A survey of spoken language corpora in Romance

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. 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that sources for authentic oral speech are insufficient, then this is due to the fact that many corpora either remain unpublished or otherwise unavailable.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...
some shortcomings of one or the other kind. 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.

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, ‘grey’ 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. 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. Needless to emphasize that it is sometimes hard to make a clear distinction between ‘dialect’ corpora and ‘general’ corpora which exhibit diatopically marked variation.

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.

---

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 Wiesmeth (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. Estudes de linguistique occitane : Le Couserans (Gasconyne 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.’
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),
2. 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;

3. 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;

4. 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;

5. 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 HAT 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;

6. 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,

7. 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


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).


1 Printed corpus. 3.820 pages, approx. 500.000 words. 85 texts available in separate fascicles. 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,’ meaning ‘published on off-line electronic media (floppy disk, CD-ROM) or downloadable via Internet,’ 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. Orthographic transcription in a simplified score notation. No translation available.

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).


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.


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


• On-line corpus. Approx. 1.1 Mio words and 800 texts. Detailed meta-information available. Marked on all diasystematic levels, but to a different degree depending on the text. Also many generally unmarked texts. Orthographic transcription; a small proportion has also been transcribed phonetically. No translation 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.


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

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


• Printed corpus. 95 pages, approx. 28.000 words. 1 text. 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). Diastratic and moderate diatopic features.


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


A survey of spoken language corpora in Romance (full version)


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'Études 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:


1 Printed corpus. 92 pages. 96 texts, some of them very short. 2 Basic meta-information available. Recordings made in Vauvert (near Nimes) 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.

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.

*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.

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 inline. 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”).

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.  

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

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

A survey of spoken language corpora in Romance (full version)

de l'Estrie. 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.


- Printed corpus. 78 pages, approx. 22,000 words. 36 texts. 
- 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. 
- Strongly marked on the diatopic level. 
- 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. 
- No translation available, but the entire vocabulary is explained in a French-English glossary that follows the corpus. 
- No recordings available.

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


- Printed corpus. 28 pages, approx. 5,500 words. 8 texts, some very short. 
- Only rudimentary meta-information available. 
- Although intended to be a corpus of Tahitian regional French, the data is only moderately diatopically marked. Most texts are interviews. 
- Orthographic transcription in linear notation. 
- No translation available. 
- No recordings available.


- Audio corpus with printed transcription booklet. 22 pages. 23 texts. 
- Basic meta-information available. 
- 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. 
- Orthographic transcription in linear notation. 
- No translation available. 
- 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.


- Printed corpus. 62 pages, approx. 22,000 words. 3 texts. 
- Only rudimentary meta-information available. Elicited conversation between students. 
- Diatopically marked (Acadian French as spoken in Louisiana), with interferences of English. 
- Orthographic transcription with slight adjustments, in linear notation. 
- No translation available. 
- No recordings available.


- Printed corpus. 245 pages. 10 texts. 
- Diatopically marked (Acadian French as spoken in Louisiana), with moderate interferences from English. 
- Orthographic transcription with very few phonetic adjustments. Simplified HIAT-oriented score notation. Intonation patterns marked in-line. 
- Interlinear translation into German. 
- Recordings may be requested from the author.


- 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). 
- Detailed meta-information available. All recordings made between 1972 and 1981 with 3 informants from French-speaking communities in Western Newfoundland. 
- Strongly marked on the diatopic (Newfoundland Acadian and Laurentidien French) and on the diaphatic level (narrative texts of the *conte* type). 
- Orthographic transcription with phonetic adjustments in linear transcription; many texts are monologues. 
- No translation available (only English texts are translated into French). 
- No recordings available.

*Comment*: Linguistically usable corpus with an ethnographic-folkloristic scope.
4. Corpora of other Gallo-Romance languages

4.1. Franco-Provençal


Printed corpus. 81 pages, approx. 40,000 words. 35 texts, arranged in 6 thematic groups. Detailed meta-information available. All recordings made with one informant, an 80 year old farmer and viticulturist from Pélussin (Dép. Loire). Diatopically and diaphatically marked. All texts are narrative in character, some being of the contes et légendes type, but most are auto-biographic. Hybrid phonetically-oriented transcription, not conform to IPA. Parallel translation into French. No recordings available.

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’Études Méridionales of Toulouse; see sect. 3.1 for the French part.


Printed corpus. 105 pages, approx. 20,000 words. 37 texts. 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). Diatopically (Gascony Occitan) and diastratically marked (conversations about traditional breeding activities).


Printed corpus. 122 pages, approx. 32,000 words. 2 texts. 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). Diatopically (occitan languedocien) and diastratically marked (workmen’s colloquial speech).


