as such are not accessible individually. 4 Orthographic transcription. Files seem to be TEI-encoded. 5 No translation available. 6 No recordings available.

Comment: This extremely useful on-line corpus, which is continuously enlarged, allows very specific searches with numerous search parameters (geographical, text-typological, thematic, and linguistic criteria) via its web interface and delivers quantitative results and concordance tables. However, no access is given to the corpus texts as such, and there is no download option for off-line analysis either.

The CREA corpus includes many other corpora (such as the Marcos Marín *et al.* [1992] or the *Habla culta* corpora) which are available in off-line versions. The same web-site gives also access to CORDE (*Corpus diacrónico del español*).

☒ Samper Padilla, José Antonio / Hernández Cabrera, Clara Eugenia / Troya Déniz, Magnolia (eds.) 1998: *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria / ALFAL. ISBN 84-89728-88-7.

1 Electronic corpus on CD-ROM. Approx. 720.000 words. 168 texts. 2 Only very basic meta-information available. 3 Being a compilation of different *Habla culta* corpora, this *Macrocorpus* shares the diasystematic preferences of the overall project (focus on diatopic variation among educated speakers in different parts of the Spanish-speaking world). Only the text type “Dialogues between an interviewer and an informant” has been included in this compilation. 4 Strict orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

Comment: This extremely useful (and affordable) CD-ROM represents a synthesis of the different European and Hispano-American *Habla culta* corpora in an easy-to-use electronic format. It contains texts (14 each) from the following cities: Bogotá, Buenos Aires, Caracas, La Paz, Las Palmas de Gran Canaria, Lima, Madrid, Ciudad de México, San José de Costa Rica, San Juan de Puerto Rico, Santiago de Chile and Sevilla. As can be seen from this list, some corpora are published on the CD-ROM for the first time, whereas for some cities which have had their corpora published in printed form before, some new texts have been included. The editors of the *Macrocorpus* have been at pains to unify the transcription and notation principles of the different corpora. The CD-ROM includes the texts in four formats (RTF, DOC [MS WORD] and HTML files, and TEI-encoded versions as used by the CREA corpus’ data-base).

☒ Sánchez, Aquilino / Cantos, Pascual (eds.) 2001: *Corpus CUMBRE del español contemporáneo de España e Hispanoamérica. Extracto de dos millones de palabras*. Madrid: SGEL. No ISBN.

1 Electronic corpus on CD-ROM. Approx. 400.000 words from oral sources. 2 No meta-information available. 3 Comparable to the CREA corpus, the CUMBRE corpus includes oral texts in both European and Overseas Spanish (approx. 50% each). All texts come from audio- or video-taped TV and radio programs. Further diasystematic parameters are impossible to check, as corpus texts as such are not accessible. 4 Orthographic transcription. 5 No translation available. 6 No recordings available.

Comment: This CD-ROM, containing a representative selection of the materials that make up the original 20-million words CUMBRE corpus (cf. Sánchez 1995 for details), comes with its own search tools which allow all types of queries generally used in corpus linguistics. Unfortunately, in the concordance results window, the maximal context of tokens that is displayed is the sentence, which reduces the analyzability of the occurrences considerably.

The CD-ROM was distributed in 2001 as a promotional special offer together with the *Gran Diccionario de uso del español actual* (ISBN 84-7143-872-0) by the same publisher.

6. Portuguese

6.1. European Portuguese

☒ ☒ Bacelar do Nascimento, Maria Fernanda / Garcia Marques, Maria Lúcia / Segura da Cruz, Maria Luísa 1987: *Português Fundamental. Vol. II, Métodos e Documentos, tomo 1, Inquérito de Frequência*. Lisbon: Instituto Nacional de Investigação Científica / Centro de Linguística da Universidade de Lisboa. No ISBN. Corpus: 79-309.

1 Printed corpus. 230 pages, approx. 100.000 words. 140 texts. 2 Complete meta-information is given within the detailed description of the collection and edition of the corpus data (pages 36-72 of the book). Recordings made between 1970 and 1974 in different parts of Portugal. 3 All texts are interviews that are slightly to moderately marked on the diatopic and diastratic level. 4 Orthographic transcription in linear notation. 5 No translation available. 6 Recordings are accessible locally at the University of Lisbon only.

Comment: This important corpus of European Portuguese will be the core component of the spoken language section of a huge *Corpus de Referência do Português Contemporâneo* (cf. Bacelar do Nascimento 2000a, 2000b). An electronic version (HTML files) of the published parts of the *Português Fundamental* corpus is available for free download at <http://www.clul.ful.pt/sectores/corpus_oral_pf_publicado.zip>.

☞ Pinto Coelho, Maria Zara *s.a.*: *A droga de boca em boca. Conversas informais sobre drogados*. Braga: Universidade do Minho / Ciências da Comunicação. Accessible for free download at <<http://natura.di.uminho.pt/ijbin/corpora/>>.

1 Electronic corpus (ASCII file). 187.000 words (1 MB). 36 texts. 2 Basic meta-information is provided in the header that precedes each text. All texts are interviews made in Braga and surroundings (Northern Portugal) in 1998 with informants of different age groups who were questioned on urban change, social problems and drugs. 3 No significant diastematic marking. 4 Orthographic transcription in linear notation. Some non-TEI conform tags are used for text structure encoding. 5 No translation available. 6 No recordings available.

6.2. Brazilian Portuguese

☞ Callou, Dinah *et al.* *s.a.*: *Norma urbana oral culta do Rio de Janeiro. Corpus comparativo da década 70-90*. Rio de Janeiro: UFRJ. Accessible (and downloadable) at <<http://www.lettras.ufrj.br/nurc-tj/>>.

