Manual of Romance Sociolinguistics
MRL 18

Manuals of Romance Linguistics

Manuels de linguistique romane
Manuali di linguistica romanza
Manuales de lingüística románica

Edited by
Günter Holtus and Fernando Sánchez Miret

Volume 18

Manual of Romance Sociolinguistics

Edited by
Wendy Ayres-Bennett and Janice Carruthers

DE GRUYTER

ISBN 978-3-11-037012-6
e-ISBN (PDF) 978-3-11-036595-5
e-ISBN (EPUB) 978-3-11-039433-7

Library of Congress Control Number: 2018934550

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at: http://dnb.dnb.de.

Typesetting: jürgen ullrich typosatz, Nördlingen
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Manuals of Romance Linguistics

The new international handbook series *Manuals of Romance Linguistics* (*MRL*) will offer an extensive, systematic and state-of-the-art overview of linguistic research in the entire field of present-day Romance Studies.

MRL aims to update and expand the contents of the two major reference works available to date: *Lexikon der Romanistischen Linguistik* (*LRL*) (1988–2005, vol. 1–8) and *Romanische Sprachgeschichte* (*RSG*) (2003–2008, vol. 1–3). It will also seek to integrate new research trends as well as topics that have not yet been explored systematically.

Given that a complete revision of *LRL* and *RSG* would not be feasible, at least not in a sensible timeframe, the *MRL* editors have opted for a modular approach that is much more flexible:

The series will include approximately 60 volumes (each comprised of approx. 400–600 pages and 15–30 chapters). Each volume will focus on the most central aspects of its topic in a clear and structured manner. As a series, the volumes will cover the entire field of present-day Romance Linguistics, but they can also be used individually. Given that the work on individual *MRL* volumes will be nowhere near as time-consuming as that on a major reference work in the style of *LRL*, it will be much easier to take into account even the most recent trends and developments in linguistic research.

MRL's languages of publication are French, Spanish, Italian, English and, in exceptional cases, Portuguese. Each volume will consistently be written in only one of these languages. In each case, the choice of language will depend on the specific topic. English will be used for topics that are of more general relevance beyond the field of Romance Studies (for example *Manual of Language Acquisition or Manual of Romance Languages in the Media*).

The focus of each volume will be either (1) on one specific language or (2) on one specific research field. Concerning volumes of the first type, each of the Romance languages – including Romance-based creoles – will be discussed in a separate volume. A particularly strong focus will be placed on the smaller languages (*linguae minores*) that other reference works have not treated extensively. *MRL* will comprise volumes on Friulian, Corsican, Galician, Vulgar Latin, among others, as well as a *Manual of Judaeo-Romance Linguistics and Philology*. Volumes of the second type will be devoted to the systematic presentation of all traditional and new fields of Romance Linguistics, with the research methods of Romance Linguistics being discussed in a separate volume. Dynamic new research fields and trends will yet again be of particular interest, because although they have become increasingly important in both research and teaching, older reference works have not dealt with them at all or touched upon them only tangentially. *MRL* will feature volumes dedicated to research fields such as Grammatical Interfaces, Youth Language Research, Urban Varieties, Computational Linguistics, Neurolinguistics, Sign Languages or Forensic Linguistics.

https://doi.org/10.1515/9783110365955-202

Each volume will offer a structured and informative, easy-to-read overview of the history of research as well as of recent research trends.

We are delighted that internationally-renowned colleagues from a variety of Romance-speaking countries and beyond have agreed to collaborate on this series and take on the editorship of individual *MRL* volumes. Thanks to the expertise of the volume editors responsible for the concept and structure of their volumes, as well as for the selection of suitable authors, *MRL* will not only summarize the current state of knowledge in Romance Linguistics, but will also present much new information and recent research results.

As a whole, the *MRL* series will present a panorama of the discipline that is both extensive and up-to-date, providing interesting and relevant information and useful orientation for every reader, with detailed coverage of specific topics as well as general overviews of present-day Romance Linguistics. We believe that the series will offer a fresh, innovative approach, suited to adequately map the constant advancement of our discipline.

Günter Holtus (Lohra/Göttingen)
Fernando Sánchez Miret (Salamanca)
May 2018

Acknowledgements

We would like to thank the following people who kindly offered us advice during the preparation of this volume: Adam Ledgeway, Mair Parry, Nigel Vincent and Christopher Pountain. A number of people helped with copyediting the text, bibliographical work, and translating the chapters by Manzano and Bergounioux/Jacobson/Pietrandrea: Sarah Brierley, Jessica Brown, Merryn Davies-Deacon, Daniel McAuley, Jessica Soltys, and particularly Aedín Ní Loingsigh. We are also grateful to the series editors, Günter Holtus and Fernando Sánchez Miret, and to the editorial team at De Gruyter, notably Gabrielle Cornefert.

Wendy Ayres-Bennett and Janice Carruthers

https://doi.org/10.1515/9783110365955-203

Table of Contents

Medium, register, text type, genre

Linguae minores / Minoritized languages: status, norms, policy and revitalization

Language contact

Introduction

Wendy Ayres-Bennett and Janice Carruthers

0 Romance sociolinguistics: past, present, future

Abstract: In this chapter we consider what it means to produce a Manual of Romance Sociolinguistics. We outline the development of Sociolinguistics and situate Romance Sociolinguistics within this, highlighting in particular its distinctive characteristics. We conclude by considering possible new directions for scholars working in the field, and argue that there needs to be more truly comparative work, which draws on the wealth and diversity of the data afforded by the very many genetically related dialects, regional varieties, minoritized and major languages of the Romance-speaking area. We contend that such studies will not only contribute to general theories of variation and change and the testing of so-called sociolinguistic universals, but also address important areas of language policy, including the maintenance and support of linguistic diversity at a time of high mobility and migration, as well as concerns about the effects of globalization and the dominance of English.

Keywords: Romance sociolinguistics, sociolinguistics, recent and current trends, new directions

1 Introduction

What does it mean to produce a Manual of Romance Sociolinguistics? In particular, how does this differ from a Manual of Sociolinguistics in general? To answer this question, we will start by outlining some of the main areas and issues which have dominated sociolinguistics, particularly in the anglophone world, over the last fifty years, as well as considering some of the more recent and emerging fields of interest (section 2). We will then examine the extent to which these fields have been taken up by Romance scholars, and try to propose some explanations as to why certain sub-disciplines are more represented than others (section 3). Crucially, in section 4, we will address what is distinctive about work in Romance sociolinguistics, and what this might in turn offer to the field more broadly. In a final section (5), we will suggest some possible new directions and fresh opportunities for scholars working in Romance sociolinguistics, whether we think of the major Romance languages or the *linguae minores*. This is particularly vital given the diverse sociocultural and linguistic landscape across the Romance-speaking world, and notably the changing population patterns in both urban and rural contexts. These rich and diverse data may serve not only to reinforce or further exemplify current theories and methodologies, but also to challenge assumptions and bring new questions into focus.

https://doi.org/10.1515/9783110365955-001

2 Trends in sociolinguistic theory and practice

2.1 Milestones and classics

Sociolinguistics as a sub-discipline of linguistics has been thriving since the 1960s. The widely-acknowledged foundational figure is William Labov, whose work on American varieties of English (for example in New York City and Martha's Vineyard) has had a profound influence internationally on our understanding of linguistic variation and language change, and on the methodologies used for speaker sampling, fieldwork and analysis (Labov 1966; 1972a; 1972b; 1994; 2001). The Labovian paradigm has been the catalyst for the development of large-scale linguistic corpora, many of which are stratified according to key speaker variables such as socio-economic status, gender and age so that statistical data can be generated and empirical evidence cited for the relationship between language variation and sociolinguistic variables.[1] Labov also led the way in developing sociolinguistic interviewing techniques, particularly as regards the need to create as "natural" a context for interviewing as possible in order to attempt to obviate the observer's paradox and to gain access to natural, everyday speech which he termed the "vernacular" and which is the desired type of speech for most early and many contemporary corpora.[2] For example, Labov encouraged selection of a location where informants are relaxed such as their home (allowing "normal" life, even where it involves interruptions, to continue around the interview), and advocated the benefits of open questions that allow the interviewee to speak at length, as well as the use of techniques such as the "danger of death" question ("have you ever been in a situation where you thought you were in serious danger of being killed?" Labov 1972a, 92) in order to reduce the speaker's self-consciousness. Linguistic data from natural speech can then be compared to more formal varieties of language or to written forms (both of which are easier to obtain than access to the vernacular), allowing quantitative work on questions of medium and register. Indeed, Labov's early work incorporated a measure of stylistic variation, testing "vernacular" interview data against speakers' pronunciation when reading prose, word lists and minimal pairs (1972a, 207–216).

One of the key concepts to emerge from Labov's work has been the "linguistic variable", i.e. a linguistic feature whose behaviour varies according to social factors such as socio-economic status or gender, where that sociolinguistic variation can be measured, quantified and tested for statistical significance. The idea of a measurable linguistic variable implies that both "actual occurrences" and "possible occurrences"

1 The extent of Labov's worldwide influence is evident in the multiplicity of variationist projects referenced in standard works on sociolinguistic theory and practice such as Trudgill (1974; 2002); Chambers/Trudgill/Schilling-Estes (2002); Milroy/Gordon (2003); Tagliamonte (2012); Bayley/Cameron/Lucas (2013); Wardhaugh/Fuller (2015).

2 Note, however, that the very concept of the "vernacular" is problematic: see Milroy/Gordon (2003, 49–50).

of the variable can be identified and counted, since researchers will usually wish to establish to what extent a particular variable occurs relative to the number of times it could have occurred. The notion of "possible occurrences" makes the assumption that more than one linguistic variable is possible in a given context. In practice, this concept is relatively easily applicable to phonological variables, where we are dealing with different realizations of the same phoneme, e.g. Labov's study of the realization or absence of post-vocalic "r" in New York speech (1966; 1972a, 43–69). Indeed, phonological variables dominate Labov's early work and much of the pioneering work internationally in variationist linguistics (e.g. Labov 1966 looked at five phonological variables in New York speech; Trudgill 1974 explored three consonantal and thirteen vowel variables in his study of Norwich City).[3] However, transferring the concept of a linguistic variable to the syntactic domain is much more complex, since two possible variants of a syntactic feature are rarely entirely semantically equivalent. For example, while we might consider occurrence or non-occurrence of the French negative *ne* as "meaning the same thing" (e.g. *je sais pas* usually means the same thing as *je ne sais pas*),[4] the choice of *je vais faire* as opposed to *je ferai* is highly likely to involve temporal, aspectual and modal factors. The two options are thus not semantically equivalent and consequently problematic to consider straightforwardly as linguistic variables whose variation could be measured to show possible differences in behaviour according to speaker variables such as age, gender, socio-economic status, etc. Nonetheless, sociolinguists working in a Labovian paradigm have indeed taken the notion of the linguistic variable into the syntactic domain, using a variety of arguments to make the case that in some instances, the notion of a syntactic variable is entirely valid, thereby allowing empirical studies of a Labovian nature.[5]

The methodologies deployed when handling speaker and stylistic variables can also be complex and problematic. For example, it is notoriously difficult to measure socio-economic status, particularly given regional variations within societies and the different patterns found in different countries, and it is particularly tricky to score women for this variable.[6] Discussions of key issues concerning speaker variables can be traced back to Labov, not least the question of whether significant differences between language usage in different age groups should be considered indicative of language change "in apparent time" or whether they simply indicate "stable variation" between the linguistic behaviour of different age groups, known as "age grading" (1972a, 1–42; 160–182), the latter suggesting that "real-time" data may be required in order to be

3 See Tagliamonte (2012, 177–205) for an update on more recent work on phonological variation.

4 Even in cases like this, linguistic factors, such as phonological considerations, can mean that not all variables are equally likely.

5 For an early discussion, see Sankoff (1980b). See also the broader discussion in Milroy/Gordon (2003, 169–197) and Tagliamonte (2012, 206–246).