Printed corpus. 24 pages, approx. 6,000 words. 1 text. Very detailed meta-information given. The recording was made in 1973 with a 62-year old wine-grower. Diatopically marked (Languedoc Occitan).

The following characteristics are shared by all of these corpora:

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


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


Electronic corpus (PDF files). 370 pages. 13 texts. Depending on the text, basic or complete meta-information is available. Recordings made in different places of Gascony (Southwestern France) in 1996; some audiotaped radio programs. 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

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).
A survey of spoken language corpora in Romance (full version)

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


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


Printed corpus. 357 pages, 178.000 words. 13 texts. 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. Strongly marked on the diastratic level, but only very slightly marked diatopically. All texts are semi-guided conversations between peers. Orthographic transcription in a linear notation following TEI recommendations for text-descriptive mark-up and including also information on extra-linguistic events. No translation available. No recordings available.

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


Printed corpus. 152 pages, approx. 55.000 words. 9 texts. 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. All texts represent casual conversations and are slightly marked diatopically (some interferences from Catalan). Orthographic transcription with phonetic adjustments in a score-like notation. Prosodic features annotated in-line. No translation available. No recordings available.

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


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


Printed corpus. 43 pages, 6,300 words. 8 texts. Basic meta-information provided. Recordings made with informants from different parts of Spain but who have been living in Germany for some time. Generally unmarked speech. Orthographic transcription in linear notation. All texts are monologues. Parallel translation into German. No recordings available.

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 diastratic levels.


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 Sevillan texts are included in the Samper Padilla et al. (1998) Macrocorpus on CD-ROM described in section 5.3.:


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


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


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.

*Comment:* See Koch / Oesterreicher (1990: 45s).


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


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


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


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


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


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).


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


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

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

- 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. As one may expect from a project stretching over the entire Spanish-speaking world, there is significant diatopic variation, but other diachronic 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.
- 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. No translation available.
- No recordings available.


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

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:
5.3. Corpora including both European and extra-European Spanish

  - 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. Basic meta-information available in the files’ headers, which may be visualized after having completed a search through the web interface. 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. Orthographic transcription. Files seem to be TEI-encoded. No translation available. 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).


Electronic corpus on CD-ROM. Approx. 720,000 words. 168 texts. Only very basic meta-information available. 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. Strict orthographic transcription in linear notation. No translation available. 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.

Electronic corpus on CD-ROM. Approx. 400,000 words from oral sources. No meta-information available. 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. Orthographic transcription. No translation available. 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


Printed corpus. 230 pages, approx. 100,000 words. 140 texts. 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. All texts are interviews that are slightly to moderately marked on the diatopic and diastratic level. Orthographic transcription in linear notation. No translation available. 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>.


- 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.
- No significant diastratic marking.
- Orthographic transcription in linear notation. Some non-TEI conform tags are used for text structure encoding.
- No translation available.
- No recordings available.

### 6.2. Brazilian Portuguese


- Electronic corpus (HTML files). Approx. 180.000 words. 37 texts.
- 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.
- All texts are interviews. Slightly to moderately marked on the diastratic levels.
- Orthographic transcription in linear notation.
- No translation available.
- 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 (overseas) 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.

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

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

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

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

- Printed corpus. 212 pages, approx. 60.000 words. 9 texts.
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-tj/>.

1 Printed corpus. 198 pages, approx. 40,000 words. 8 texts.

1 Printed corpus. 252 pages, approx. 75,000 words. 12 texts.

1 Printed corpus. 150 pages, approx. 90,000 words. 12 texts.

The following characteristics are shared by all NURC corpora:
- Basic, sometimes detailed meta-information given in the headers that precede the transcription texts.
- Texts marked (slightly) as far as diatopic variation is concerned, while other diasyntactic features are weak.
- Orthographic transcription in linear notation.
- No translation available.
- 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. General meta-information available. Texts were recorded in 1986/87 in different parts of the city of Fortaleza. Texts marked on the diatopic and diastratic level, but homogeneous on the diaphatic level. All are semi-guided interviews. Orthographic transcription in mostly linear notation, with passages noted score-like. The notation system as proposed by Marcuschi (1986) has been applied. No translation available. 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

1 Electronic corpus (ASCII files) on four CD-ROMs. 92,000 words. 86 texts. 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. 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. Strict orthographic transcription in linear notation. No translation available. 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 Barme (to appear).
7. Corpora of other Ibero-Romance languages

7.1. Catalan

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 focusing 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.

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.

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 Montuïri (Majorca). 3 Strong diatopical 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.

