Johannes Kabatek (Tübingen)
Claus D. Pusch (Freiburg im Breisgau)
Wolfgang Raible (Freiburg im Breisgau)

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

One of the main achievements of structuralism is said to be the distinction between diachronic and synchronic linguistics and the preference given to the synchronic perspective, thus shifting the interest of linguists from the description and explanation of instability in language towards an approach that views language as a rather stable system with in-built complex but balanced functionalities; a view that only the synchronic snapshot may catch. This approach has been further developed in subsequent schools of structuralist linguistics, namely in generative theories which have arrived at highly sophisticated formal representations of how language functions at a given point of time. These formalizations, although capable of further development and undergoing indeed constant revision and updating, are by themselves rather static when trying to model the processes which are going on when language is used.

Such approaches to language in the structuralist tradition face a crucial dilemma: in order to be able to describe language as a *système où tout se tient*, they have to postulate a state of language devoid of its quite obvious inherent dynamics and tendencies of change. How should we conceive of the interrelation between the layers of a well-organized language system identified by the structuralist in the diachrony of a given language? How can axiomatic stability and empirical instability be reconciled?

Some linguists try to resolve this dilemma identifying variation as the culprit leading to language change while others emphasize language acquisition and its different stages as reminiscence of former linguistic changes before reaching the comparatively stable system that the synchronicist is interested in. The problem faced by structuralism regarding language change may in part be attributed to the fundamental question of what counts as evidence in linguistics (cf. Brend / Sullivan / Lommel eds. 2002; Penke / Rosenbach 2004). Traditionally, the empirical basis for synchronic structuralism, particularly within the generative paradigm, used to be above all grammaticality-judgement data gathered through introspection, elicitation or consultation of native speakers of the language. However, these forms of data collection are restricted to *hic et nunc* stages of the linguist's or informant's language. Even relatively recent stages of language are inaccessible to introspection or elicitation. Who is able to judge a given construction produced ten or twenty years earlier in our L1 as grammatical? Who could decide whether a given morphologic or

lexical unit of our mother tongue was equally frequent in our childhood or adolescence?

Whereas judgment data lack reliability for stages of the language that lie only some years back, they are clearly beyond our competence for earlier stages of a language. However, formally oriented structuralists as well as functional linguists, namely those whose interest lies in the area of cognitive grammar, face the problem of the validity of data. Functionalists are by definition more at ease with structural and competence-based variation,[1] but – as with all scientific scholars – their aim is to formulate generalizations and to test their validity, which requires them to make the same critical reflections about the nature and the treatment of the data these generalizations are based upon.

These reflections are particularly necessary for cognitivists approaching language under the heading of pan-chronicity thus overcoming the Saussurean dichotomy emphasized at the beginning; grammaticalization theory is a case in point. Although grammaticalization theory wants to reconcile (synchronic) stability and (diachronic) instability of the structures found in language, important issues such as the status of markedness and frequency are still controversial (cf. Bybee ed. 2001). However, the point is again that the cognitive processes that might explain how language works are already difficult to access when recent stages of a given language are taken into account; they are inaccessible – and assumptions about these processes and conceptualizations become speculative – when we talk about remote stages of the language.

Diachronic linguists in the philological tradition have always been aware of the fact that their evidence is mostly indirect and heterogeneous but necessarily related to texts; and, according to O. Fischer (2004), contemporary diachronicists working within current linguistic paradigms should take the same stand: "The historical linguist", this author emphasizes, "has only one firm source of knowledge and that is the historical documents [sic]." And she continues:

> His prime concern is an accurate description of the data in the documents, in their context, which he investigates in order to understand the regularities underlying the data and the changes that take place. In doing this he may (or rather, must) make use of insights provided by other disciplines. He must not only turn to language acquisition studies, but also to sociolinguistics, generative grammar, cognitive grammar, discourse analysis, optimality theory etc., and in addition make use of insights drawn from synchronic variation and typological comparison [...]. These insights, however, are not *primary* data to be used in the same way as the written documents themselves. (Fischer 2004: 730; italics as in the original text)

The idea of making use of language data that have been produced independently from the linguist who uses them with a certain guiding interest in mind,

1 See, for instance, Raible (2004), who shows that variation on a given level of language can be seen as having a common denominator, or a unifying principle, in an invariant on the next higher level, thus presupposing a hierarchy of levels with realizations subject to a variation that corresponds to a fundamental principle of any language system, viz. creativity.

and which are therefore external to the linguistic analyses as such, is the very basis of what corpus linguistics is about; and the inevitability of external sources and indirect evidence in the case of remote stages of language(s) makes corpus linguistics a most suitable – one would almost claim: the only suitable – methodological option for diachronic studies.