6 See the discussion in Milroy/Gordon (2003, chapter 2).

certain that change is occurring.[7] Moreover, the whole notion of stylistic variation is controversial, with a wide range of approaches developing after Labov's early work. These range from Coupland's (1980) innovative approach to style as "audience design" to highly quantitative work on register and genre, spearheaded by Biber (Biber 1984; Biber/Conrad 2009).[8] From Labov onwards, there is also a clear understanding amongst sociolinguists that speaker and stylistic variables interact with each other in highly complex ways.[9]

A second influential approach in sociolinguistics is the methodology developed by James and Lesley Milroy in their pioneering research on Belfast speech (Milroy/Milroy 1978; Milroy 1987). Two aspects of their approach have been particularly influential. The first is the development in sociolinguistics of a research technique more widely used in anthropology, i.e. participant observation. Rather than interviewing speakers, with all the problems generated by the complexities of the observer's paradox, the Milroys use participant observation, a technique whereby the fieldworkers get to know the speakers as individuals and as a community, to the point where they win the trust of the speakers they are recording and are able to leave the recording technology running in the background for long periods in the speakers' homes, thereby gaining access to unguarded, natural speech. There are, of course, drawbacks to this type of fieldwork methodology, not least the fact that there will inevitably be sections of indecipherable discourse (e.g. due to noises or speaker overlap) and the fact that it is very time-consuming and therefore expensive. However, the losses in these respects are offset by the gains in terms of access to the vernacular. The second area of influence from the Milroys is their innovative concept of "linguistic networks", whereby, in the case of their Belfast study, speakers in three localized inner-city communities (Ballymacarrett, Clonard and the Hammer) were scored according to the intensity of their everyday contact with, and amongst, members of their local community. Rather than measuring more standard variables such as age, gender and socio-economic status, the Milroys tested the strength of particular linguistic variables – in this case, marked features of Belfast speech – against the network scores of speakers, finding in general that the stronger the network score for a given speaker, the more likely that speaker was to use linguistic features strongly marked as belonging to Belfast.

From the early major projects in sociolinguistic variation, a certain number of sociolinguistic universals have emerged which have shaped debate internationally over the last fifty years. Amongst these are the "gender paradox" regarding women's speech and their tendency to adopt "perceived prestige" forms, which means that at times they lead linguistic change whilst at other times they resist it.[10] Another key

7 For full discussions of these issues, see the texts cited in footnote 1.

8 See the discussion in Milroy/Gordon (2003, 198–222).

9 See Schilling-Estes (2002) and the comments on style, sex and social class in Tagliamonte (2012, 35).

10 See, for example, women's avoidance of non-standard *-n* for *-ing* in Trudgill (1972) and their role in leading change in the Northern Cities Vowel Shift (Labov 2001, 285–290). For an excellent synopsis of

concept is that of "change from below", where change is initiated below the level of consciousness, usually in the vernacular, and spreads throughout the linguistic system (Labov 1972a, 178–180; Trudgill 1974, 90–132).[11]

2.2 Recent and current trends

We make no claim to give a comprehensive account of current trends in sociolinguistics. There are, nevertheless, some recent and ongoing developments which, in opening up fresh areas of research, merit mention because they have been taken up by scholars working on Romance varieties and are therefore represented in this volume.

The fundamentals of sampling and fieldwork methodology have not undergone dramatic transformation in the past number of years, although there has been an increased emphasis, particularly in conversational analysis, on obtaining "authentic" oral data in naturally-occurring contexts, rather than creating "false" contexts through techniques such as interviews (see the discussion in Mondada 2005). One significant change is the current greater awareness of the ethical dimension to data collection. As recently as 10–15 years ago, it was still possible to collect language data while interviewees were largely unaware of the linguistic purpose of the exercise. There was a good reason for this widespread practice: making speakers conscious of the linguistic motivation increases the chances of the observer's paradox coming into play and decreases the researcher's chances of gaining access to the vernacular. Now, with greater awareness of the need to treat data respectfully and sensitively, and the requirement from the higher education sector and funders that ethical questions be handled properly, all researchers build an important ethical dimension into their data collection, and guidelines on good practice are widely available.[12] Clearly, ethical issues are particularly acute when working with certain groups such as children or vulnerable adults, but the minimum level of ethical consent is now to give informants full details of the project in which they are participating (including details of personal metadata and linguistic data storage) and to gain their formal "informed consent". This can impact on the question of preventing the observer's paradox but other strategies, such as creating a maximally relaxed atmosphere, can be used to help mitigate the downsides. In contexts of naturally occurring oral data, ethical practice also involves getting to know those involved in the activity being recorded, and following up on data collection to mitigate against any potential sense of exploitation (Mondada 2005).

---

the possible reasons for women's tendency to opt for overtly prestigious forms, see Tagliamonte (2012, 32–34). Cheshire (2002) offers a discussion of more problematic elements of established sociolinguistic notions about women's speech.

11 See the discussion in Wardhaugh/Fuller (2015, 214–216) and Tagliamonte (2012, 27–29).

12 See BAAL (2016) and LSA (2009).

With the impact of the digital revolution, there has been a seismic shift in our approach to storing, accessing and annotating linguistic data. The widely-accepted norm currently is for corpora to be transcribed and digitized in a format such as XML, annotated using a variety of annotation systems (e.g. Part of Speech (POS) taggers, Text Encoding Initiative (TEI) conventions) and, where possible, to be made available online using a Creative Commons licence which allows wide access and citation for the purposes of scholarly research. The first chapter in this volume (↗1 Annotating oral corpora) discusses these developments in detail, but we note for now that one of the key goals in sociolinguistics in the last 20 years has been the development of digitization and annotations systems that are both compatible (and therefore comparable) internationally and fit for purpose in terms of long-term data preservation. It is a field where technology moves very quickly and where developments have facilitated the creation of a multiplicity of corpora – both synchronic and diachronic – worldwide, dramatically enhancing the range and reliability of both qualitative and quantitative sociolinguistic and historical sociolinguistic analyses.[13] Quantitative methods have continued to develop with increasing sophistication (see ↗2 Quantitative approaches for modelling variation and change), using statistical modelling and toolkits such as Goldvarb, Rbrul and R.[14]

New developments in technology have not only facilitated the sophisticated digitization and annotation of corpora; they have also enabled the growth of oral corpora, with increasingly unobtrusive high-quality recording possibilities and easier access to publicly available oral material such as radio data. In complementing written corpora, new oral corpora have created the possibility of incorporating data of multiple genres, text types, media and registers into mega-corpora such as the BNC. What started as a discussion of the Labovian concept of "style", has now opened up much greater possibilities for analysis of medium, register, text type and genre which have moved well beyond a simple "oral-written" dichotomy and into the field of multidimensional analysis (see Biber/Conrad 2009; ↗12 Oral genres). Given the digital nature of much social media, the explosion in communication through Facebook, Twitter, texting etc. has created a readily available dataset of new varieties, which are equally not easily classifiable in simple binaries such as oral and written. This renewed focus on questions related to Labov's original concept of "style" also sits well with what has been termed the "third wave" of sociolinguistics (Eckert 2012), where variations in stylistic practice are understood as having a clear social-semiotic value. These issues are mentioned in D'Agostino/Paternostro (↗7 Speaker variables and

13 For just two of the best-known English-language digitized corpora, see the British National Corpus (BNC, http://www.natcorp.ox.ac.uk, last access 18.02.2018) and the Corpus of Contemporary American English (COCA, http://corpus.byu.edu/coca, last access 18.02.2018). For a wide selection of linguistic corpora, see the Oxford Text Archive (https://ota.ox.ac.uk, last access 18.02.2018).

14 For an excellent discussion of the issues, including recent "mixed effects" modelling, see Tagliamonte (2012, 120–161).

their relation to language change, section 5) and in Zafiu's contribution (↗13 Register and text type, section 2.1).

In terms of innovatory trends in sociolinguistic analysis, a number of areas spring to mind. Two of these are connected to questions around regional varieties and *linguae minores*. First, new perspectives have been opened up around the concept of "dialect levelling". Whereas in the past, sociolinguists interested in regional variation would be most likely to focus on the strikingly regional features of a given variety, subsequent research, set in the context of globalization, mass communication and increasing mobility, has explored "dialect levelling", i.e. "the reduction or attrition of marked variants" (Trudgill 1986, 98).[15] Second, the field of language revitalization has grown substantially. This is a vast field within sociolinguistics and involves political questions around language policy, particularly as regards *linguae minores*. Recent areas of interest include discussion of the issues raised by "new speakers" (i.e. speakers of "new" standardized, often urban varieties which can be linguistically distant from native speaker varieties)[16] as well as the use of language in the public space, e.g. on signage, street names etc.[17]

3 Positioning Romance sociolinguistics

To what extent has Romance sociolinguistics adopted these major trends seen in the broader field? In this section we will consider first the uptake of Labovian-style quantitative studies, before turning to the more recent themes and approaches outlined in the previous section.

3.1 The impact of Labovian-type approaches on Romance sociolinguistics

Compared with work on English, there is a relative lack of major studies across the Romance domain which adopt a classic Labovian variationist approach, although as the chapter in this volume by D'Agostino/Paternostro shows (↗7 Speaker variables and their relation to language change), some of the key ideas in Labovian theory (e.g.

15 For explorations of dialect levelling in British English, see Williams/Kerswill (1999) and Kerswill (2003).

16 See for example the special issue of the *International Journal of the Sociology of Language* (2015, vol. 231) containing a series of articles on new speakers of a range of minoritized languages. See also the New Speakers Network, funded by COST: http://www.nspk.org.uk (last access 18.02.2018).

17 Much research on language in the public space draws on Linguistic Landscape theory. See Landry/Bourhis (1997); Kallen (2010); Blommaert (2013); the new journal, *Linguistic Landscape* (Benjamins); ↗20 Revitalization and the public space.

the gender paradox referred to above) can be discerned in earlier work on the Romance languages. That said, the variationist studies on Romance that do exist have made extremely important contributions to the field from the 1970s through to contemporary research across a range of linguistic variables, including phonological variation,[18] morphosyntactic variation,[19] stylistic variation,[20] as well as issues around gender variation.[21] Indeed, analyses of syntactic variation in both French and Spanish were crucial in early discussions of whether or not it was possible to measure syntactic variation.[22] This volume showcases current research for which Labovian methodologies have been the springboard; Mooney's chapter (↗2 Quantitative approaches for modelling variation and change) demonstrates how a Labovian quantitative paradigm can shed new light on a regional minoritized language; while Poplack et al. (↗8 Variation and grammaticalization in Romance) extend the strengths of this type of approach into comparative work across several Romance languages; and Armstrong/Mackenzie (↗6 Speaker variables in Romance) interrogate the complexities of the relationship between certain speaker variables and the cultural context and ideologies in which they operate. Researchers working on the Romance languages drawing on the Milroys' participant observation methodology are much less numerous, but have again made important contributions.[23]

One possible reason for the relative lack of research in these paradigms in the European countries where Romance languages are spoken may be linked to the availability of translations, since the primary methodological sources are all published in English. Nevertheless, it is important to note that, while fewer translations of the Milroys' work are available, many of Labov's key works have been translated into both French and Spanish. Other reasons seem more plausible. The first relates to the different concerns across the Romance-speaking nations, notably in Europe, which reflect the varying linguistic, historical and cultural traditions of those countries. Notable amongst these is the interest in dialects and regional languages, where the primary concern has been documentation and analysis of the multiple varieties rather than variation within "major" languages according to the type of speaker variables central to Labovian approaches, i.e. age, gender, socio-economic status etc. Indeed, this emerges strongly in the discussion in Jones/Parry/Williams (2016), where the

18 Important early work includes Cedergren (1973) on Spanish. For further discussion and recent examples, see the contributions on Spanish of Medina-Rivera (2011); Lipski (2011) and Samper Padilla (2011); see Carmo/Tenani (2013) for a recent Portuguese example.

19 See Poplack (2011) for a comparative Romance study; see Coveney (1996) and Ashby (2001) for French examples; and for recent examples from Spanish see Schwenter (2011); Bentivoglio/Sedano (2011); Serrano (2011).

20 For example, Armstrong (2001); Medina-Rivera (2011); Massot/Rowlett (2013); Kabatek (2016).

21 See for example Parry (1991); Fresu (2006); Holmquist (2011).

22 See the debate in Lavandera (1978); see also Sankoff/Vincent (1980) on retention and loss of negative *ne* in French; Sankoff/Thibault (1980) on *avoir*/*être* alternation.