1 Electronic corpus (HTML files). Approx. 180.000 words. 37 texts. 2 Complete meta-information provided in the text headers and accompanying material. 11 texts come from fieldwork carried out in the 1970's, whereas the remainder was recorded in 1992 and 1996. 3 All texts are interviews. Slightly to moderately marked on the diastematic levels. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

Comment: An interesting and perfectly accessible corpus that belongs conceptually to the NURC series (see below) but takes a longitudinal perspective: some of the informants interviewed in the 1970's were recorded a second time in the 90's; these recordings were complemented with interviews of new informants to cover all age groups. The corpus is supposed to be used for the elaboration of a *Gramática do Português falado*.

The following nine corpora are presented together because they form part of a coordinated research project, which came to be known as 'Projecto Norma Urbana Culta' (NURC), and are therefore very similar in structure and corpus design. The NURC project, initiated in the early 1970's, aimed at extending the research in (over-seas) urban varieties as carried out within the Hispanic *Habla culta* project (see section 5.2.) to five Brazilian towns (see Castilho 1990 for details on the project's history). Therefore, the guidelines developed for the *Habla culta* initiative were also adopted for the NURC corpora. As in the *Habla culta* corpora, NURC texts are arranged according to degrees of formality and elaboration, communicative situation and the informants' gender and age group.

☞ Castilho, Ataliba Teixeira de / Preti, Dino (eds.) 1986: *A linguagem falada culta na cidade de São Paulo: materiais para o seu estudo. Vol. 1: Elocuções formais*. São Paulo: T.A. Queiroz. ISBN 85-85008-47-4.

1 Printed corpus. 85 pages, approx. 27.000 words. 6 monological texts.

☞ Castilho, Ataliba Teixeira de / Preti, Dino (eds.) 1987: *A linguagem falada culta na cidade de São Paulo: materiais para o seu estudo. Vol. 2: Diálogos entre dois informantes*. São Paulo: T.A. Queiroz. ISBN 85-85008-70-9.

1 Printed corpus. 242 pages, approx. 60.000 words. 6 texts.

☞ Preti, Dino / Urbano, Hudinilson (eds.) 1990: *A linguagem falada culta na cidade de São Paulo: materiais para o seu estudo. Vol. 3: Diálogos entre informante e documentador*. São Paulo: T.A. Queiroz. ISBN 85-85008-85-7.

1 Printed corpus. 142 pages, approx. 38.000 words. 9 texts.

☞ Callou, Dinah (ed.) 1992: *A linguagem falada culta na cidade do Rio de Janeiro: materiais para o seu estudo. Vol. 1: Elocuções formais*. Rio de Janeiro: UFRJ / Faculdade de Letras. No ISBN.

1 Printed corpus. 107 pages, approx. 30.000 words. 6 texts.

☞ Callou, Dinah / Lopes, Célia Regina (eds.) 1993: *A linguagem falada culta na cidade do Rio de Janeiro: materiais para o seu estudo. Vol. 2: Diálogos entre informante e documentador*. Rio de Janeiro: UFRJ / Faculdade de Letras. No ISBN.

1 Printed corpus. 212 pages, approx. 60.000 words. 9 texts.

Callou, Dinah / Lopes, Célia Regina (eds.) 1994: *A linguagem falada culta na cidade do Rio de Janeiro: materiais para o seu estudo. Vol. 3: Diálogos entre dois informantes*. Rio de Janeiro: UFRJ / Faculdade de Letras. No ISBN.

1 Printed corpus. 287 pages, approx. 80.000 words. 7 texts.

Comment: An electronic version (HTML files) of the three NURC-RJ (Rio de Janeiro) corpora is accessible (and downloadable with the browsers' 'save' option) at <<http://www.letras.ufrj.br/nurc-rj/>>.

Hilgert, José Gaston (ed.) 1997: *A linguagem falada culta na cidade de Porto Alegre: materiais para o seu estudo. Vol. 1: Diálogos entre informante e documentador*. Passo Fundo / Porto Alegre: Editora da Universidade Federal do Rio Grande do Sul / Ediupf. No ISBN.

1 Printed corpus. 198 pages, approx. 40.000 words. 8 texts.

Mota, Jacyra / Rollemberg, Vera (eds.) 1994: *A linguagem falada culta na cidade de Salvador. Materiais para seu estudo. Vol. 1: Diálogos entre informante e documentador*. Salvador: Universidade Federal da Bahia / Instituto de Letras. No ISBN.

1 Printed corpus. 252 pages, approx. 75.000 words. 12 texts.

Sá, Maria da Piedade Moreira de *et al.* (eds.) 1996: *A linguagem falada culta na cidade do Recife: materiais para seu estudo. Vol. 1: Diálogos entre informante e documentador*. Recife: Universidade Federal de Pernambuco. No ISBN.

1 Printed corpus. 150 pages, approx. 90.000 words. 12 texts.

The following characteristics are shared by all NURC corpora:

2 Basic, sometimes detailed meta-information given in the headers that precede the transcription texts. 3 Texts marked (slightly) as far as diatopic variation is concerned, while other diasystematic features are weak. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

Silva de Aragão, Maria do Socorro / Soares, Maria Elias (eds.) 1996: *A linguagem falada em Fortaleza. Diálogos entre informantes e documentadores (Materiais para estudo)* (= Projeto Dialetos sociais cearenses). Fortaleza: Universidade Federal do Ceará. No ISBN.

1 Printed corpus. 468 pages, approx. 115.000 words. 18 texts. 2 General meta-information available. Texts were recorded in 1986/87 in different parts of the city of Fortaleza. 3 Texts marked on the diatopic and diastratic level, but homogeneous on the diaphatic level. All are semi-guided interviews. 4 Orthographic transcription in mostly linear notation, with passages noted score-like. The notation system as proposed by Marcuschi (1986) has been applied. 5 No translation available. 6 No recordings available.