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 diatopical 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.

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.

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.

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.
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.


Comment: See Koch / Oesterreicher (1990: 36).


Electronic corpus (ASCII and WAVE files). Approx. 40.000 words. 15 texts. 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. 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). Orthographic transcription in linear notation. No translation available. The recordings of all AVIP texts are downloadable in WAVE format.


Comment: See Koch / Oesterreicher (1990: 36)


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

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.
\[\text{Printed corpus. 16 pages. 13 texts.}\]

Leissing-Giorgetti, Sonja / Vicari, Mario (eds.) 1975: Valle Onsernone – Centovalli – Valle Verzasca TI (= Dialetti della Svizzera italiana; 3). No ISBN.
\[\text{Printed corpus. 27 pages. 14 texts.}\]

Vicari, Mario (ed.) 1978: Locarnese – Terre di Pedemonte TI (= Dialetti della Svizzera italiana; 4). No ISBN.
\[\text{Printed corpus. 32 pages. 16 texts.}\]

Vicari, Mario (ed.) 1980: Valle Riviera – Bellinzonese TI (= Dialetti della Svizzera italiana; 5). No ISBN.
\[\text{Printed corpus. 29 pages. 32 texts.}\]

Vicari, Mario (ed.) 1983: Malcantone (Cantone Ticino) (= Dialetti della Svizzera italiana; 6). No ISBN.
\[\text{Printed corpus. 39 pages. 13 texts.}\]

The following characteristics are shared by all “Dialetti della Svizzera italiana” corpora:
\[\text{Detailed meta-information is given. Recordings were made in the late 1960’s and the 1970’s. Diatopic variation, \textit{corresponding to the collection’s objective, is obvious. Diastratic variation, however, is weak. Text types vary, with most texts belonging to the autobiographic \textit{récits de vie} genre. Orthographic (dialectal) transcription and a parallel phonetic transcription in linear notation. A parallel translation into Standard Italian is provided, resulting in a 3-columns layout of the corpus data. Recordings of all transcribed texts are available on the vinyl records that go with the edition.}\]

\[\text{Comment: See Koch / Oesterreicher (1990: 37)}\]

\[\text{Comment: See Koch / Oesterreicher (1990: 37s).}\]

\[\text{Printed and electronic corpus, the electronic versions being included in a CD-ROM that is attached to the volume. 317 pages, approx. 58.000 words. 49 texts. Complete meta-information available. Texts recorded in Firenze and surrounding Tuscany between 1977 and 1998 (with most texts taped in the 90’s). Some texts marked on the diatopic level. As the corpus embraces a high number of conceptually divergent text types and documents very different communicative settings, its significant diastratic and – most important – diaphasic variation comes as no surprise. Orthographic transcription in linear notation following the \textit{chat} conventions (cf. MacWhinney 2000; De Cat / Plunkett in this volume [CD-ROM]). No translation available. Short audio samples (of not more than 2 min length) from selected texts are included as WAVE files on the CD-ROM. Comment: This corpus volume represents a part of the \textit{Lablita} corpus described in detail in the corresponding introductory volume (Cresti 2000). The CD-ROM contains the transcripts in more than one file format (ASCII, RTF; furthermore some samples in specific formats produced by analysis tools).}\]

\[\text{Comment: See De Mauro \textit{et al}. (1992) and Schneider (in this volume) for a detailed description.}\]

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

9.1. Sardinian

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

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

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

  - 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:

- 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.
- Strongly marked diatopically but homogeneous on the diaphatic level, with all texts being of the *conte et légende* type.
graphic transcription in linear notation. A parallel translation into Italian is available for all texts. 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


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

1. 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 Ebneter / Toth (1995) corpus which contains the complete texts.
2. Strong diatopic features corresponding to the dialectologic scope of the series.
3. 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.
4. A parallel translation into German is provided, resulting in a 3-columns layout of the corpus data (except for the Ebneter / Toth 1995 volume, were the different versions are printed one after the other).
5. 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 Mariojeanu (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 deploiring the absence of a representative corpus of colloquial non-dialectal Romanian.
A survey of spoken language corpora in Romance (full version) 21


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 Iaşi (Romania), Chişinău and Bălţi (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 diastatic 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


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 autobiographical 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:


1 Printed corpus. 19 pages. 2 texts.


1 Printed corpus. 26 pages. 4 texts.

The following characteristics are shared by both corpora:

1 Basic meta-information given. 2 Different, rather unmarked text types. 3 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.


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.


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

Beeching, Kate to appear: Gender, politeness and pragmatic particles in French. Amsterdam / Philadelphia: Benjamins.