23 See for example Pooley (1996); Klein (1989); Vietti (2002) and McAuley (2017).

authors highlight the dearth of Labovian-type approaches in Italy (where interest in dialects and regional varieties is paramount), as well as the distinction in Spain between bilingual areas (where regional concerns are paramount) and monolingual areas (where the small number of Labovian studies are based). The Romance tradition of dialectology has strongly influenced the shape of sociolinguistics in the field and we will return to the importance of regional languages and dialects as a distinctive feature of Romance sociolinguistics in section 4 below.

The second relevant factor, perhaps the most striking, is that where Labovian methodologies have been applied to Romance data, it has often been by scholars emerging from the anglophone tradition of sociolinguistic methodology, who either work, or received their research training, in Canada, the United States or the UK.[24] In practice, the bulk of this Romance variationist research relates to French and Spanish. In view of Labovian methodology's connections to these three countries, this may be because French is an official language in Canada (and crucially is the first language of Quebec); Spanish is widespread as a heritage language in North America which is of course positioned geographically "next door" to a continent where Spanish is the dominant language; and French and Spanish are by far the most widely taught and researched Romance languages in the UK university system.

3.2 Recent and current trends in Romance sociolinguistics

The newer sociolinguistic trends discussed in 2.2 above are also, not surprisingly, influencing current research on the Romance languages. As was the case regarding the influence of Labov, certain approaches have had more traction in the Romance field than others, and many of the chapters in this volume illustrate the trends that are currently shaping Romance sociolinguistics. Moreover, recent growth in some fields can be seen as seamlessly linked to Labovian concepts or, in some cases, as bringing together phenomena that have been discussed since the 1960s with the opportunities provided by new technologies, e.g. in work on spoken varieties, on creoles and endangered varieties, on new varieties, or on register, medium and text type.

In terms of methodology, we can see the importance in contemporary research of building major corpora, both written and spoken, which are digitized and annotated using internationally-recognized good practice. In their detailed discussion of current methodological issues (↗1 Annotating oral corpora), Bergounioux/Jacobson/Pietrandrea cite many of the major digitized corpora of French, Spanish, Portuguese and

24 For example William Ashby, Zoe Boughton, Aidan Coveney, Beatriz Lavandera, Gillian and David Sankoff, Rena Torres Cacoullos, and a number of contributors to this volume.

Italian. Corpora are also being created, digitized and annotated in the minoritized Romance languages,[25] opening up new possibilities for research, particularly where there is a quantitative component in the analysis, as is evident from Mooney's discussion (↗2 Quantitative approaches for modelling variation and change). New opportunities for comparative research have been created by the growth of large-scale corpora: the contribution by Poplack et al. (↗8 Variation and grammaticalization in Romance), which looks at the evolution of the subjunctive across several Romance languages, is an excellent example of the superior level of analysis, both qualitative and quantitative, that can be achieved when large annotated corpora are available. Since there is now a considerable lapse of time since the early transcribed corpora in the 1970s, new corpora are being developed that allow "real-time" analysis of data, as opposed to Labovian "apparent-time" analysis, where evidence of linguistic change hinges on age differences. For example, the Orléans corpus (ESLO) that was originally created between 1968 and 1974 has now been supplemented by a contemporary corpus from 2008 onwards,[26] where a number of speakers from the original corpus feature again, forty years later, thereby allowing real-time analysis of linguistic change. Similarly, the original Montreal corpus recorded in 1971 has been supplemented by corpora in 1984 and 1995, again allowing the possibility of real-time analysis.[27] It is important, however, to note that we are still dealing with very short timespans for the tracing of major linguistic changes, as Ayres-Bennett points out (↗9 Historical sociolinguistics and tracking language change). As a result, those working in historical sociolinguistics are both exploiting existing corpora and creating new databases of textual sources which seek to meet the challenge of finding appropriate sources for what Labov famously termed the "bad data" problem (Labov 1972c, 98; 1994, 10–11). In Ayres-Bennett's contribution, the focus is on the extent it is possible to exploit large-scale multi-genre databases and corpora to track innovation and the spread of change through different text types or genres. By contrast, Anipa (↗10 Speaker-based approaches to past language states) offers a different solution to the problem of sources for socio-historical linguistics, arguing for what he terms the micro-framework, which focuses on the linguistic usage of individual writers of literary texts.

As mentioned in section 3.2 above, the growth in high-quality oral corpora around the world has led to a much larger volume of research on contemporary spoken varieties, which in turn has facilitated an increase in research on medium,

25 For a pre-2003 list, see Pusch/Raible (2002). Recent corpora include, for Occitan http://redac.univ-tlse2.fr/bateloc (last access 18.02.2018); for Galician http://ilg.usc.es/gl/node/1016 (last access 18.02.2018); for Catalan http://nlp.ffzg.hr/resources/corpora/cawac (last access 18.02.2018).

26 See http://eslo.huma-num.fr (last access 18.02.2018). Ashby (2001) has also exploited two 'real-time' corpora from 1976 and 1995 to analyse the retention/loss of negative *ne* in French.

27 See for example Sankoff/Blondeau (2007).

register and text type. This growth in oral corpora has impacted positively across the Romance domain. In addition to multiple studies of specific oral phenomena (lexical, morphosyntactic and phonetic/phonological) in particular Romance languages, we have seen the development of comparable oral corpora in several Romance languages (see Cresti/Moneglia 2005) and the analysis of specific phenomena across several Romance languages, such as Philippe Martin's recent work on intonation in Romance (Martin 2015). Oral corpora have also been crucial for investigating creoles and endangered varieties, where spoken data are vital for full documentation. In such cases, particular methodological issues arise, such as the presence of multiple varieties and the paucity of speakers (see Repetti's discussion of endangered varieties, ↗4 Fieldwork and building corpora for endangered varieties), the problematic and complex relationship (both politically and linguistically) with a standard language (see ↗4 Fieldwork and building corpora for endangered varieties, and also Sippola's discussion of creole ↗3 Collecting and analysing creole data), and the question of the researcher's status as "insider" or "outsider" (discussed in both these chapters). Changes in practice around research ethics in collecting oral corpora are also reflected strongly in the development of corpora of creoles and endangered languages, where standard research techniques such as interviews can pose particular problems, both cultural (see ↗3 Collecting and analysing creole data) and practical (see Repetti's comments on elderly speakers, ↗4 Fieldwork and building corpora for endangered varieties). As noted, the growth of oral corpora has greatly enhanced research on media, register and genre. Carruthers (↗12 Oral genres) demonstrates how the availability of digitized Spanish corpora has been the catalyst for multidimensional work on style, genre and register and how corpora of oral French can shed new light on the concept of genre. Both Carruthers and Zafiu (↗13 Register and text type) touch on the problematic definitional issues in this field and on the complex relationships between genre, text type and register, with Zafiu discussing a series of case studies from Romanian. More broadly, research on medium, register and genre has great potential for future developments in several fields. For example, questions relating to oral and written media are central to the use and status of creole, as is clear in Ludwig's discussion of Martinican and Guadeloupean creole (↗15 Medium and creole). Recent and current research also explores the complexities of new varieties that are emerging in electronic media, as reflected in Kallweit's discussion of Spanish (↗14 New Media). It will be interesting to see how Romance sociolinguistics develops elements of the so-called "third wave" in sociolinguistics (see 2.2 above) in the ways in which it takes research forward on questions of style, medium, register and genre.

A very significant area of growth, particularly for French and to some extent Spanish, involves the exploration of linguistic phenomena in superdiverse urban contexts where the "melting pot" represented by large cities raises fascinating linguistic issues relating not only to the influence of migration, multiculturalism and multilingualism, but also to questions of peer-group, community, identity and social

cohesion.[28] Gadet/Hambye (↗27 The metropolization of French worldwide) explore the complexities of metropolization, comparing linguistic patterns for French in Europe, North America and Africa. They demonstrate the importance of factors such as the level of plurilingualism amongst speakers, the status of the various languages in contact, as well as their social and educational position. These questions are further complicated by different types of migration patterns which produce a range of different permutations involving Romance languages: e.g. large urban contexts to which Romance speakers migrate where the main language of the city is a Romance one (e.g. a Mexican migrating to Spain); similar contexts where the main language is not Romance (a lusophone migrating to London); and Romance contexts where a variety of Romance and non-Romance languages are present through migration (a multicultural setting in Marseille where migrant languages include varieties of Arabic, sub-Saharan languages and possibly Occitan). As Mar-Molinero/Paffey explain (↗28 Transnational migration and language practices), this field is also leading to new theoretical developments, notably the emergence of concepts such as "translanguaging", and "metrolingualism", and the creation of large urban corpora and international collaborations.[29] Mar-Molinero/Paffey also explore the further complexities connected to different stages of migration (e.g. the concept of "return" migrants or the possibility of secondary migration to one destination followed by another). The linguistic outcomes of immigration in Italy are examined by Goglia (↗26 Code-switching and immigrant communities), whose contribution exemplifies code-switching by immigrant speakers and its function in interactions with Italians, in "in-group" interactions, and where it involves Italo-Romance dialects. Schulte (↗22 Romance in contact with Romance) explores contact-induced structural change between Romance languages in three different situations; alongside recent migration-based contact, he considers long-term contact situations in Spain and in the New World. Finally, it is important to note that questions of language contact between Romance and non-Romance languages, so central to explorations of superdiverse urban settings, are also increasingly researched in other contexts. Indeed, Bullock (↗25 Language contact in a rural community) stresses the importance of not neglecting the rural context, particularly where we find isolated rural communities speaking what is effectively a minority language, even if elsewhere it is a major world language. Through the case study of Frenchville, a French-speaking community in Pennsylvania, she explores the high levels of inter- and intra-speaker variability in this setting. In her contribution, Escobar (↗23 Language contact between typologically different languages) considers

28 See, for example, Armstrong/Jamin (2002); Trimaille (2004); Trimaille/Billiez (2007); Pooley (2009); Gasquet-Cyrus (2004; 2009); Zentella (2009); Gadet (2013); Hambye/Gadet (2014); Padilla/Azevedo/Olmos-Alcaraz (2015); Lynch (forthcoming).

29 For example, there is close collaboration and comparative work between the "London Multicultural English project" and Françoise Gadet's project on Parisian speech (www.mle-mpf.bbk.ac.uk, last access 18.02.2018), both of which have generated substantial corpora of urban speech.

functional transfer in highly bilingual communities in South America, between the Romance language of the colonizer/state, i.e. Spanish, and a non-Romance indigenous language, in this case Quechua. McLaughlin (↗24 When Romance meets English) investigates the differences in terms of the effects of language contact with English across different Romance languages, i.e. French, Spanish and Italian, and highlights the question of "Romance in contact with English" as a major area for future research, given the increasingly global status of English.

4 Distinctive features of Romance sociolinguistics

In addition to the recent and current trends in Romance sociolinguistics discussed in the previous section, which might be viewed as emerging from major international developments since the 1960s, Romance scholars have developed their own distinctive approaches to the Romance languages, informed by the work of different linguists and paradigms. We will mention just three representative examples here.

The first, "variational linguistics", arose in the Scandinavian tradition. In this approach the language system of a community is described as a "language architecture" with different diasystems, including diatopic, diastratic, diaphasic, diamesic, as well as diachronic, varieties. The terms diatopic (spatial) and diastratic (social) variation were first introduced by the Norwegian linguist Flydal (1951) and, three years later, Weinreich (1954) proposed the term "diasystem". Subsequently, to cite Völker (2009, 32), Coșeriu "a repris, unifié, modifié et surtout promu les instruments terminologiques proposés par Flydal et Weinreich en confirmant l'usage des termes *diasystème*, *diatopique* et *diastratique* [...] et en introduisant une dimension nouvelle [...] *diaphasique*". The diaphasic is used to refer to variation according to different styles or registers used in different communicative settings (see, for example, Coșeriu 1981). The work of Coșeriu, who held a chair in Tübingen, has been especially significant for German Romanists and Hispanists, but it is also exploited in discussions of other Romance languages and varieties.[30] In this volume, the influence of "variational linguistics" and the concept of diasystems can be seen, for instance, in ↗5 Romance dialectology; ↗7 Speaker variables and their relation to language change; ↗13 Register and text type; ↗18 The languages and dialects of Italy.