Comment: Although this corpus does not form part of the NURC project, it takes over its general perspective and the principles of sample and corpus constitution. However, informants are overwhelmingly members of lower social classes.

6.3. Corpora including both European and extra-European Portuguese

Bacelar do Nascimento, Maria Fernanda (ed.) 2001: *Português Falado. Documentos autênticos. Gravações audio com transcrição alinhada*. Lisbon: Centro de Linguística da Universidade de Lisboa / Instituto Camões. No ISBN.

1 Electronic corpus (ASCII files) on four CD-ROMs. 92.000 words. 86 texts. 2 Basic meta-information given in the transcription files' headers. Recordings were made from the 1970's up to the 90's and include 30 texts from Portugal, 20 from Brasil, 25 from Portuguese-speaking countries in Africa and 11 from Portuguese outposts in Asia. 3 Diatopic and diastratic marking is weaker than one would expect for a sample of spoken Portuguese with a world-wide coverage. Diaphatic features are also weak, most texts being interviews with one informant. 4 Strict orthographic transcription in linear notation. 5 No translation available. 6 The complete recordings of all texts are included as WAVE files on the CD-ROMs. Text and sound are aligned, with a moving gray bar indicating in the transcription file the progress of the sound file.

Comment: Even if this sample was published primarily as a didactic tool for teaching Portuguese, its usefulness for research purposes is obvious.¹¹

11 A 40.000-word corpus containing talkshow transcripts recorded in 1997/98 in Portugal and Brazil will be published in printed form during 2002 in Barne (to appear).

7. Corpora of other Ibero-Romance languages

7.1. Catalan

📖 Berkenbusch, Gabriele 1988: *Sprachpolitik und Sprachbewußtsein in Barcelona am Anfang dieses Jahrhunderts. Versuch einer Rekonstruktion auf der Grundlage mündlicher und schriftlicher Quellen am Beispiel des Erziehungswesens*. Frankfurt am Main: Lang. ISBN 3-8204-1472-X. Corpus: 334-456.

1 Printed corpus. 115 pages, approx. 28.000 words. 5 texts. 2 Detailed meta-information is given in the headers preceding the transcriptions. Recordings made in 1983. 3 Sociolinguistic interviews focussing on life stories. No significant diatopic or diastratic characteristics. 4 Orthographic transcription in HIAT-like score notation. Detailed annotation, e.g. prosodic features. 5 No translation available. 6 No recordings available.

📖 Guàrdia, Roser (ed.) 1985: *Katalanisch I. Debats radiofònics "Parlem-ne" (Ràdio Quatre)* (= Bielefelder Text-Corpora romanischer Sprachen; 6). Bielefeld: Universität / Fakultät für Linguistik und Literaturwissenschaft. No ISBN; typescript.

1 Printed corpus. 144 pages, approx. 30.000 words. 3 texts. 2 Only basic meta-information available. All texts are audio-taped radio programs recorded in September 1983 in Catalonia. 3 Semi-spontaneous radio discussions, with weak diatopic marking. 4 Orthographic transcription in a predominantly linear notation. 5 No translation available. 6 No recordings available.

📖 Miralles, Joan 1995: *Un poble, un temps*. Palma de Mallorca: Miquel Font. ISBN 84-7967-050-9. Corpus: 67-427.

1 Printed corpus. 267 pages, approx. 70.000 words. 21 texts. 2 Complete meta-information is given. All informants, who were interviewed in 1969 and in the early 70's, lived in the village of Montuiri (Majorca). 3 Strong diatopic features of Majorcan Catalan. 4 Orthographic transcription with many adjustments to reproduce the phonetic peculiarities of Balearic Catalan; linear notation. The transcriptions seem to be slightly polished up for the convenience of the reader. 5 No translation available. 6 No recordings available.

📖 Montoya Abat, Brauli 2000: *Els alacantins catalanoparlants: una generació interrompuda* (= Biblioteca de Dialectologia i Sociolingüística; 7). Barcelona: Institut d'Estudis Catalans. ISBN 84-7283-520-0. Corpus: 157-190.

1 Printed corpus. 33 pages, approx. 13.000 words. 9 texts. 2 Basic meta-information available. Recordings made in 1993/94 with Catalan-speaking informants from the Eastern Spanish town of Alacant (Alicante). Some videotaped TV programs. 3 Strong diatopic features and many interferences from Spanish. 4 Hybrid orthographic transcriptions with many adjustments to represent phonetic features. Some parts of the texts which are of particular phonological interest are also transcribed in a parallel phonetic transcription. Linear notation. 5 No translation available. 6 No recordings available.

📖🎧 Veny, Joan / Pons i Griera, Lúdia 1998: *Atles Lingüístic del Domini Català. Etnotextos del català oriental* (= Biblioteca de Dialectologia i Sociolingüística; 5). Barcelona: Institut d'Estudis Catalans. ISBN 84-7283-435-2. *Comment:* See Pons / Massanell (in this volume).

7.2. Galician

The CORGA reference corpus of contemporary Galician, which offers a very convenient on-line access (cf. <<http://corpus.cirp.es/corga>>), includes 'oral' as one of the selection features listed in the query interface, but spoken language texts do not seem to be available in the data-base yet.

🎧📖 Fernández Rei, Francisco / Hermida Gulías, Carme (eds.) 1996: *A nosa fala. Bloques e áreas lingüísticas do galego*. Santiago de Compostela: Consello da Cultura Galega / Arquivo sonoro de Galicia. ISBN 84-87172-11-3.

1 Printed corpus. 92 pages. 43 texts. 2 Only basic meta-information is provided. Texts were recorded in all parts of the Galician-speaking territory between 1974 and 1995. 3 Interviews ("ethnotexts"), only parts of which are actually transcribed. Strong diatopic variation, corresponding to the dialectologic scope of the volume. 4 Orthographic transcription in linear notation. 5 No translation available. 6 The corpus volume is published in a box with 3 audio-cassettes including the full recordings of the transcribed texts.