A clarification seems to be adequate at this point: unlike Fischer in the above-quoted passage, we have avoided to speak of 'historical linguistics' up to now, preferring the term of 'diachronic linguistics'. This choice is voluntary. Even if sometimes diachronic linguistics is understood to be the comparative study of synchronic cuts at different moments in the history of a language as opposed to language history in a broader sense, including external aspects, it is widely accepted that 'historical linguistics' evokes studies dedicated to remote stages of a given language, or analyses embracing long periods of time, starting maybe at this language's origins and ending at the present stage. Although our claim about the particular suitability of a corpus-based approach holds also for this type of diachronic study, which may be characterized as *longue durée*-diachrony, the before-mentioned problems faced by formalist and functionalist linguists when studying less remote layers of language and even its recent stages, indicate that there is no real *methodological* difference between diachronic research irrespective of its extension in time. Obviously, there *are* differences, on both the data side and the expectable results, according to the time span taken into account: long-term change may profoundly modify the language involved, even altering its typologic structure. In this case a corpus-linguistic treatment requires databases with an important historical depth including evidence that comes restricted to written sources. Mid- and short-term change, on the other hand, will manifest itself in a more subtle way. Data bases used for such a purpose may be historically flatter. Nevertheless, they have to be carefully designed as far as conceptional and medial parameters of the included text types are concerned. It goes without saying that this proviso is not limited to diachronic studies of this mid- and short-term type, but as the availability of text material is an ever increasing limiting factor as one goes back in time, conceptional and textual variation and representativity is less easily achieved for remote stages of the language.

Availability turns out to be a limiting factor for corpus-based methodological approaches in studies of language change in another respect: although diachronic linguistics has always and necessarily been much more oriented towards (written) textual data than synchronic linguistics, diachronic corpora in the modern sense of the term, i.e. machine-readable, structured collections of texts of different age within the selected diachronic level (short-, mid- or long-term), do not seem to be more readily available than corpora of synchronic (spoken and written) language data, at least as far as Romance linguistics is concerned. There is, however, a basic stock of completed diachronic text corpora for a vast field of Romance languages (cf. Lucía Megías 2002 – the currently most exhaustive guide to historical texts in electronic form, although limited to literary genres –, but also Kunstmann 2000 – confined to Old and Middle French but less restrictive concerning text types –,

Kunstmann / Martineau / Forget eds. 2003 and Pusch forth.), and, as will become clear from the contributions in the first section of this volume, an impressive and ever increasing number of projects are currently underway or near completion.

It is worth noting that the problematic issues, which in part explain the delay of the Romance research community in gaining access to electronic diachronic corpora, are quite similar to the challenges presented by the machine-readable documentation and presentation of spoken language data as illustrated in the first volume of this series (cf. Pusch / Raible eds. 2002): again, questions of classification of texts, of conceptual balancing (namely regarding text types and discourse traditions; cf. Kabatek forth.), of inclusion of contextual meta-data, issues related to the handling of polymorphism and the definition of adequate mark-up conventions, especially concerning tag-set design, and unresolved technical and juridical problems of unrestricted public accessibility of the data prevail in the discussion.

Following the basic structure of the first specimen in our series, which has just been alluded to, the present volume contains two sections, the first of which is devoted to methodological and technical issues, with special emphasis on recent and current diachronic corpus projects and tools for creating or analyzing such corpora.