The second, termed linguistic ecology,[31] has been particularly influential in France. In his 1999 work, *Pour une écologie des langues du monde*, Calvet criticizes what he considers the dominant model in linguistics of considering language as an abstract system. For Calvet, language is rather "un ensemble de pratiques et de représentations" (1999, 165). As a result, the ecolinguistic approach consists in study-

30 See, for example, D'Achille (2008) on Italian or Verjans (2014) on French.

31 The term "ecology of language" is borrowed from Einar Haugen (see, for instance, Haugen 1972).

ing "les rapports entre les langues et leur milieu, c'est-à-dire d'abord les rapports entre les langues elles-mêmes, puis entre ces langues et la société" (1999, 17). Emphasis then is placed on looking at the complex nature of the social and communicative dimensions of language. Linguistic ecology aims to explain social communication as a whole, considering the factors which explain the revitalization, maintenance or loss of languages. As Françoise Gadet (2009) notes, the umbrella term *Écologie des langues* aims to introduce a certain unity into a somewhat heterogeneous domain, comprising sociolinguistics, dialectology, creolistics, language contact, plurilingualism, etc. (many of which are represented in this volume), although it may itself run the risk of perpetuating the idea of norms and variations from the norm.

Perhaps most important in the Romance context has been the emphasis on dialectology and the related field of the production of dialect atlases. In other words, the diatopic long held sway in Romance studies over other types of *dia-* variations. This is particularly true for Italy where, as Parry (Jones/Parry/Williams 2016, 616) observes, diatopic variation is the primary dimension of study in the case of sociolinguistic variation, the diverse dialects interacting everywhere with diastratic, diaphasic and diamesic variation. Starting from the last third of the nineteenth century, dialect studies became a vital part of historical-comparative studies of the Romance languages, and the atlases that were produced in the last decades of the nineteenth century and the first decades of the twentieth century, such as the *Petit Atlas phonétique du Valais roman* by Gilliéron (1880), the *Petit Atlas linguistique d'une région des Landes* by Millardet (1910), and above all, the *Atlas linguistique de la France* (Gilliéron/Edmont 1902–1910), came to inspire a whole series of atlases of different Romance varieties in the period 1910–1940.[32] As Swiggers observes (2010), from the 1970s dialectology started to incorporate a sociolinguistic perspective, so that maps began to include information about the age, sex, social class and education of the speakers. Indeed, Manzano (↗5 Romance dialectology) demonstrates how important the field of dialectology has been since the nineteenth century in helping to shape the newer field of sociolinguistics.

The study of diatopic variation in Romance – of dialects, regional varieties, regional and minoritized languages, etc. – has proved to be a particularly rich and productive area for consideration of a number of key issues, not least because of the genetic relationship between the languages and dialects across Romania. The volume of languages and speakers involved also makes the Romance-speaking area a particularly fruitful observatory for such issues, compared to other areas, where there may only be a constellation of two or three related languages (e.g. Frisian/Dutch/German; English/Scots). The chapters by Ramallo on Spain (↗17 Linguistic diversity in Spain) and Berruto on Italy (↗18 The languages and dialects of Italy) are particularly instructive in this regard, where the majority of the multiple languages and varieties attested are linguistically related to each other within the Romance family.

32 See Swiggers (2010); ↗5 Romance dialectology.

Moreover, the richness of the linguistic landscape means that study of the Romance languages does not just give us an abundance of data, but also adds an additional layer of theoretical complexity. Manzano (↗5 Romance dialectology) and Berruto (↗18 The languages and dialects of Italy), for instance, both open up the thorny issues around what should be termed a language or a dialect, while Ramallo (↗17 Linguistic diversity in Spain) relates these issues to the status of the different languages and dialects in Spain. A number of scholars have in recent years re-opened the question of the definition of a regional variety,[33] notably in relation to French. In terms of the analysis of Romance varieties, there are a number of clear trends. As we have already noted, there is growth in the related area of the study of dialect levelling, particularly with respect to French, Spanish and Italian. Discussions centre not only on the phenomenon of levelling but also on the complexities involved in different parts of Romania, with their differing histories in terms of the relationships between languages, dialects and regional varieties.[34] There is also debate around dialect and standard convergence where minoritized varieties are concerned (see Cerruti/Regis 2014).

In the light of the multiplicity of regional languages and dialects, study of the Romance-speaking area inevitably brings to the fore important questions about the status of languages, of linguistic ideologies and policies. We have placed particular emphasis in this volume on the *linguae minores* and important questions about their revitalization. Linked to revitalization is also the complex area of "new speakers" (see also 2.2 above) who speak varieties which can differ substantially from those spoken by native speakers and who, in some cases such as Galician and Rhaeto-Romance, constitute in fact the dominant group of speakers.[35] The vulnerability of many of the minoritized varieties means that questions of language planning and policy are never far away.[36] In this volume, a range of language policy issues are discussed, from policy within a single country where Romance languages and dialects dominate (Ramallo, ↗17 Linguistic diversity in Spain), to comparative policy across several countries (Bochmann, ↗16 Language policies in the Romance-speaking countries of Europe), to policy in countries where Romance languages are minoritized relative to other languages (Grünert, ↗19 Multilingualism in Switzerland). Two crucial areas of policy are discussed in individual chapters, notably educational policy (see Ghimenton/Depau's discussion of practice in France and Italy, ↗21 Revitalization and education), and the question of "language in the public space" (see Blackwood's discussion of Corsican and Niçois, ↗20 Revitalization and the public space, where "Linguistic Landscape" methodology is particularly productive; see also note 17 above). Conver-

33 For example, Boughton (2005); Hornsby (2006); Armstrong/Boughton (2009); Mooney (2016).

34 See Kabatek (2016).

35 For interesting recent and current discussions of the authority and legitimacy of new varieties, see Kasstan on Francoprovençal (2018) and O'Rourke/Ramallo (2013) on Galician. See also the discussion in Kabatek (2016, 632).

36 See García (2011); Mar-Molinero/Paffey (2011); Soria (2015); Mooney (2015) and Joubert (2015).

sely, the role of standard languages and norms has been central to the debate in Italy with the *questione della lingua*, and in France with its tradition of prescriptivism and purism; in his contribution to this volume, Costa-Carreras (↗11 Variation and prescriptivism) discusses recent work on prescriptivism, a subject which has at times been considered unfashionable – or even shunned – by descriptive linguists.

5 Opportunities and new directions

Romance linguistics as a discipline – and particularly, within this, Romance philology – has traditionally derived its strength and originality from the genetic relatedness of the Romance varieties studied and, consequently, the comparative perspective afforded by the data. This comparative approach has led to great advances in several areas of linguistics, including phonology, morphology, syntax, and, of course, historical linguistics, as well as in linguistic theory more generally (see, for instance, Maiden/Smith/Ledgeway 2011–2013; Ledgeway/Maiden 2016).

It is therefore all the more striking that, as yet, a truly comparative Romance perspective is largely lacking in the field of Romance sociolinguistics. The vast majority of studies published to date consider an issue in relation to a particular Romance language or variety. Indeed, whilst for some chapters it was relatively easy to include data from a range of languages and to introduce a comparative dimension, in several instances this was difficult, or even impossible, because of the concentration of research by most scholars on one particular language. Our volume, then, reflects the fact that truly comparative work is, as yet, at a relatively early stage of development. One reason for this is that, for some languages, there is still a dearth of work on that individual language from a sociolinguistic perspective, and such work might be thought to be an essential prerequisite for comparative studies. The different level of treatment for the different Romance languages and varieties means that there is still a great deal of unexploited data, for instance in the case of Romanian. A second reason lies in the fact that, as we have noted, large corpora are needed for sociolinguistic research, and comparable corpora are needed for the different languages to facilitate comparative work. Ideally, there would be international cooperation in order to achieve a sufficiently high level of comparability in terms of speaker variables, annotation etc., so that we can be sure of comparing like with like, but the creation and annotation of databases is costly and time-consuming, and depends on the availability of funding as well as on a favourable intellectual and political context.

A good example of a welcome initiative is the creation of comparable historical corpora for Spanish and Portuguese cited by Ayres-Bennett (↗9 Historical sociolinguistics and tracking language change), but there is tremendous potential for further development of a Romance perspective. We have already mentioned the benefits of comparative corpora that are evident in the contribution by Poplack et al. (↗8 Variation and grammaticalization in Romance) yet this chapter also highlights the varia-

bility of the corpora which currently exist, and the challenges that this heterogeneity poses. We will cite just three cases of where the adoption of a comparative Romance perspective could be highly beneficial. First, a comparative study of dialect levelling across the Romance domain would seem to be highly desirable, particularly given the different status and vitality of the varieties, the varying strength of the substratum, etc. Second, a comprehensive comparative study of the richness and diversity of the linguistic landscape across the Romance area with its different dialects, regional varieties, regional languages, etc. and the large number of speakers involved makes it, in our view, a particularly fertile test bed for research in this area. Third, Romania, with its many minoritized varieties, would seem to be an excellent laboratory for examining different pathways of change, ranging from the traditional course of a dialect giving way to a regional variety, through the potential integration of regional features into new urban varieties, to the revitalization of a variety through new speakers (and the dangers this brings of alienating native speakers).

Why are such comparative Romance sociolinguistic studies necessary? The wealth and diversity of the data afforded by Romance means that such studies can contribute importantly to the testing of the so-called sociolinguistic universals, as well as contributing to general theories of variation and change, etc. However, the implications and impact of such work goes well beyond the interest to academic specialists in sociolinguistics or language variation and change. At a time of high mobility and migration, increasing globalization and concerns about the dominance of English, the status and vitality of Romance varieties raises important questions concerning national, regional and local identity. Much work remains to be done on the impact that the high volume of population movements will have on the rich tapestry of dialects, regional varieties, regional and standard languages across the Romance-speaking area. Policy on language revitalization and the protection of linguistic diversity, for instance, needs to be underpinned by evidence-based research. Comparative data from a number of related languages and situations can only serve to strengthen the arguments in support of individual languages and dialects.

6 Bibliography

Armstrong, Nigel (2001), *Social and Stylistic Variation in Spoken French. A Comparative Approach*, Amsterdam/Philadelphia, Benjamins.

Armstrong, Nigel/Boughton, Zoe (2009), *Perception and Production in French Dialect Levelling*, in: Kate Beeching/Nigel Armstrong/Françoise Gadet (edd.), *Sociolinguistic Variation in Contemporary French*, Amsterdam/Philadelphia, Benjamins, 9–24.

Armstrong, Nigel/Jamin, Mikaël (2002), *"Le Français des banlieues": Uniformity and Discontinuity in the French of the Hexagon*, in: Kamal Salhi (ed.), *French in and out of France: Language Policies, Intercultural Antagonisms and Dialogues,* Frankfurt/Bern, Lang, 107–136.

Ashby, William (2001), *Un nouveau regard sur la chute du "ne" en français parlé tourangeau: S'agit-il d'un changement en cours?*, Journal of French Language Studies 11(1), 1–22.

BAAL [The British Association for Applied Linguistics] (2016), *Recommendations on Good Practice in Applied Linguistics*, https://baalweb.files.wordpress.com/2016/10/goodpractice_full_2016.pdf (last access 20.02.2018).

Bayley, Robert/Cameron, Richard/Lucas, Ceil (edd.) (2013), *The Oxford Handbook of Sociolinguistics*, Oxford, Oxford University Press.

Beeching, Kate/Armstrong, Nigel/Gadet, Françoise (edd.) (2009), *Sociolinguistic Variation in Contemporary French*, Amsterdam/Philadelphia, Benjamins.

Bentivoglio, Paola/Sedano, Mercedes (2011), *Morpho-Syntactic Variation in Spanish-Speaking Latin America*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 168–186.

Biber, Douglas (1984), *Variation across Speech and Writing*, Cambridge, Cambridge University Press.

Biber, Douglas/Conrad, Susan (2009), *Register, Genre and Style*, Cambridge, Cambridge University Press.

Blommaert, Jan (2013), *Ethnography, Superdiversity and Linguistic Landscapes: Chronicles of Complexity*, Bristol, Multilingual Matters.

Boughton, Zoe (2005), *Accent Levelling and Accent Localisation in Northern French: Comparing Nancy and Rennes*, Journal of French Language Studies 15(3), 235–256.

Calvet, Louis-Jean (1999), *Pour une écologie des langues du monde*, Paris, Plon.