📖📄 Kabatek, Johannes 1996: *Die Sprecher als Linguisten. Interferenz- und Sprachwandelphänomene dargestellt am Galicischen der Gegenwart* (= Beihefte zur ZRPh; 276). Tübingen: Niemeyer. ISBN 3-484-52276-3. Corpus: 209-434. Also available in the Galician translation of this monograph: Kabatek, Johannes 2000: *Os falantes como lingüistas. Tradición, innovación e interferencias no galego actual*. Vigo: Edicións Xerais de Galicia. ISBN 84-8302-530-2. Corpus: 271-447. The transcripts are downloadable as word processing files at <<http://www.uni-erfurt.de/sprachwissenschaft/romanistik/corpus1.html>>.

1 Printed corpus. 223 pages, approx. 80.000 words. 26 texts. 2 Detailed meta-information on informants and recording situation is provided. All texts were recorded in 1993 in Santiago de Compostela. 3 Rather moderate diatopic marking, due to the study's aim and to informants' age. Most texts are guided interviews with either students of Galician philology or broadcast professionals. 4 Orthographic transcription in mostly linear notation. 5 No translation available. 6 No recordings available.

8. Corpora of spoken Italian

Much of the published spoken language data for Italian has been collected for the purpose of dialectologic studies. Phonetic transcription is still – though gradually less – common for this kind of corpora. As explained in sect. 1, dialectologic text collections have not been systematically included in the following corpus list.

📖 Arnuzzo, Anna Maria 1976: *Rilievi di italiano popolare nel Basso Monferrato*; in: *Problemi di morfosintassi dialettale. Atti dell'XI Convegno del C.S.D.I.* Pisa: Pacini, 83-105. No ISBN. Corpus: 93-104.
Comment: See Koch / Oesterreicher (1990: 36).

📖 Bertinetto, Pier Marco / Leoni, Federico Albano / Locchi, Donatella / Refice, Mario (eds.) *s.a.*: *AVIP – Archivio delle Varietà di Italiano Parlato*. Pisa et al.: Consorzio AVIP. Downloadable at <<ftp://ftp.cirass.unina.it/cirass/pub/avip/>>.

1 Electronic corpus (ASCII and WAVE files). Approx. 40.000 words. 15 texts. 2 Extensive meta-information for some recordings available in separate (downloadable) header files, whereas other texts come with only basic or no meta-information. Recordings made in the late 1990's in Pisa, Naples and Bari. 3 Moderately marked on the diatopic level. All texts are semi-spontaneous 'map task' dialogues, where informants are invited to describe a path along different landmarks on a map (cf. Thompson / Anderson / Bader 1995 for the experimental design in a very similar English project). 4 Orthographic transcription in linear notation. 5 No translation available. 6 The recordings of all AVIP texts are downloadable in WAVE format.

📖 Bianconi, Sandro ²1980: *Lingua matrigna. Italiano e dialetto nella Svizzera italiana*. Bologna: Il Mulino. No ISBN. Corpus: 176-188.
Comment: See Koch / Oesterreicher (1990: 36)

📖 Bruschi, Renzo 1980: *Venticinque racconti popolari nel vernacolo di Sorifa (con un'appendice di indovinelli e filastrocche)*. *Testi dialettali in trascrizione fonetica del territorio di Nocera Umbra* (= Opera del vocabolario dialettale umbro; 6). Perugia: Istituto di Filologia Romanza. No ISBN. Spontaneous oral corpus: 13-53.

1 Printed corpus. 35 pages. 25 texts of unequal length, some very short. 2 Geographical meta-information available, but the corpus lacks completely of details on the informants. 3 Strongly marked on the diatopic level. Contrary to what the title may suggest, the texts are not predominantly of the folk-tale type. 4 Non-API conform phonetic transcription in linear notation. 5 No translation available. 6 No recordings available.

The following series of corpora – all of them on Swiss Italian dialects – are presented together because they have many methodological features in common. All the booklets and accompanying records are part of the “Dialetti della Svizzera Italiana” collection, itself a subseries to “Dialetti Svizzeri. Dischi e testi dialettali editi dall'Archivio fonografico dell'Università di Zurigo” (“Schweizer Dialekte in Text und Ton”), published by the *Phonogrammarchiv* Zurich (see also sect. 9.2). The booklets were printed in Lugano by Mazzuconi:¹²

12 In order not to overcharge this list with corpora referring to Swiss-Italian dialects, the early recordings and transcriptions by the *Phonogrammarchiv* of Zurich University, many of which would not be considered ‘spoken language corpora’ in the modern sense, are not listed here. See *Phonogrammarchiv Zürich* (2001) for a complete inventory. – The ‘Dialetti della Svizzera Italiana’ series is continued by the ‘Documenti orali della Svizzera italiana’ collection, presented at the end of this section.

☞📖 Camastral, Peter / Leissing-Giorgetti, Sonja (eds.) 1974: *Valle Maggia TI* (= *Dialetti della Svizzera italiana*; 2). No ISBN.

1 Printed corpus. 16 pages. 13 texts.

☞📖 Leissing-Giorgetti, Sonja / Vicari, Mario (eds.) 1975: *Valle Onsernone – Centovalli – Valle Verzasca TI* (= *Dialetti della Svizzera italiana*; 3). No ISBN.

1 Printed corpus. 27 pages. 14 texts.

☞📖 Vicari, Mario (ed.) 1978: *Locarnese – Terre di Pedemonte TI* (= *Dialetti della Svizzera italiana*; 4). No ISBN.