The first contribution by Christiane MARCHELLO-NIZIA serves as an introductory chapter to that section, where the author describes from her own experience as a long-time diachronic linguist the transition from manual to computer-aided methods of analyzing historical texts, illustrating the new possibilities brought about by technical progress with the example of demonstrative pronouns in former stages of French. Dieter WANNER, in the following paper, has chosen examples from Old Spanish in order to show how hitherto inaccessible new insights can be achieved through the use of electronic corpora, but he also demonstrates how the technical insufficiencies of currently available diachronic corpora limit their use for certain research questions. Harald VÖLKER, in a stimulating and somehow provocative contribution, argues that hypertextuality – typically associated with modern electronic texts, particularly when published in the World Wide Web – is not a feature brought about by the electronic media but is inherent to any, and more specifically: to any historical, document.

The next two articles deal with the use of parallel translation corpora in diachronic linguistics: Peter STEIN and Georg A. KAISER both consider such data as useful for comparative studies of Romance languages but as particularly suitable for syntactic studies; they advocate the use of largely distributed and translated texts such as Livy's historiographic work *Ab urbe condita* (Stein) or the Bible (Kaiser) to assure representativity and pan-Romance coverage. Harald THUN addresses the important issue of how to gain access to the spoken manifestations of former stages of a language and examines the reliability of orality in literary texts, with the example of early 20th century Argentinean Gaucho literature at hand.

After these five papers having dealt with more general methodological questions, we turn to specific corpora and analysis tools in the individual Romance languages, advancing, within these individual languages, from more remote to more recent stages. This section is led off with Alexei LAVRENTIEV's paper on the encoding problems that had to be resolved for the *Banque du Français Médiéval* project, an XML-based Old French corpus already mentioned in Marchello Nizia's contribution. The articles by Anne-Christelle MATTHEY, Achim STEIN / Martin-D. GLESSGEN and Martin-D. GLESSGEN / Matthias KOPP present a corpus of Old French charters called *Les Plus Anciens Documents Linguistiques de la France*, for which the software package TUSTEP (*TUebinger System von TextverarbeitungsProgrammen*) has been put to use. Gleßgen, Stein and Kopp concentrate on the specificities of semi-automatic lemmatizing of Medieval texts and on the purpose-built tool PHOENIX, whereas Matthey illustrates the query and analysis options of this corpus through the study of selected charters from North-Eastern France. Hiltrud GERNER, from the ATiLF (former: INaLF) research group, presents the current state of a Middle French lexicographic corpus, the *Base des Lexiques du Moyen Français*. Patrice BRASSEUR uses the linguistic atlas of Normandy as an example in order to show how an electronic data-base of lemmatized lexical types as documented in sources such as linguistic atlases may render these dialectologic sources more accessible for new research perspectives.

The following two papers are dedicated to Italian corpora elaborated in order to document language change on two different diachronic levels. Paul VIDESOTT describes the design and constitution of his *Corpus Scriptologicum Padanum*, a collection of non-literary texts in Northern Italian varieties from the earliest testimonies up to the 16th century. Antonietta SCARANO and Sabrina SIGNORINI describe two oral corpora that will allow the comparison of spoken Italian at a distance of three decades: the electronic version of the so-called *Stammerjohann* corpus, created in the 1960's (the first modern Italian spoken-language corpus in general), and the recently published C-ORAL-ROM corpus from the 1990's.

The eight subsequent papers treat corpora and corpus-linguistic tools for the Ibero-Romance languages. Mark DAVIES illustrates the structure of his impressive 100-million word tagged corpus of Spanish that extends from the origins of the language up to the 20th century and constitutes the most comprehensive electronic resource for the diachronic study not only of Spanish, but of any Romance language. Davies gives examples for query strategies which allow convenient and extremely rapid searches in this freely accessible data-base accessible through the WWW. Gloria CLAVERÍA and Joan TORRUELLA explain the creation of a lexicographic database of a somehow different kind: in the context of an electronic edition of the leading Spanish etymological dictionary *Diccionario Crítico Etimológico Castellano e Hispánico* by Joan Corominas, these authors are organizing the documentary data used by the Catalan philologist for his dictionary in a relational database that will allow easy access to the textual evidence underlying Corominas' lexical entries. Giorgio PERISSINOTTO shows how his data-base of documents illus-

trating the material culture of the Southwestern parts of today's United States of America can give a vivid picture of daily life when this area in the late 18th and early 19th century was under Spanish rule.