Carmo, Márcia Cristina do/Tenani, Luciani Ester (2013), *The Pretonic Mid-Vowels in the Variety of the Northwest of São Paulo: A Sociolinguistic Analysis*, Alfa: Revista Linguística 57(2), 607–637.

Cedergren, Henrietta J. (1973), *The Interplay of Social and Linguistic Factors in Panama*, doctoral thesis, Ithaca, NY, Cornell University.

Cerruti, Massimo/Regis, Ricardo (2014), *Standardization Patterns and Dialect/Standard Convergence: A North-Western Italian Perspective*, Language in Society 43, 83–111.

Chambers, Jack/Trudgill, Peter/Schilling-Estes, Natalie (edd.) (2002), *The Handbook of Language Variation and Change*, Oxford, Blackwell.

Cheshire, Jenny (2002), *Sex and Gender in Variationist Research*, in: Jack Chambers/Peter Trudgill/Natalie Schilling-Estes (edd.), *The Handbook of Language Variation and Change*, Oxford, Blackwell, 423–443.

Coșeriu, Eugenio (1981), *Los conceptos de "dialecto", "nivel" y "estilo de lengua" y el sentido propio de la dialectología*, Lingüística española actual 3, 1–32.

Coupland, Nikolas (1980), *Style-Shifting in a Cardiff Work-Setting*, Language in Society 9(1), 1–12.

Coveney, Aidan (1996), *Variability in Spoken French. Interrogation and Negation*, Bristol, Intellect.

Cresti, Emanuela/Moneglia, Massimo (2005), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages,* Amsterdam/Philadelphia, Benjamins.

D'Achille, Paolo (2008), *Le varietà diastratiche e diafasiche delle lingue romanze dal punto di vista storico: Italiano*, in: Gerhard Ernst et al. (edd.), *Romanische Sprachgeschichte. Ein internationales Handbuch zur Geschichte der romanischen Sprachen*, vol. 3, Berlin/New York, De Gruyter, 2334–2355.

Díaz-Campos, Manuel (ed.) (2011), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell.

Eckert, Penelope (2012), *Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation*, Annual Review of Anthropology 41, 87–100.

Flydal, Leiv (1951), *Remarques sur certains rapports entre le style et l'état de langue*, Norsk Tidsskrift for Sprogvidenskap 16, 240–257.

Fresu, Rita (2006), *"Gli uomini parlano delle donne, le donne parlano degli uomini": Indagine sociolinguistica in un campione giovanile di area romana e cagliaritana*, Rivista italiana di dialettologia 30, 23–58.

Gadet, Françoise (2009), *Sociolinguistique, écologie des langues, et cetera*, Langage et société 129, 121–135.

Gadet, Françoise (2013), *Collecting a New Corpus in the Paris Area: Intertwining Methodological and Sociolinguistic Reflections*, in: Mari C. Jones/David Hornsby (edd.), *Language and Social Structure in Urban France*, Oxford, Legenda, 162–171.

García, Ofelia (2011), *Planning Spanish: Nationalising, Minoritising and Globalising Performances*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 667–685.

Gasquet-Cyrus, Médéric (2004), *The Sociolinguistics of Marseille*, International Journal of the Sociology of Language 169, 107–123.

Gasquet-Cyrus, Médéric (2009), *Territorialisation, stigmatisation et diffusion: L'Accent "quartiers nord" à Marseille*, in: Thierry Bulot (ed.), *Formes et normes sociolinguistiques (ségrégations et discriminations urbaines)*, Paris, L'Harmattan, 209–222.

Gilliéron, Jules (1880), *Petit Atlas phonétique du Valais roman (sud du Rhône)*, Paris, Champion.

Gilliéron, Jules/Edmont, Edmond (1902–1910), *Atlas linguistique de la France*, 12 vol., Paris, Champion.

Hambye, Philippe/Gadet, Françoise (2014), *Contact and Ethnicity in "Youth Language" Description: In Search of Specificity*, in: Robert Nicolai (ed.), *Questioning Language Contact: Limits of Contact, Contact at its Limits*, Leiden/Boston, Brill, 183–216.

Haugen, Einar (1972), *The Ecology of Language: Essays by Einar Haugen*, edited by Anwar S. Dil, Stanford, Stanford University Press.

Holmquist, Jonathan (2011), *Gender and Variation: Word-Final /s/ in Men's and Women's Speech in Puerto Rico's Western Highlands*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 230–243.

Hornsby, David (2006), *Redefining Regional French. Koinéization and Dialect Levelling in Northern France*, Oxford, Legenda.

Jones, Mari C. (ed.) (2015), *Policy and Planning for Endangered Languages*, Cambridge, Cambridge University Press.

Jones, Mari C./Hornsby, David (edd.) (2013), *Language and Social Structure in Urban France*, Oxford, Legenda.

Jones, Mari C./Parry, Mair/Williams, Lynn (2016), *Sociolinguistic Variation*, in: Adam Ledgeway/Martin Maiden (edd.), *The Oxford Guide to the Romance Languages*, Oxford, Oxford University Press, 611–623.

Joubert, Aurélie (2015), *Occitan. A Language that Cannot Stop Dying*, in: Mari C. Jones (ed.), *Policy and Planning for Endangered Languages*, Cambridge, Cambridge University Press, 171–187.

Kabatek, Johannes (2016), *Diglossia*, in: Adam Ledgeway/Martin Maiden (edd.), *The Oxford Guide to the Romance Languages*, Oxford, Oxford University Press, 624–633.

Kallen, Jeffrey (2010), *Changing Landscapes. Language, Space and Policy in the Dublin Linguistic Landscape*, in: Adam Jowalski/Crispin Thurlow (edd.), *Semiotic Landscapes: Language, Image, Space*, London, Continuum, 41–58.

Kasstan, Jonathan (2018), *Exploring Contested Authenticity among Speakers of a Contested Language*, Journal of Multilingual and Multicultural Development, https://www.tandfonline.com/doi/full/10.1080/01434632.2018.1429451 (last access 05.03.2018).

Kerswill, Paul (2003), *Dialect Levelling and Geographical Diffusion in British English*, in: David Britain/Jenny Cheshire (edd.), *Social Dialectology. In Honour of Peter Trudgill*, Amsterdam/Philadelphia, Benjamins, 223–243.

Klein, Gabriella (ed.) (1989), *Parlare in città: studi di sociolinguistica urbana*, Galatina, Congedo.

Labov, William (1966), *The Social Stratification of English in NY City*, Washington, DC, Center for Applied Linguistics.

Labov, William (1972a), *Sociolinguistic Patterns*, Philadelphia, University of Pennsylvania Press.

Labov, William (1972b), *Language in the Inner City*, Philadelphia, University of Pennsylvania Press.

Labov, William (1972c), *Some Principles of Linguistic Methodology*, Language in Society 1, 97–120.
Labov, William (1994), *Principles of Linguistic Change*, vol. 1: *Internal Factors*, Oxford, Blackwell.
Labov, William (2001), *Principles of Linguistic Change*, vol. 2: *Social Factors*, Oxford, Blackwell.
Landry, Rodrigue/Bourhis, Richard Y. (1997), *Linguistic Landscape and Ethnolinguistic Vitality. An Empirical Study*, Journal of Language and Social Psychology 16(1), 23–49.
Lavandera, Beatriz (1978), *Where Does the Sociolinguistic Variable Stop?*, Language in Society 7(2), 171–182.
Ledgeway, Adam/Maiden, Martin (edd.) (2016), *The Oxford Guide to the Romance Languages*, Oxford, Oxford University Press.
Lipski, John M. (2011), *Socio-Phonological Variation in Latin American Spanish*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 72–97.
LSA (2009), *The Linguistic Society of America Ethics Statement*, https://www.linguisticsociety.org/resource/ethics (last access 13.02.2018).
Lynch, Andrew (ed.) (forthcoming), *Spanish in the Global City*, London, Routledge.
Maiden, Martin/Smith, John Charles/Ledgeway, Adam (edd.) (2011–2013), *The Cambridge History of the Romance Languages*, vol. 1: *Structures*, vol. 2: *Contexts*, Cambridge, Cambridge University Press.
Mar-Molinero, Clare/Paffey, Darren (2011), *Linguistic Imperialism: Who Owns Global Spanish?*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 747–764.
Martin, Philippe (2015), *The Structure of Spoken Language: Intonation in Romance*, Cambridge, Cambridge University Press.
Massot, Benjamin/Rowlett, Paul (edd.) (2013), *L'Hypothèse d'une diglossie en France*, Special Issue, Journal of French Language Studies 23.
McAuley, Daniel (2017), *L'Innovation lexicale chez les jeunes des quartiers urbains pluriethniques: "C'est banal, ouèche"*, in: Mireille Bilger/Françoise Mignon/Laurie Buscail (edd.), *Langue française mise en relief: Aspects grammaticaux et discursifs*, Perpignan, Presses Universitaires de Perpignan, 175–186.
Medina-Rivera, Antonio (2011), *Variationist Approaches: External Factors Conditioning Variation in Spanish Phonology*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 36–53.
Millardet, Georges (1910), *Petit Atlas linguistique d'une région des Landes*, Toulouse, Privat.
Milroy, Lesley ([2]1987, [1]1980), *Language and Social Networks*, Oxford, Blackwell.
Milroy, Lesley/Milroy, James (1978), *Belfast: Change and Variation in an Urban Vernacular*, in: Peter Trudgill (ed.), *Sociolinguistic Patterns in British English*, London, Arnold, 19–36.
Milroy, Lesley/Gordon, Matthew (2003), *Sociolinguistics. Method and Interpretation*, Oxford, Wiley.
Mondada, Lorenza (2005), *Constitution de corpus de parole-en-interaction et respect de la vie privée des enquêtés: Une démarche réflexive. Rapport sur le projet "Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité" financé par le Programme Société de l'Information/Archivage et patrimoine documentaire*, http://icar.univ-lyon2.fr/projets/corinte/documents/Projets/rapport_juridique_mond05.pdf (last access 13.02.2018).
Mooney, Damien (2015), *Confrontation and Language Policy: Non-Militant Perspectives on Conflicting Revitalisation Strategies in Béarn, France*, in: Mari C. Jones (ed.), *Policy and Planning for Endangered Languages*, Cambridge, Cambridge University Press, 153–170.
Mooney, Damien (2016), *Southern Regional French. A Linguistic Analysis of Language and Dialect Contact*, Oxford, Legenda.
O'Rourke, Bernadette/Ramallo, Fernando (2013), *Competing Ideologies of Linguistic Authority amongst New Speakers in Contemporary Galicia*, Language in Society 42, 287–305.

Padilla, Beatriz/Azevedo, Joana/Olmos-Alcaraz, Antonia (2015), *Superdiversity and Conviviality: Exploring Frameworks for Doing Ethnography in Southern European Intercultural Cities*, Ethnic and Racial Studies 38, 621–635.

Parry, Mair (1991), *Evoluzione di un dialetto*, Rivista italiana di dialettologia 14, 7–39.

Pooley, Tim (1996), *Chtimi: The Urban Vernaculars of Northern France*, Clevedon, Multilingual Matters.

Pooley, Tim (2009), *The Immigrant Factor in Phonological Levelling,* in: Kate Beeching/Nigel Armstrong/Françoise Gadet (edd.), *Sociolinguistic Variation in Contemporary French*, Amsterdam/Philadelphia, Benjamins, 63–76.

Poplack, Shana (2011), *Grammaticalization and Linguistic Variation*, in: Bernd Heine/Heiko Narrog (edd.), *Handbook of Grammaticalization*, Oxford, Blackwell, 209–224.

Pusch, Claus/Raible, Wolfgang (edd.) (2002), *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache/Romance Corpus Linguistics: Corpora and Spoken Language*, Tübingen, Narr.

Samper Padilla, José Antonio (2011), *Sociophonological Variation and Change in Spain*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 98–120.

Sankoff, Gillian (ed.) (1980a), *The Social Life of Language*, Philadelphia, University of Pennsylvania Press.

Sankoff, Gillian (1980b), *Above and beyond Phonology in Variable Rules*, in: Gillian Sankoff (ed.), *The Social Life of Language*, Philadelphia, University of Pennsylvania Press, 81–93.