1 Printed corpus. 32 pages. 16 texts.

☞📖 Vicari, Mario (ed.) 1980: *Valle Riviera – Bellinzonese TI* (= *Dialetti della Svizzera italiana*; 5). No ISBN.

1 Printed corpus. 29 pages. 32 texts.

☞📖 Vicari, Mario (ed.) 1983: *Malcantone (Cantone Ticino)* (= *Dialetti della Svizzera italiana*; 6). No ISBN.

1 Printed corpus. 39 pages. 13 texts.

The following characteristics are shared by all “*Dialetti della Svizzera italiana*” corpora:

2 Detailed meta-information is given. Recordings were made in the late 1960’s and the 1970’s. 3 Diatopic variation, corresponding to the collection’s objective, is obvious. Diastratic variation, however, is weak. Text types vary, with most texts belonging to the autobiographic *récits de vie* genre. 4 Orthographic (dialectal) transcription and a parallel phonetic transcription in linear notation. 5 A parallel translation into Standard Italian is provided, resulting in a 3-columns layout of the corpus data. 6 Recordings of all transcribed texts are available on the vinyl records that go with the edition.

📖 Collovà, Patrizio / Petrini, Dario 1981-82: *Lingua, dialetto e commutazione di codice. Interazioni verbali in un negozio del Luganese. Rivista Italiana di Dialettologia* 5-6, 257-293. Corpus: 259-278.

Comment: See Koch / Oesterreicher (1990: 37)

📖 Cresti, Emanuela 1987: *L’articolazione dell’informazione nel parlato*; in: Accademia della Crusca (ed.): *Gli italiani parlati. Sondaggi sopra la lingua di oggi*. Firenze: Accademia della Crusca, 27-90. No ISBN. Corpus: 72-88.

Comment: See Koch / Oesterreicher (1990: 37s).

📖📧 Cresti, Emanuela (ed.) 2000: *Corpus di italiano parlato. Vol. II: Campioni*. Firenze: Accademia della Crusca. ISBN 88-87850-01-1.

1 Printed and electronic corpus, the electronic versions being included in a CD-ROM that is attached to the volume. 317 pages, approx. 58.000 words. 49 texts. 2 Complete meta-information available. Texts recorded in Firenze and surrounding Tuscany between 1977 and 1998 (with most texts taped in the 90’s). 3 Some texts marked on the diatopic level. As the corpus embraces a high number of conceptually divergent text types and documents very different communicative settings, its significant diastratic and – most important – diaphasic variation comes as no surprise. 4 Orthographic transcription in linear notation following the CHAT conventions (cf. MacWhinney 2000; De Cat / Plunkett in this volume [CD-ROM]). 5 No translation available. 6 Short audio samples (of not more than 2 min length) from selected texts are included as WAVE files on the CD-ROM.

Comment: This corpus volume represents a part of the LABLITA corpus described in detail in the corresponding introductory volume (Cresti 2000). The CD-ROM contains the transcripts in more than one file format (ASCII, RTF; furthermore some samples in specific formats produced by analysis tools).

📖 De Mauro, Tullio *et al.* 1994: *Lessico di frequenza dell’italiano parlato*. Rome: EtasLibri. ISBN 88-453-0574-0. Corpus on two floppy disks enclosed with the book.

Comment: See De Mauro *et al.* (1992) and Schneider (in this volume) for a detailed description.

📖📧 Gavioli, Laura / Mansfield, Gillian 1991: *The PIXI Corpora. Bookshop encounters in English and Italian* (= *Testi e Discorsi*; 10). Bologna: CLUEB. ISBN 88-491-0387-5. Italian corpus: 93-262.

1 Printed and electronic corpus containing English and Italian samples. The following information refers to the Italian sample only: 174 pages, approx. 30.000 words. 142 texts, some of them very short. The electronic version (ASCII files) is available via ftp or download from the *Oxford Text Archive*, free of charge. A written request has to be sent to OTA; see <<http://ota.ahds.ac.uk>> for details. 2 Very few meta-information is given; even the places of the recordings have been anonymized. 3 As the title suggests, all texts are service encounters. No significant markings on any diasystematic level. 4 Orthographic transcription in linear notation. 5 No translation available. 6 No recordings available.

📖 Haase, Martin 1999: *Dialektdynamik in Mittelitalien. Sprachveränderungsprozesse im umbrischen Apenninenraum* (= Romanica et Comparatistica; 33). Tübingen: Stauffenburg. ISBN 3-86057-083-8. Corpus: 338-374.

1 Printed corpus. 36 pages. 15 texts. 2 Complete meta-information is given for each transcription. Recordings were made between 1993 and 1995 in different mountain villages of Central Umbria. 3 Strong diatopic features in the speech of most informants. 4 Orthographic transcription with slight phonetic adjustments, following the recommendations of the *Rivista Italiana di Dialettologia*; linear notation. 5 No translation available. 6 No recordings available.

📖 Rizzi, Elena / Vincenzi, Giuseppe Carlo 1987: *L'italiano parlato a Bologna. Fonologia e morfosintassi*. Bologna: CLUEB. No ISBN. Corpus: 57-114.

Comment: See Koch / Oesterreicher (1990: 41s).

📖 Rovere, Giovanni 1977: *Testi di italiano popolare. Autobiografie di lavoratori e figli di lavoratori emigrati. Analisi sociolinguistica*. Rome: Centro Studi Emigrazione. No ISBN.

Comment: See Koch / Oesterreicher (1990: 41).

📖 Stammerjohann, Harro 1970: Strukturen der Rede. *Studi di Filologia Italiana* 28, 295-397. Corpus: 358-394.

Comment: See Koch / Oesterreicher (1990: 42).

📖🎧 Vicari, Mario (ed.) 1992: *Valle di Blenio. Prima parte* (= Documenti orali della Svizzera italiana. Trascrizioni e analisi di testimonianze dialettali; 1). Bellinzona: Cantone Ticino / Dipartimento dell'istruzione e della cultura. No ISBN. Corpus: 55-178.