While Perissinotto's contribution already goes beyond the mere description of the corpus data and exemplifies the results that can be drawn from these data, the following article by the TICA (*Tractament Informàtic del Català Antic*) research group is again more technical in scope and presents the functions of a finite-state-based lemmatizer tool designed primarily for the morphologic analysis of Old Catalan but which may be applied to other stages of the language (or other languages) as well. Joan VENY and Àngels MASSIP, in the second article devoted to Catalan, describe the elaboration of a database containing samples of non-literary dialectal texts from all major Catalan dialects, embracing the period between the 15th and the 20th century. In the last paper of this unit, the Barcelona-based "*Llengua i Publicitat*" research team explains the structure and scope of a media corpus containing TV advertising spots collected over a period of 10 years and its possible uses for studies of short-term language change in a regional language as is Catalan.

The subsequent contribution by Eckhard BICK and Marcelo MÓDOLO gives a general overview of a journalistic corpus that forms part of an on-going mega-corpus project, the *Projeto Para a História do Português Brasileiro*. The partial corpus presented here includes editorials and readers' letters published in 19th century Brazilian newspapers. Apart from the data, its lemmatization with the PALAVRAS tool is described. Marisol LÓPEZ treats the conception and the data collection for the CORGA corpus, a 25-million word reference corpus of modern Galician which, once completed, will document this regional language of Spain in its written and spoken form between the mid-1970's and the beginning of the 21st century. The author highlights the problem of encoding graphic variation in the corpus texts, a challenge for any corpus representing remote stages of a language but, as the CORGA example proves, although a difficult issue for contemporary corpora of languages for which the process of corpus planning is still on-going.

The final article in the first section of the present volume, signed by Steven DECORTE, Tilly DUTILH-RUITENBERG and Truus KRUYT, goes beyond the geographical boundaries of Romance and describes a corpus designed for long-term diachronic studies in a Germanic language, Dutch. The authors give detailed explanations of the problems they faced when deciding upon the tagset to be used for the mark-up of this multi-modular corpus documenting Dutch between the 8th century and the present.

It should have become obvious from this cursory overview of the contributions included in the first section of this book that it is not always possible to clearly separate the technical and methodological presentation of the data and the corpus-linguistic resources, on the one hand, and the aims and results of the diachronic studies and analyses for which these resources have been gathered, on the other. As a matter of fact, although some of the corpus projects and tools described here are self-contained, most data-bases and text collections have been purpose-built with a specific research interest in mind; a

circumstance that does not necessarily reduce the usefulness of these data for and their applicability to other than the original research target. However, one should bare in mind this sometimes intricate relation between the data and these research questions when passing on to the second section of the volume. This second section focuses primarily on selected historical periods or on individual aspects of short-, mid- and long-term language change in Romance languages, which are analyzed on the basis of either publicly available diachronic corpora or, again, purpose-built original data collections. Contrary to the first section of the book, the articles in this section are not generally aligned according to languages but grouped together in respect to thematic criteria.

The first contribution in this section deals with the language out of which all contemporary Romance tongues evolved, i.e. Latin: Viara BOUROVA re-examines, on the basis of the ample *CLCLT–5 Library of Latin Texts* corpus, the uses of the 'infinitive + HABERE' periphrasis from which the Romance conditional form may have originated. Johanne PEEMÖLLER is interested in the question to which extent graphical variation and scriptural instability in ancient manuscripts allows hypotheses about language change that was underway at the moment of the creation of these texts, using the earliest textual sources of French to support her claims. Pierre KUNSTMANN also uses Old French texts to exemplify the types of variation that the diachronic linguist might find when analyzing ancient documents with corpus-linguistic methodology and tools; his focus is on adjectives and pronouns expressing indefinite quantity. The subsequent two papers remain in the realm of Gallo-Romance and discuss the issue of the word-order typological status of these languages. Martin G. BECKER raises once more the much debated question of whether French originally was a verb-second language that changed, through the Middle French period, to rigid SVO order, and, after close examination of the *Corpus d'Amsterdam*, comes to a highly differentiated conclusion. Ioanna SITARIDOU's article is devoted to the behavior of Old French regarding the pro-drop parameter – an issue intimately related to the often suggested but controversial V2 character of this language stage – and compares it to that of Old Occitan, a much more prototypical null-subject language.