Sankoff, Gillian/Blondeau, Hélène (2007), *Language Change across the Lifespan: /r/ in Montreal French*, Language 83(3), 560–588.

Sankoff, Gillian/Thibault, Pierrette (1980), *The Alternation between the Auxiliaries "avoir" and "être" in Montréal French*, in: Gillian Sankoff (ed.), *The Social Life of Language*, Philadelphia, University of Pennsylvania Press, 311–346.

Sankoff, Gillian/Vincent, Diane (1980), *The Productive Use of "ne" in Spoken Montréal French*, in: Gillian Sankoff (ed.), *The Social Life of Language*, Philadelphia, University of Pennsylvania Press, 295–310.

Schilling-Estes, Natalie (2002), *Investigating Stylistic Variation*, in: Jack Chambers/Peter Trudgill/Natalie Schilling-Estes (edd.), *The Handbook of Language Variation and Change*, Oxford, Blackwell, 375–401.

Schwenter, Scott A. (2011), *Variationist Approaches to Spanish Morphosyntax*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 123–147.

Serrano, María José (2011), *Morphosyntactic Variation in Spain*, in: Manuel Díaz-Campos (ed.), *The Handbook of Hispanic Sociolinguistics*, Oxford, Blackwell, 187–204.

Soria, Claudia (2015), *Assessing the Effect of Official Recognition on the Vitality of Endangered Languages: A Case Study from Italy*, in: Mari C. Jones (ed.), *Policy and Planning for Endangered Languages*, Cambridge, Cambridge University Press, 123–137.

Swiggers, Pierre (2010), *Mapping the Romance Languages of Europe*, in: Peter Auer et al. (edd.), *Language and Space: An International Handbook of Linguistic Variation*, Berlin/New York, De Gruyter, 269–300.

Tagliamonte, Sali A. (2012), *Variationist Sociolinguistics. Change, Observation, Interpretation*, Oxford, Blackwell.

Trimaille, Cyril (2004), *Études de parlers de jeunes urbains en France: Éléments pour un état des lieux*, Cahiers de sociolinguistique 9, 99–134.

Trimaille, Cyril/Billiez, Jacqueline (2007), *Pratiques langagières de jeunes urbains: Peut-on parler de "parler"?*, in: Enrica Galazzi/Chiara Molinari (edd.), *Les Français en émergence*, Frankfurt/Bern, Lang, 95–109.

Trudgill, Peter (1972), *Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich*, Language in Society 1(2), 179–195.

Trudgill, Peter (1974), *The Social Differentiation of English in Norwich*, Cambridge, Cambridge University Press.
Trudgill, Peter (1986), *Dialects in Contact*, Oxford, Blackwell.
Trudgill, Peter (2002), *Sociolinguistic Variation and Change*, Edinburgh, Edinburgh University Press.
Verjans, Thomas (2014), *"Système de possibilités" et changement linguistique*, in: Wendy Ayres-Bennett/Thomas M. Rainsford (edd.), *L'Histoire du français: États des lieux et perspectives*, Paris, Classiques Garnier, 305–320.
Vietti, Alessandro (2002), *Analisi dei reticoli sociali e comportamento linguistico di parlanti plurilingui*, in: Silvia Dal Negro/Piera Molinelli (edd.), *Comunicare nella torre di Babele: Repertori plurilingui in Italia oggi,* Roma, Carocci, 43–61.
Völker, Harald (2009), *La Linguistique variationnelle et la perspective intralinguistique*, Revue de linguistique romane 73, 27–76.
Wardhaugh, Ronald/Fuller, Janet M. ([7]2015, [1]1986), *An Introduction to Sociolinguistics*, Oxford, Blackwell.
Weinreich, Uriel (1954), *Is a Structural Dialectology Possible?*, Word 10, 388–400.
Williams, Ann/Kerswill, Paul (1999), *Dialect Levelling: Change and Continuity in Milton Keynes, Reading and Hull*, in: Paul Foulkes/Gerard Docherty (edd.), *Urban Voices*: *Accent Studies in the British Isles*, London, Arnold, 141–162.
Zentella, Ana Celia (ed.) (2009), *Multilingual San Diego: Portraits of Language Loss and Revitalization*, San Diego, University Readers.

Methodological issues

Gabriel Bergounioux, Michel Jacobson and Paola Pietrandrea

1 Annotating oral corpora

Abstract: Focusing primarily on oral corpora, this chapter examines annotation as a means of standardizing transcription, identification and categorization. Annotation is a sequence of characters inserted into a text to annotate a particular phenomenon. Annotation is performed at the start of an operational workflow in order to enrich and document the contents. This chapter identifies three types of annotation, based on the degree to which they interact with the base file: embedded/online; stand-off/standalone; multi-tiered/ interlinear. It also argues that sociolinguistic annotation suffers from an absence of consensus with regard to the categories that should occur as tags.

Keywords: oral corpora, transcription, annotation, tagging, hearer variation

1 Introduction

Spoken language has been studied from two opposing directions: (i) the analysis of speech as a signal; experimental phonetics, for instance, is intrinsically linked to dialectology and to areal variation; and (ii) anthropological research which focuses on content (what is said) rather than on form (how something is said).

The study of linguistic variation, at all levels, requires large sets of data and internal comparison of the data. As such, sociolinguistics has given rise to a methodology for languages with a written tradition that combines philological techniques of text collection with data collection methods from field linguistics. The *corpus* has been developed as the best way to present these collected resources in a systematic way. The range of different types of corpora has required standardization of the principles of *transcription*, of *identification* and of *categorization*. These principles have come to be standardized in the form of annotation as a technique for quantification and taxonomy.

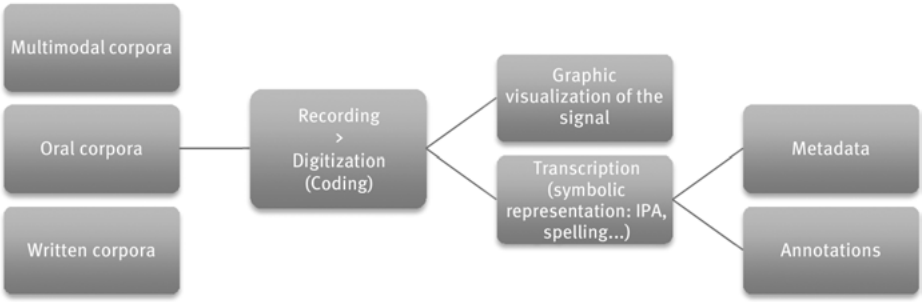
Although this chapter adopts a broad definition of *annotation* that includes transcription, we will examine the practice in a narrower sense, i.e. as the insertion of linguistic information in a corpus using standardized and explicit rules.

To start with, we can distinguish three broad types of corpora: (i) written; (ii) multimodal; and (iii) oral.

It is the research aims that govern the data format, which in turn conditions the type of representation chosen. Multimodal corpora are essential for studies on children's language acquisition, whilst written corpora are necessary for philological studies. This chapter focuses on oral corpora.

https://doi.org/10.1515/9783110365955-002

Figure 1: Operational workflow



The chart in Figure 1 can be completed (i) by considering the range of transcription forms with regard to the speech that has been produced, from IPA to logograms; and (ii) including the subsequent steps of tools, archiving, distribution and analysis.

Understood in this way, annotation, along with transcription, can be taken both as a (more or less automated) technique and a (predominantly empirical) method that lies at the intersection of the humanities (surveys, fieldwork) and computer science, particularly Natural Language Processing (NLP, with regard to automation). From both a humanities and computer science perspective, it is necessary to explain the principles behind the annotation, i.e. the adoption of a procedure, if only to distinguish different uses that appear in the same form.

Example 1: Three annotations for an utterance

Transcription: "Y a pas de bug"

Annotations: (i) <Pronounced: /byg/>

(ii) <Alternative: "bugs">

(iii) <POS: N / Emprunt / Informatique>

This article focuses, in turn, on (i) the creation and processing of corpora; (ii) the annotation itself, including the role of the researcher, the process, the tools and the typology used; and (iii) an evaluation of the current state of practice and suggestions for ways forward.

2 Principles of annotation

2.1 What is annotation?

In the ordinary sense of the word, annotation is a sequence of characters inserted into a text, whether that text has been transmitted directly in written form or has been transcribed from oral data. In the first case, the act of annotation falls under the

philological traditions of marginalia, glosses and scholia, i.e. of comments with variable functions such as correcting a word or adding a translation or opinion. In this respect, the annotation is a later addition to the document and remains separate from it.

Transcription of oral corpora can in fact be considered as the first form of annotation of the source data. In languages with a written tradition, at least for the Romance languages, a very large number of speakers have direct access to this means of representation.

In linguistics, annotation tends to have a rather more restricted meaning and is used to refer to a set of tags, located outside of the text itself but within the written document. There are two approaches to annotation. The first approach, inherited from editing, inserts the tags into the running text. Suppressing the annotations in this case does not prevent a return to the original text. The second approach, used for instance in relational databases, structures the document itself in such a way that the original content of the text can only be distinguished from the added comments once the tags have been interpreted.

2.2 Why annotate?

Annotation is necessary whenever the editor of a written document or the producer of a transcription needs to make note of a particular phenomenon or to draw the reader's attention to a particular phenomenon. Adding footnotes (or references) is the primary scientific tool used in philology, playing on a visual distribution that dissociates the text from the commentary. In oral corpora, preference is given to a multilinear notation that separates different levels – as Boas (1911) demonstrated for aligned translations – distinguishing the written representation of the utterance and the indications provided by the editor.

This manner of proceeding, which is obligatory for languages with only an oral tradition was also implemented for the spoken form of written languages. In the preparation of sociolinguistic or dialectal corpora, the transcription of language acquisition data (in the mother tongue or as a learner) or data on language disorders meant compensating in writing for certain types of information that were lost in the conversion to graphical form, such as all the indications provided by the signal itself.

2.3 How to annotate

Various developments in computer science such as:

- proprietary languages such as IBM's GML (Generalized Markup Language),
- the normalization of a generic SGML format (Standard Generalized Markup Language) within ISO (the International Organization for Standardization),

– and lastly, in 1996 the appearance of the first XML specifications (eXtensible Markup Language)

enabled the separation of:
– the logical structure of the document that can or must be defined within a schema such as DTD (Document Type Definition) for SGML, and
– the representation (or physical structure) of the document whose derivational rules can be specified in a style sheet.

This distinction between the two levels was intended to clarify, in the editing process, the respective roles of the different actors. Producing schemas and style sheets is the editor's task in order to ensure better control over the validation and formatting processes. The use of these tools by the writers assists them in inputting the data.

SGML gave birth to different applications. HTML (*Hypertext Markup Language*), EAD (*Encoded Archival Description*) and TEI (*Text Encoding Initiative*) are the most widely-used applications in digital humanities. XML is more readily compatible with the internet.

2.4 Metadata and annotations

Annotating a text involves adding information expected to be lacking, but potentially useful or even necessary, for the intended audience. Linguistic practice has been transformed by digital file transfer. There are two complementary strategies: either (i) the information is given by the metadata, recorded in a separate file; or (ii) the information is inserted into the text itself.

The distribution of data as either metadata or annotation is variable: if the encoded information relates to the whole of the corpus or text (e.g. speaker ID, situation, text genre), then it will generally be placed in the metadata. In contrast, if the information relates to a word (or a short sound sequence such as a click, filler, false start or idiom), then it will be encoded through annotation.

The need to add certain information in a form that can be exploited (for example lemmatization or frequency statistics) led to the introduction of the notion of *tagging* in linguistics. An example of a written tagged corpus is the Brown corpus (Greene/Rubin 1971), and an example of an oral tagged corpus is the LLC (London-Lund Corpus) taken from the Survey of English Usage Corpus (Crystal/Quirk 1964). In the field of Romance linguistics, the first example of tagging in an oral corpus was in Italian on the LIP corpus (De Mauro et al. 1993).

The example below, taken from the digital version of the London-Lund Corpus, shows coding of information of different natures that can be considered as tagging. These tags are, in order of appearance on each line: *Text category*, *Text within category*, *Identifier*, *Tone unit number* [...], *Speaker identity* [...], *Text*. The transcription

(*Text*) uses written conventions to indicate non-verbal elements (laughter, telephone ringing, etc.), pauses, intonation, etc. (for a list of conventions, see: http://clu.uni.no/icame/london-lund/index.htm).