1 Printed corpus. 33 pages. 19 texts.

📖🎧 Vicari, Mario (ed.) 1995: *Valle di Blenio. Seconda parte* (= Documenti orali della Svizzera italiana. Trascrizioni e analisi di testimonianze dialettali; 2). Bellinzona: Cantone Ticino / Dipartimento dell'istruzione e della cultura. No ISBN. Corpus: 51-213.

1 Printed corpus. 35 pages. 20 texts.

Comment: For the remainder of the parameters, these two corpora are very similar to the 'Dialecti della Svizzera italiana' corpora, presented above and to which they constitute a continuation, with the exception that these two transcription volumes contain even more detailed meta-information, extremely accurate ethnographic documentation (including drawings and photos), and a linguistic analysis of each text. The recordings are available on vinyl records or audio cassettes alternatively.

9. Corpora of other Italo-Romance languages

9.1. Sardinian

The following three titles form a coherent collection of Sardinian folk tale corpora and are therefore presented together:

🎧📖 Sanna, Enedina (coord.) 1996: *Contami unu contu. Racconti popolari della Sardegna. Vol. 1: Racconti del Logudoro*. Alghero: Archivi del Sud. No ISBN.

1 Audio corpus with recordings on an audio CD and companion transcription booklet. 66 min of recordings. 10 texts.

🎧📖 Sanna, Enedina (coord.) 1997: *Contami unu contu. Racconti popolari della Sardegna. Vol. 2: Baronie*. Alghero: Archivi del Sud. No ISBN.

1 Audio corpus with an audio CD and companion transcription booklet. 68 min of recordings. 16 texts.

🎧📖 Sanna, Enedina (coord.) 1998: *Contami unu contu. Racconti popolari della Sardegna. Vol. 3: Campidano*. Alghero: Archivi del Sud. No ISBN.

1 Audio corpus with an audio CD and companion transcription booklet. 70 min of recordings. 7 texts.

The subsequent information is valid for all *Contami unu contu* booklets:

2 General meta-information is provided for each text together with a folklore-typological comment. Most recordings were made in the 1970's by dialectologists and students from Cagliari University. 3 Strongly marked diatopically but homogeneous on the diaphatic level, with all texts being of the *conte et légende* type. 4 Ortho-

graphic transcription in linear notation. **5** A parallel translation into Italian is available for all texts. **6** Recordings of all transcribed texts are available on the audio CDs, to which the booklets constitute a mere annex.

Comment: Although this corpus is limited to semi-spontaneous narrative text types, the recordings are easily available so that it can be used as a sample of traditional spoken Sardinian and hereby justifies its inclusion in the present survey.

9.2. Reto-Romance

📖 Ebnetter, Theodor / Toth, Alfred 1995: *Romanisch im Boden, in Trin und in Flims* (= Schweizer Dialekte in Text und Ton / Romanisch und Deutsch am Hinterrhein / GR; 7). Zürich: Phonogrammarchiv der Universität Zürich. No ISBN. Corpus: 217-619.

📖 Printed corpus. 174 pages. 19 texts of unequal length.

📖🎧 Solèr, Clau / Ebnetter, Theodor 1983: *Romanisch am Heinzenberg / Mantogna* (= Romanisch...; 1). Zürich: Phonogrammarchiv der Universität Zürich. No ISBN. Corpus: 24-48.

📖 Printed corpus. 22 pages. 22 texts.

📖🎧 Solèr, Clau / Ebnetter, Theodor 1988: *Romanisch im Domleschg* (= Romanisch...; 3). Zürich: Phonogrammarchiv der Universität Zürich. ISBN 3-907538-02-1. Corpus: 148-234.

📖 Printed corpus. 44 pages. 27 texts.

📖🎧 Solèr, Clau 1991: *Romanisch im Schams* (= Romanisch...; 5). Zürich: Phonogrammarchiv der Universität Zürich. ISBN 3-907538-08-0. Corpus: 184-380.

📖 Printed corpus. 84 pages. 45 texts.

Characteristics shared by all “Romanisch und Deutsch am Hinterrhein” corpora are the following:

2 Basic or detailed meta-information available. Recordings made between 1978 and approx. 1983 with the very last fluent speakers in some valleys of Northern Grisons where Reto-Romance has arrived at the ultimate stages of language death. Only (rather short) extracts of the recordings, selected on the basis of their linguistic or ethnographic interest, are transcribed, except for the Ebnetter / Toth (1995) corpus which contains the complete texts. **3** Strong diatopic features corresponding to the dialectologic scope of the series. **4** Phonetic transcription, adapted from the transcription system used in the AIS (*Sprach- und Sachatlas Italiens und der Südschweiz*), and a parallel orthographic transcription in linear notation. **5** A parallel translation into German is provided, resulting in a 3-columns layout of the corpus data (except for the Ebnetter / Toth 1995 volume, where the different versions are printed one after the other). **6** For the volumes marked with the corresponding symbol, recordings are available on audio-cassettes.

10. Corpora of other Romance languages

10.1. Romanian

Cresti (2000: 15) mentions that the Romanian Academy of Sciences is preparing a reference corpus of spoken Romanian but seems to suggest that, for the time being, no other corpora were publicly available. Such a statement cannot be true as soon as dialect corpora are taken into consideration. As a matter of fact, dialectologically motivated collections of spoken language samples have been established in Romania for almost a century now. However, many of these samples do not correspond to ‘spoken language corpora’ in the modern sense and are therefore difficult to evaluate in the light of the descriptive parameters used here (apart from being sometimes difficult to access). For details on the tremendously rich bibliography of Romanian dialect text anthologies, readers are referred to Caragiu Marioțeanu (1989; on dialectologic sources of spoken language data) and to Winkelmann / Lausberg (2001: 1010-1018; on sources related to linguistic atlas projects). Apparently, Cresti is right in deploring the absence of a representative corpus of colloquial non-dialectal Romanian.