The following five contributions analyze aspects of the puzzling grammar of Romance pronouns in different Gallo- and Ibero-Romance languages: Andrés ENRIQUE-ARIAS studies the position of object clitics with respect to finite and non-finite verb forms in the history of Spanish and argues for their status having changed from pronominal to affixal, the whole process being driven, among other things, by prosodic constraints. Susann FISCHER extends these reflections to Old Catalan and – starting from a generative theoretical base – concludes that functional reasons related to information structure play an important role in clitic placement tendencies in this language, too. Her contribution includes some fundamental thoughts on the status of corpus data, and diachronic corpus data in particular, in Generative Grammar. Marc-Olivier HINZELIN provides us with a study of the distributional patterns of clitic pronouns comparable to that of Fischer but dedicated to Old Occitan; the author detects parallelisms between Old Occitan and Old Ibero-Romance, although the available data do not seem to allow a definite conclusion.

The two successive contributions by Ana Maria MARTINS and Mark DAVIES regard the multi-functional pronominal element *se* in Portuguese and Spanish. Martins asks whether the impersonal *se* constructions in Modern Portuguese should be interpreted as active constructions, as opposed to the passive *se* construction in Old Portuguese from which they originated. In this context, she emphasizes the importance of contemporary variational data from dialectal and other non-standard corpora to enlighten the diachronic pathways of pronominal *se*. Davies takes into account the whole range of constructions that involve pronominal *se* in the history of Spanish, including impersonal, decausative and causative, "root reflexive" and modalizing uses, and retraces their quantitative development through the last centuries on the basis of the 100-million word *Corpus del Español*, already presented by the author in his contribution to the first section of this volume.

In the paper signed by Cristina BOSCO and Carla BAZZANELLA, we leave the area of pronominal grammar and turn to Italian discourse markers. Using the data contained in the *Padua* / *Italant* corpus of 13th- and 14th-century Italian, and other corpora of old and contemporary Italian, the authors describe, as a case study for diachronic subjectification, or modal shift, according to their terminology, the development of the meaning of *allora* from temporal and correlative to modal. The following three articles have in common that they shed light on relations of determination and specification which exist within the Romance noun phrase. Elisabeth STARK, on the basis of a selection of historic texts stretching from 13th- to 16th-century Italian, argues for the development of a nominal classification system in this (and other) Romance language(s) as a corollary to the loss of other previously marked features of nominal inflection, with this process involving the grammaticalization of the indefinite and the partitive article. Anne Moseng KNUTSEN and Katja PLOOG's paper is devoted to the multifunctional element *là* in French, which in the African French variety of Abidjan / Ivory Coast is developing from the two-fold use as adverb and localizing nominal determiner towards a three-fold system which includes an additional function of post-nominal *là*, that of a marker of definiteness. Chad LANGFORD and Kathleen M. O'CONNOR investigate the diachronic changes of the placement of color adjectives in the Modern French noun phrase with a special focus on the four color adjectives white, black, green and gray, for which a detailed quantificational analysis elaborated on the basis of the ATILF / INALF research group's FRANTEXT corpus is provided. The authors conclude that an overall decline in pre-nominal color adjective placement is observable but that this is not a steady and even decline and that the different color adjectives do not show a totally parallel development either.

The following two articles deal with word formation and, more specifically, with denominal and deverbal derivation. Elmar EGGERT tries to identify, on the basis of a huge toponymical corpus, the regularities of formation and adaptation of French adjectives derived from place names of towns and villages which designate the inhabitants of these settlements. Alfonso GALLEGOS treats deverbal nominalization in technical registers of Spanish between its

classic and modern period and pays special attention to the nominalizing affix *-do*. The author insists on the fact that the genesis and diachronic expansion of this affixal marker can only be understood when the historical context of the texts where this element appears, and the internal and external features of the underlying discursive tradition are taken into account.