Example 2

```
1 3   9 1420 1 1  b   20*[mhm]* /
1 3   9 1410 1 1(A   11^th\en they _said# /
1 3   9 1430 1 1 A   11well "^now that you`ve done th/ese# /
1 3   9 1440 1 1 A   11and they`ve been "^s\o succ/essful# /
1 3   9 1450 1 1 A   11we`d ^like you to do our s\uper# /
1 3   9 1460 1 1 A   11^alpha:m\atic# /
1 3   9 1470 1 1 A   11or ^s/omething# /
1 3   9 1480 1 1 A   11and ^this is one of th/ese# /
1 3   9 1490 1 1 A   11that ^goes s/ideways# /
1 3   9 1500 1 1 A   11and ^fr/ontwards# /
1 3   9 1510 1 1 A   11and em^br/oiders# /
1 3   9 1520 1 1 A   11and *^d/arns# /
1 3   9 1530 1 1 A   11and sews* ^b\uttons on# /
1 3   9 1540 1 1 b   20*( - laughs) yes* /
1 3   9 1550 1 1(A   11- - and I ^s=aid# /
```

In the example below, drawn from the *Vienna-Oxford International Corpus of English* (VOICE: see http://ota.ox.ac.uk/desc/2542), the formatting of the text into lexical units or the markup on each unit of its form, lemma and part of speech uses XML syntax with conventions from the *Text Encoding Initiative* (TEI), an international programme that standardizes the principles of formatting and exchange of digital texts within ISO standards.

Example 3

```
<u who="#LEcon351_S5" xml:id="LEcon351_u_17">
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#NPfNP" lemma="austria">austria</w>
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#RBfRB" lemma="very">very</w>
    <w ana="#JJfJJ" lemma="cold">cold</w>
    <w ana="#CCfCC" lemma="and">and</w>
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#PAfPA">_0</w>
    <w ana="#JJfJJ" lemma="hot">hot</w>
```

```
    <w ana="#RBfRB" lemma="enough">enough</w>
    <w ana="#PAfPA">_0</w>
</u>
```

As we can see in these two examples, the annotation of a text or transcription takes the form of enriching the original information with new types of information that, depending on the era, have used different transcription conventions, such as structured text or XML syntax.

2.5 Range of approaches

Proposals by computer scientists were not always easy to reconcile with the expectations of linguists, which explains certain delays (Léon 2015). Enabling researchers from mathematics (logic or calculus) or from electronics to collaborate with philologists or anthropologists was not always self-evident.

In linguistics, the act of using corpora had a different meaning depending on whether the corpus was understood as the record of a culture or as a resource for linguistic analysis. Requirements for annotation differ noticeably depending on whether corpora are understood as a reservoir of examples or a set of data to explain, whether they are used to criticize theories or for empirical approaches such as the "usage-based model" in construction grammar.

A further difference lies in the manner in which annotations are produced. Annotations can be produced manually, semi-automatically (pre- or post-edition) or automatically. Automatic annotation, whether a symbolic or machine-learning method is used, presupposes an underlying theory that provides the relevant tools for annotation and pre-annotation and allows the results to be evaluated (Fort 2012). Manual annotation is considered the most reliable but cannot be extended to large quantities of data and, if the effort is collective, raises problems of inter-annotator agreement. Automatic methods are primarily judged in terms of the number of corrections required when it comes to post-editing the results.

For example, in order to transcribe the corpus *Enquête Sociolinguistique à Orléans* forty years after data collection, the cost of manual transcription was compared to that of automatic transcription plus post-editing. In the end, it was decided not to choose automatic transcription, because the time gained by automatic transcription was lost through the effort required during the manual post-editing stage.

In the *Enquête Sociolinguistique à Orléans* corpus, organizing the work led to the definition of several levels of annotation:

– Level zero (T0) uses minimal conventions and has as its main objectives: (i) to enable navigation within the signal (synchronized to enable multiple replays); and (ii) to suggest a transcription for each word, including false starts and disfluencies, with encoding similar to normal writing conventions in order to

make reading and editing easier. A rule in the *Guide de transcription* specifies that a transcriber must not listen to the same segment more than twice.
– Level one (T1) aims to produce a transcription that can be exploited for advanced linguistic analysis by providing a fine-grained written transcription including corrections and theoretical choices, adding specific coding for prosody, multi-transcription, etc., and providing a multi-tiered annotation.

For annotation in T0, three transcriptions were systematically produced:
– an A version, i.e. a raw transcription undertaken as quickly as possible (cost: 10 times the listening time);
– a B version, i.e. rereading the A version by another transcriber (cost: 5 times the listening time);
– a C version, i.e. correction of the B version by an experienced annotator (cost: 5 times the listening time).

All three versions are preserved so that differences between individuals and groups can be studied.

The difficulty of harmonizing the work of different researchers engaged with the same corpus can be found on a much larger scale as soon as it is a question of bringing together annotations across languages, countries and types of data. There is currently a debate between those who advocate good practice and those who support normalization driven by *a priori* universal principles such as how to define a standard respecting both the constraints of the object of study and scientific principles. For an example of standardization using current ISO norms under the auspices of the TEI, see Stührenberg (2012).

3 Key criteria in creating corpora

Annotation is performed at the start of an operational workflow that gives rise to the enrichment of the contents and their possible exploitation (e.g. counting, analysis, patterning, collocation extraction, concordances). The consideration of examples and the running of training tests in NLP open up questions about the granularity of attributes and data sampling.

3.1 Size

At what point can we consider that a sample of a language is representative of all its uses? This question, which has been raised for the lexicon (see for instance the creation of BASIC English, Ogden 1930; or of *Français fondamental*, Gougenheim et al. 1964) and for syntax, notably in generative grammar, took as points of reference

frequency of use, the structure of the language and their correlation (Zipf's law). Social variation, needing to integrate a range of parameters, i.e. dialectal, diaphasic and diastratic, proved to be less secure in its classification criteria and more demanding quantitatively. The quantitative aspect is the first one that should be considered, as it conditions all the other aspects. The increase in computer memory and the decrease in hardware costs have enabled the requirements in this domain to be progressively raised (see tables).

Table 1: Written corpora

Brown Corpus	1,000,000 words	1961
Frantext (French)	300,000,000 words	1975
British National Corpus	100,000,000 words	1995
CORIS-CODIS (Italian)	130,000,000 words	2001
Corpus de Referencia del Español Actual	160,000,000 words	2008
Reference Corpus of Contemporary Portuguese	300,000,000 words	2012
frTenTen (French)	10,000,000,000 words	2012

Table 2: Oral corpora

ESLO 1 and 2 (French)	7,000,000 words	1969
London Lund Corpus of Spoken English	500,000 words	1990
LIP (Italian)	500,000 words	1993
CLAPI (French)	2,500,000 words	2005
Corpus de Referencia del Español Actual	9,000,000 words	2008
Reference Corpus of Contemporary Portuguese	1,600,000 words	2012

We could also mention the University of Barcelona's *Corpus de Català Contemporani*, the *SyMiLa* Occitan corpora at the Université de Toulouse Jean-Jaurès and the *Thesoc* corpus at the Université de Nice Sophia Antipolis.

3.2 Balancing different considerations

A second criterion relates to the balance between different practices and different contexts. As a general rule, a collection of written documents is less an absolute reflection of the language than of certain uses of that language, e.g. literary uses such as *Frantext*, or journalistic uses such as the corpus from *Le Monde* newspaper or the

French corpus at the University of Leipzig (see Wortschatz/French Leipzig). The choice of resources is determined by accessibility. It is easier for instance to start by working on typed texts than on handwritten texts, and on documents that are understandable by those involved in processing the text rather than on medical texts for example, without actually excluding the less accessible texts from the investigation.

Apart from music, most audio documents that are stored in sound archives relate either to exotic languages – where they compensate for the lack of written documents – or to official public speeches. Creating oral corpora has allowed techniques from field linguistics and dialectology to be applied to everyday urban conversations in a manner approaching sociological methods.

The situation differs depending on the language and country. French corpora appear to be less varied than those of other Romance languages. Excluding some uses of a language reflects different power relations. Catalan and Spanish, European Portuguese and Brazilian Portuguese, Italian and its dialects, the great difference between written and spoken French – all represent case studies that contribute to the discussion on how to define a *reference corpus*. The answers differ from one country to another and sometimes lead to paradoxical results. For instance, LIP (Corpus *Lessico di frequenza dell'Italiano Parlato*) which aimed to provide an inventory of the different uses of spoken Italian, in fact demonstrated a tendency towards standardization.

3.3 Conditions of use

Availability of the corpus is a major obstacle, either because the researchers (or institutions) do not allow distribution of the corpus, or because no means of preservation has been guaranteed (see the investigations of the *Groupe Aixois de Recherche en Syntaxe* – GARS). Gaining authorization to access the text – whether that authorization falls under copyright or the protection of informants – is a particularly difficult obstacle to overcome in the case of spoken corpora as the possibility of vocal recognition infringes on anonymization and older studies did not seek the consent of the speakers recorded. The same goes for metadata. The possibility of matching recordings to speakers, situations, dates, etc. that are available in the form of a dataset puts gathering the information necessary to do research face to face with strict legal requirements.

Another difficulty is inherent to the formats and tools used. Whether it is material or support, software or language, the obsolescence of equipment and systems remains a challenge. The choice of available and lasting formats and tools is even more complex in the case of audio documents where the primary source itself may disappear or become inaccessible.

4 Uses of annotation

In the usual understanding of the term, the secondary character of annotation and its interpolation distinguishes annotation on the one hand from transcription, and on the other hand from metadata. Leech (2005) has suggested characterizing annotation in terms of the opposing pairs {transcription vs annotation} and {representation vs interpretation}.

In linguistics, the goal of annotation is to integrate information in the corpus using descriptors that make searches conducted for a particular piece of research either more efficient or, indeed, even possible (see example 4).

Example 4: Annotation image

```
<anchor id="u-trigger-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il m'a dit
<anchor xml:id="u-trigger-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
<anchor xml:id="u-target_portion-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il travaillait pas
<anchor xml:id="u-target_portion-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
```

4.1 Transcription formats and instructions

The first step in the annotation process is to specify the data involved in the working hypothesis. Certain types of data are more widely used than others, e.g. identification of *Parts of speech* (POS), of morphemes or of semantic properties. No type of data is obligatory or excluded as long as it records properties of the language.

Once the type of data has been decided on, it is necessary to fix the conventions that put these data in a format that is simultaneously (i) distinctive, i.e. defines as many classes as required by the research; (ii) extensive, i.e. expands the annotation to the level of detail required; (iii) unitary, i.e. ensures that the same phenomenon is consistently described in the same way; (iv) economical, i.e. no element in the annotation is redundant or superfluous; and (v) explicit, i.e. every element of annotation has to be identified in an associated document.

Encoding should (i) be limited in terms of number of characters; (ii) rank the information provided in a hierarchy; and (iii) be accessible, i.e. save time during the training process and during memorization. In practice, this last condition means using well-established abbreviations, e.g. /N = noun/ in POS tagging.

Annotation is time-consuming. It can be undertaken by people other than the researcher due to its repetitive nature. The time spent in annotating is gained, however, during the exploitation phase, where the trend is for the speed of execution to be inversely proportional to the time of preparation as shown in the *Computational Analysis of Present-Day American English* (Kučera/Nelson/Carroll 1967) on the Brown Corpus, which explored a range of linguistic, psychological, statistical and sociological elements in the corpus, or the sociological analyses on the ESLO corpus

(Bergounioux 2016). The larger the corpus and the more the advantages of annotation can be used in other studies and by other researchers, whether linguists or not (e.g. statisticians, computer scientists, sociologists), the larger the profit. Also important is the definition of criteria that permit both critical analysis and a reassessment of earlier choices.

The manner in which the tasks are carried out is defined by the research aims, the selection of data and of the properties associated with the data, and by the choice of strings of characters that index the data using tags. Annotating a corpus also requires writing a manual or a guide that makes explicit the rules followed when creating the corpus, and provides comments. An example of such a manual can be found at http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V4_mai2013.pdf for the ESLO corpus.