☞ Bochmann, Klaus / Dumbrava, Vasile (eds.) 2000: *Limba Română vorbită în Moldova istorică. Vol. 2. Texte*. Leipzig: Leipziger Universitätsverlag. ISBN 3-934565-90-5.

1 Printed corpus. 295 pages, approx. 95.000 words. 48 texts. 2 Detailed meta-information is given in the transcription headers. Most recordings were made in 1997/98 in the cities of Iași (Romania), Chișinău and Bălți (Republic of Moldova). 3 As suggested by the title, this corpus is supposed to document the Moldavian varieties of Romanian as spoken in Northeastern Romania and Bessarabia (now Republic of Moldova); however, texts are ordered according to three levels of formality, so that diatopic marking varies, as diastratic and diaphasic features do. 4 Most texts are transcribed orthographically, with minor adjustments to reflect pronunciation and intonation patterns, using linear notation with score-style alignment where speakers overlap. 6 texts are transcribed phonetically, following the *Atlasul lingvistic român* conventions. 5 No translation available. 6 No recordings available.

Comment: Although this corpus has a clear dialectologic scope, it qualifies as a spoken language corpus in the current sense thanks to the wide range of text types and registers which are documented.

10.2. Romance-based creoles

☞ Bollée, Annegret / Rosalie, Marcel (eds.) 1994: *Parol ek memwar. Récits de vie des Seychelles* (= Kreolische Bibliothek; 13). Hamburg: Buske. ISBN 3-87548-075-9.

1 Printed corpus. 119 pages, approx. 32.000 words. 7 texts. 2 Complete meta-information is given in the header which precedes each text. All recordings made between 1980 and 1983 with elderly informants (60 to 96 years). 3 Diastratically marked due to the chosen age group. As the title of the corpus suggests, most texts are auto-biographical narrations. 4 Orthographic transcription in linear notation. Transcriptions are structured to match thematic units. 5 Parallel translation into French. 6 No recordings available.

The following two small-scale corpora have a common conception and therefore are grouped for a brief mention:

☞ Kriegel, Sibylle 1996: *Diathesen im Mauritius- und Seychellenkreol* (= ScriptOralia; 88). Tübingen: Narr. ISBN 3-8233-4578-8. Corpus: 188-211.

1 Printed corpus. 19 pages. 2 texts.

☞ Michaelis, Susanne 1994: *Komplexe Syntax im Seychellen-Kreol. Verknüpfung von Sachverhaltsdarstellungen zwischen Mündlichkeit und Schriftlichkeit* (= ScriptOralia; 49). Tübingen: Narr. ISBN 3-8233-4264-9. Corpus: 180-211.

1 Printed corpus. 26 pages. 4 texts.

The following characteristics are shared by both corpora:

2 Basic meta-information given. 3 Different, rather unmarked text types. 4 Orthographic transcription in score notation where necessary. 5 No translation available except for one audio-taped radio news program in the Michaelis (1994) corpus. 6 No recordings available.

☞ ☞ Ludwig, Ralph / Telchid, Sylviane / Bruneau-Ludwig, Florence (eds.) 2001: *Corpus créole. Textes oraux dominicains, guadeloupéens, guyanais, haïtiens, mauriciens et seychellois. Enregistrements, transcriptions et traductions* (= Kreolische Bibliothek; 18). Hamburg: Buske. ISBN 3-87548-290-5.

1 Printed corpus. 127 pages. 17 texts. 2 Very detailed meta-information available. The majority of the recordings were made in the (late) 1980's. 3 Text types vary from informal to highly formal oral speech. Some recordings (namely those from Haiti) come from audio-taped radio programs. As the corpus title suggests, diatopic variation is high. 4 Orthographic transcription – respecting the transliteration systems in use in the different Creole-speaking areas – in a highly accurate HIAT score notation. 5 Translations into French following the transcripts. 6 All recordings included on the two audio CDs enclosed in the volume.

☞ Neumann, Ingrid 1985: *Le créole de Breaux Bridge, Louisiane. Etude morphosyntaxique – textes – vocabulaire* (= Kreolische Bibliothek; 7). Hamburg: Buske. ISBN 3-87118-697-X. Corpus: 353-441.

1 Printed corpus. 39 pages. 14 texts. 2 Complete meta-information provided in different sections of the book. Recordings made in 1979/80. 3 Corpus consists of spontaneous casual conversations and semi-spontaneous orally performed folk-tales. 4 Hybrid orthographic transcription which notes many phonetic features; linear notation. 5 Parallel translation into French. 6 No recordings available.