The notion of discourse traditions and the dynamics and inertia which they bring about in the respective text types, are also crucial for Andreas WESCH who describes administrative and juridical texts written during Spain's colonial expansion into the Americas in the 15th to 17th century from the perspective of historical pragmatics. Wesch's paper is the first of four which deal not only with variation and change in time but also in space and, as far as Wesch's and the two subsequent articles are concerned, with variational patterns that emerge when European varieties of Romance languages are compared to their overseas counterparts. Andre KLUMP analyses the phonetic features of the Spanish spoken in the 16th and 17th century on the Caribbean island of Hispaniola, confirming thereby – albeit cautiously – the hypothesis of the Andalusian origin of these sound features. Frank JODL's contribution focuses on the morphosyntax of European and Brazilian Portuguese; the author claims that the contemporary divergent distribution of the future subjunctive forms in the protasis of conditional clauses in these two varieties is significantly different from the uses in earlier stages of this language and that Brazilian still retains these older patterns. Carsten SINNER's paper also emphasizes morphosyntactic features and presents a longitudinal research project which aims at describing the on-going changes in a contact variety of European Spanish, i.e. the Spanish spoken in Catalonia, in comparison with non-contact Castilian varieties of that language.

The volume concludes with a contribution on the historiography of linguistic thought, as Christophe REY investigates the intertextual relations that exist between different encyclopedic works of the period of the French Enlightenment, dedicated to the phonetic description of languages.

Most of the contributions assembled here are revised and updated versions of papers presented at the *2nd Freiburg Workshop on Romance Corpus Linguistics* organized by the Department of Romance Languages of Albert-Ludwigs University at Freiburg im Breisgau in mid-September 2003. The editors owe a debt of gratitude to the German Research Council DFG (*Deutsche Forschungsgemeinschaft*, Bonn) and to the rectorate and the International Office of Albert-Ludwigs University for their financial support without which the workshop could not have been carried out. We would also like to express our thanks to Susan Flocken (Freiburg) and Perrine Wieber (Paris / Freiburg) for their help in the editorial process, and to our interlocutors at Narr Publishing, Gunter Narr and Jürgen Freudl, with whom producing this volume was "business as usual", i.e. a smooth and pleasant affair.[2]

2 A companion web page, where additional material to some articles of this volume are available and where updates to the published contributions may be posted, is available on the <http://www.corpora-romanica.net> web site.

References

Brend, Ruth M. / Sullivan, William J. / Lommel, Arle R. (eds.) 2002: *What constitutes evidence in linguistics?* (LACUS Forum; 28). Houston TX: The Linguistic Association of Canada and the United States.

Bybee, Joan (ed.) 2001: *Frequency and the emergence of linguistic structure* (Typological Studies in Language; 45). Amsterdam / Philadelphia: Benjamins.

Fischer, Olga 2004: What counts as evidence in historical linguistics? *Studies in Language* 28, 710–740.

Kabatek, Johannes forthcoming: Tradições discursivas e mudança lingüística; to appear in: Lobo, Tânia (ed.): *Para a história do Português Brasileiro*. Salvador: EDUFBA (also available at <http://www.kabatek.de/discurso/itaparica.pdf>).

Kunstmann, Pierre 2000: Ancien et moyen français sur le Web : textes et bases de données. *Revue de Linguistique Romane* 64, 17–42.

— / Martineau, France / Forget, Danielle (eds.) 2003: *Ancien et Moyen Français sur le web. Enjeux méthodologiques et analyse du discours*. Ottawa: Editions David.

Lucía Mejías, José Manuel 2002: *Literatura románica en internet. T. 1: Los textos* (Guia de recursos en internet; 1). Madrid: Castalia.

Mukherjee, Joybrato 2002: The scope of corpus evidence; in Brend / Sullivan / Lommel (eds.), 103–114.

Penke, Martina / Rosenbach, Anette 2004: What counts as evidence in linguistics? *Studies in Language* 28, 480–526.

Pusch, Claus D. forthcoming: Els corpus diacrònics de les llengües romàniques. Aspectes tipològics, metodològics i tècnics; to appear in *Caplletra*.

— / Raible, Wolfgang (eds.) 2002: *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache / Romance corpus linguistics: corpora and spoken language* (ScriptOralia; 126). Tübingen: Narr.

Raible, Wolfgang 2002 [2004]: El espacio y el juego de la variación en el lenguaje. *Función* 25–26, 11–20.