Bibliography

- Bacelar do Nascimento, Maria Fernanda 2000a: Corpus de Référence du Portugais Contemporain; in Bilger (ed.), 25-29.
 — 2000b: O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito; in: Gärtner, Eberhard / Hundt, Christine / Schönberger, Axel (eds.): *Estudos de gramática portuguesa (I)* (= Biblioteca Luso-Brasileira; 12). Frankfurt am Main: TFM, 185-200.
- Barme, Stefan to appear: *Corpus des phonisch-nähesprachlichen Brasilianisch und europäischen Portugiesisch*. Germersheim / Mainz: Gutenberg-Universität / Centro de Estudios Latinoamericanos and Institut für Romanistik.
- Beeching, Kate to appear: *Gender, politeness and pragmatic particles in French*. Amsterdam / Philadelphia: Benjamins.
- Bergounioux, Gabriel / Baraduc, Jean / Dumont, Céline 1992: L'étude socio-linguistique sur Orléans (1966-1991). 25 ans d'histoire d'un corpus. *Langue française* 93, 74-93.
- Bilger, Mireille 1996: Corpus de portugais et d'espagnol. *Revue Française de Linguistique Appliquée* 1, 124-130.
 — (ed.) 2000: *Corpus. Méthodologie et applications linguistiques* (= Les français parlés – Textes et études; 3). Paris: Champion.
- Blanc, Michel / Biggs, Patricia 1971: L'enquête socio-linguistique sur le français parlé à Orléans. *Le Français dans le Monde* 85, 16-25.
- Blanche-Benveniste, Claire / Jeanjean, Colette 1987: *Le français parlé. Transcription et édition*. Paris: Didier.
- Boisvert, Lionel / Laurendeau, Paul 1988: Répertoire des corpus québécois de langue orale. *Revue Québécoise de Linguistique* 17, 241-262.
- Caragiu Marioțeanu, Matilda 1989: Rumänisch. Areallinguistik I. Dakorumänisch / Les aires linguistiques I. Dacoroumain; in: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian (eds.): *Lexikon der Romanistischen Linguistik*. Vol. 3. Tübingen: Niemeyer, 405-423.
- Castilho, Ataliba Teixeira de 1990: O português culto falado no Brasil (História do projecto NURC/BR); in: Preti, Dino / Urbano, Hudinilson (eds.): *A linguagem falada culta na cidade de São Paulo. Vol. IV: Estudos*. São Paulo: T.A. Queiroz / FAPESP, 141-202.
- Coseriu, Eugenio 1988: 'Historische Sprache' und 'Dialekt'; in: Albrecht, Jörn (ed.): *Energie und Ergon. Sprachliche Variation – Sprachgeschichte – Sprachtypologie. Band I: Schriften von Eugenio Coseriu (1965-1987)*. Tübingen: Narr, 45-61.
- Cresti, Emanuela 2000: *Corpus di italiano parlato. Volume I: Introduzione*. Firenze: Accademia della Crusca.
- De Mauro, Tullio et al. 1992: Il lessico di frequenza dell'italiano parlato: LIP; in: Moretti, Bruno / Petrini, Dario / Bianconi, Sandro (eds.): *Linee di tendenza dell'italiano contemporaneo. Atti del XXV Congresso Internazionale di Studi della Società di Linguistica Italiana (Lugano 1991)* (= Pubblicazioni della Società di Linguistica Italiana; 33). Rome: Bulzoni, 83-118.
- Ehlich, Konrad 1993: HIAT: A transcription system for discourse data; in: Edwards, Jane A. / Lampert, Martin D. (eds.): *Talking data. Transcription and coding in discourse research*. Hillsdale NJ: Erlbaum, 123-148.
- Hazaël-Massieux, Marie-Christine 1996: Les corpus créoles. *Revue Française de Linguistique Appliquée* 1, 103-110.
- Koch, Peter / Oesterreicher, Wulf 1985: Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36, 15-43.
 — / — 1990: *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch* (= Romanistische Arbeitshefte; 31). Tübingen: Niemeyer.
- Lope Blanch, Juan M. 1986: *El estudio del español hablado culto: Historia de un proyecto*. México: Universidad Nacional Autónoma de México.
- MacWhinney, Brian 2000: *The CHILDES Database: Tools for Analyzing Talk. Vol. 2: The Data*. Hillsdale, NJ: Erlbaum.
- Marcuschi, Luiz Antônio 1986: *Análise da conversação*. São Paulo: Atica.
 — 2000: Língua falada e língua escrita no português brasileiro: distinções equivocadas e aspectos descuidados; in: Große, Sybille / Zimmermann, Klaus (eds.): *O português brasileiro: pesquisas e projetos* (= Biblioteca Luso-Brasileira; 17). Frankfurt a. M.: TFM, 11-57.
- Quilis, Antonio 1985: El estudio coordinado de la lengua española hablada en Hispanoamérica y en España; in: *Actes du XVIIe Congrès International de Linguistique et Philologie Romanes*, vol. 7. Aix-en-Provence: Université de Provence, 317-328.
- Phonogrammarchiv Zürich (ed.) 2001: Textpublikationen des Phonogrammarchivs 1914-2000. Zürich: Phonogrammarchiv der Universität Zürich (<<http://www.phonogrammarchiv.unizh.ch/publikationen/textpublikationen.pdf>>).
- Pusch, Claus D. 2001: *Morphosyntax, Informationsstruktur und Pragmatik. Präverbale Marker im gaskognischen Okzitanisch und in anderen Sprachen* (= ScriptOralia; 124). Tübingen: Narr.
- Sánchez, Aquilino 1995: *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.
- Stein, Achim 1995: Maschinenlesbare Korpora für das Französische. *Zeitschrift für französische Sprache und Literatur* 105, 1-25.
- Thompson, Henry S. / Anderson, Anne H. / Bader, Miles 1995: Publishing a spoken and written corpus on CD-ROM: the HCRC Map Task experience; in: Leech, Geoffrey / Myers, Greg / Thomas, Jenny (eds.): *Spoken English on computer. Transcription, mark-up and application*. Harlow / New York: Longman, 168-180.
- Voghera, Miriam 1996: Corpora dell'italiano. *Revue Française de Linguistique Appliquée* 1, 131-134.

-
- Winkelmann, Otto / Lausberg, Uta 2001: Romanische Sprachatlanten; in: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian (eds.): *Lexikon der Romanistischen Linguistik. Vol. 1:2: Methodologie / Méthodologie*. Tübingen: Niemeyer, 1004-1068.
- Zeidler, Heidemarie 1982: *Das 'Français fondamental (1er degré)'. Entstehung, linguistische Analyse und fremdsprachen-didaktischer Standort* (= Heidelberger Beiträge zur Romanistik; 12). Frankfurt a. M. et al.: Lang.