4.2 Criticisms and sociolinguistic uses

The first type of criticism that has been levelled against annotation concerns the access to documents. Annotation increases the file size and the finished product can be off-putting as annotations clutter up the text and break up the linearity of the original text. In particular, there is a great deal of variation from one programme to another – often the result of conflicting directions and approaches – and, within the same programme, from one annotator to another. Moreover, mistakes made in each of the phases reduce the reliability of the whole document, meaning that to improve the quality of the results, a further phase of post-editing and correction is required. However, this extra stage then reduces the comparative advantage in terms of time.

Sociolinguistic annotations label characterizations linked to change (of language or dialect, of use or register), non-standard uses or uses that are innovative with respect to the norm and typical forms of a culture or subculture. These concern: (i) information about dialectal features, code-switching, etc.; (ii) details about register – as generally used in dictionaries or in uses that are characteristic of a social milieu or an age group, etc.; (iii) the marking of errors such as incorrect constructions or hypercorrection; (iv) the means by which subjective categories (positive and negative terms, classifications based on a particular social group) and expressions of identity are included; and (v) forms of address and reformulations, etc.

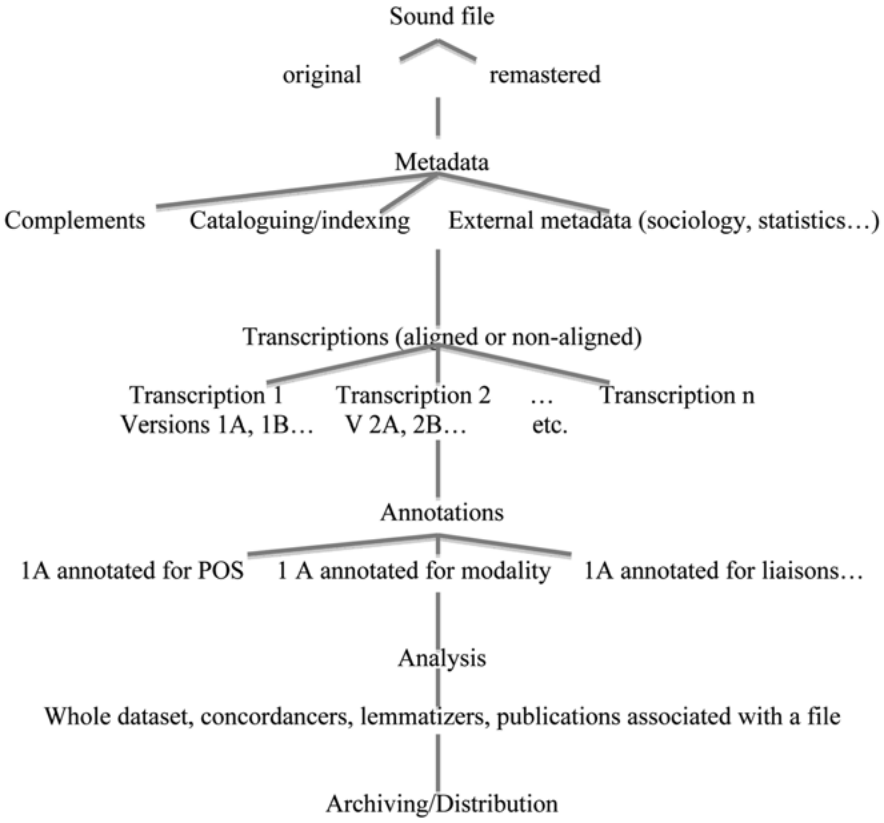
There are two parameters at issue in all of these cases. The first parameter relates to the manner in which properties of the language are realized in speech. The second parameter focuses on how a social group expresses itself, i.e. in terms of its cultural properties. Whatever the content of the annotation, all encoding lies at the intersection of these two domains.

5 Annotation instructions

Certain types of annotation are specific to written language, e.g. spelling, and some to speech, e.g. liaison. Other types of annotation apply to both written and oral language, e.g. POS, modality, language register. Despite apparent distinctions, different types of annotation are homogeneous regardless of the type of data collection – from retrieving data of a widely-used language from the Web to short studies on small linguistic groups without a written tradition.

5.1 File generation

Figure 2: File generation and the place of annotation



As a general rule, annotation is conceived of as an operation that takes place *a posteriori* on a text or transcription, after metadata has been provided and before scientific analysis. This conception of annotation situates it at a later stage than data

formatting and at a prior stage to the different types of analysis. The analyses, which are dependent on instrumentation, are achieved through a version of transcription that can continue to evolve in parallel (versioning). The division proposed by Habert (2005) between *instruments* and *tools* is arbitrated through the possibility of automating the annotation and the database management system (DMS).

Correspondance between products can be achieved in the following manner (each different file is identified by the first number, different versions of the same file have an extra letter added, and enrichments are identified by means of a number after the letter).

Table 3: Operation and production

Operation	Product
gathering/digitizing of the signal	Digital audio file
cataloging/indexing/metadata	Text file 1
transcription/coding/aligning	Text file 2 (versioning 2A, 2B, 2C…)
annotation/instrumentation	Text file 2 annotated (ex. 2B1, 2C1, 2C2…)
tools/analysis	Research investigations
maintenance/distribution	Text files 1 and 2 (with a hyperlink to analyses)

5.2 Three types of annotation

If it is possible to get to the point of analysing, archiving or distributing a corpus without carrying out annotation, is it possible to anticipate certain types of annotation when writing the metadata or transcribing the recordings? There is an intersection between annotation and what should be contained in the metadata. Thematic divisions (*topics*) can appear in a separate file or be inserted as an annotation. In general, metadata includes global *external* information such as the identity of speakers, the date of recording, digital formats or owners. In contrast, annotations primarily concern *internal* information that relates to short elements included in the signal (e.g. a noise, a click), and concerns units that range from false starts and words (with tagging and lemmatization) to the phrase (coreference, tree diagrams) or to a speech turn.

We can distinguish three types of annotation, based on the degree to which they interact with the base file.

Type 1: Embedded/online

Example 5

```
<annotatedU end="#T175" start="#T174" wh="spk1" xml:id="au72">
    <u>
        <seg xml:id="s343">alors ils faisaient comme ça euh <pause type="-
        short"/>et je me suis
        rendue compte que ça n'allait pas </seg>
        <anchor synch="#T183"/>
        <seg xml:id="s344">parce que moi je je lisais et je lisais un rigue </seg>
        <seg xml:id="s345">euh la première ligne </seg>
    </u>
    <spanGrp>
        <span type="com" target="#344">mot italien = ligne</span>
    </spanGrp>
</annotatedU>
```

(taken from http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=myauto-links:exemples_codage_teiml.pdf

Type 2: Stand-off/Standalone

Example 6

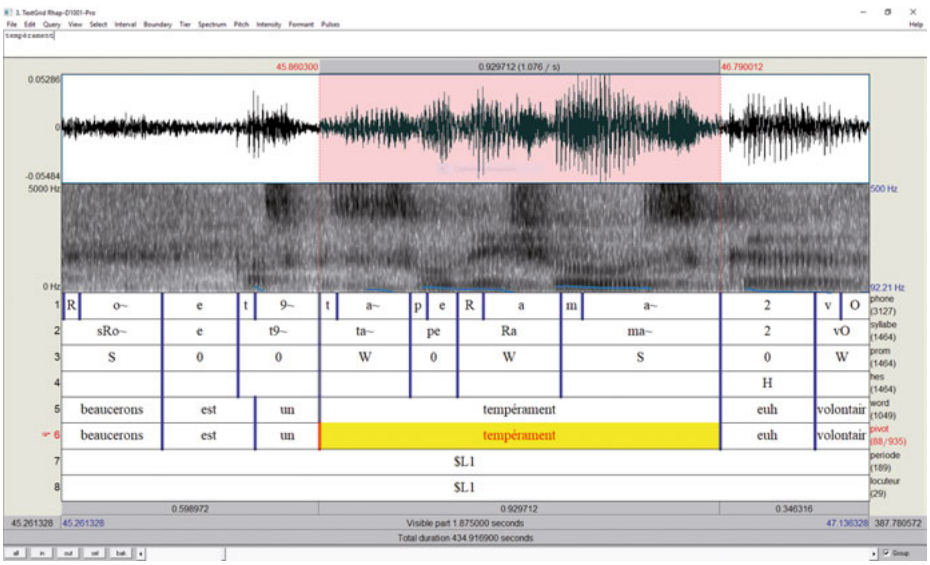
```
<texte>alors ils faisaient comme ça euh et je me suis
rendue compte que ça n'allait pas parce que moi je je lisais et je lisais un rigue
euh la première ligne</texte>

<annotatedU end="#T175" start="#T174" wh="spk1" xml:id="au72" xmlns:
xi="http://www.w3.org/2001/XInclude">
    <u>
        <seg xml:id="s343">
            <xi:include href="texte.xml" xpointer="xpointer(substring(., 1, 32))"/
            >
            <pause type="short"/>
            <xi:include href="texte.xml" xpointer="xpointer(substring(., 33,
            48))"/>
        </seg>
        <anchor synch="#T183"/>
        <seg xml:id="s344">
            <xi:include href="texte.xml" xpointer="xpointer(substring(., 82,
            48))"/>
```

```
            </seg>
            <seg xml:id="s345">
                <xi:include href="texte.xml" xpointer="xpointer(substring(., 130,
                30))"/>
            </seg>
            <spanGrp>
                <span type="com" target="#344">mot italien = ligne</span>
            </spanGrp>
        </u>
    </annotatedU>
```

Type 3: Multi-tiered/interlinear

Example 7



Embedded annotation results in a self-sufficient document because the whole set of annotations are collected together within a single document. This can reduce legibility, however. In contrast, stand-off annotation increases legibility of the master document but makes the stand-off annotations dependent on the document that contains them. This dependency between the documents requires coordination and can make maintenance more complex. Indeed, dependency relies on logical links (positions or identifiers) that have to be preserved throughout any changes that might take place in the data. Multi-tiered or interlinear annotations are used when several points of view or levels of analysis are required for the same phenomenon within a single annotation.

A non-phonetic transcription anticipates annotation, even if it is only by segmenting the signal into words or applying rules that are not representative of the orthography of the person being interviewed. For instance, is one justified in indicating plural agreement on the past participle in a phrase such as *les soucis que ça m'a faits* if the speaker has a low level of education? The answer depends on the principles that have been set out in the transcription manual corresponding to the corpus.

5.3 Instructions

Annotating a corpus needs to meet a number of interdependent requirements:
- Separating the file that is being worked on (transcription file and audio file) from the annotated files. *Separability* implies *reversibility*, i.e. it is necessary to be able to move between an annotated file and the preceding version.
- Ensuring the *replicability* of files and their preservation in a state that allows *validity*, i.e. identity and legibility in the evolving formats of computer science.
- Guaranteeing *accessibility* to data that need to be quickly retrievable, i.e. *traceable*, and available in a form that allows easy manipulation by prioritizing intuitivity and affordance.
- Ensuring the longevity of the resource and its interoperability.
- Facilitating the application of tools and *instrumentation*, and allowing *incrementation*.

Conceived of in this way, annotation is a process that, at each stage, extends the resource by transforming it without ever erasing a previous state (*versioning*).

6 Annotation formats

Annotation is an operation that is carried out with a particular aim in mind. It is never a primary state of the data. Rather, annotation is carried out on a resource that has already been specially prepared. Annotation can be visualized in different ways, often – but not exclusively – in written form, *linear* and *segmented*. The text file offers page layout, global volume (number of characters or spaces), metadata, etc. In this way, annotation is both a method of analysis and an NLP tool. The spectrum of representation formats goes from an analogue format (for phonetic analyses) to symbolic notations (for syntactic or stylistic studies). Sociolinguistic studies tend to favour aligning the transcription with the signal and providing extensive comments within tags. This compromise ensures maximum legibility by distinguishing between an orthographic tier and information that can be retrieved via queries in order to carry out analyses.

6.1 Principles

Leech (2005) laid down four principles on which corpus instrumentation should be based: (i) provide access to the metadata and the processing; (ii) make the choices underlying each operation explicit; (iii) ensure replicability of results, i.e. the instrument should, under identical conditions, enable the same conclusions to be reached; and (iv) be able to be verified through processes that are independent of the observer (a recurrent difficulty, one of the aims of which is to overcome inter-annotator disagreement).

Annotation is executed following conventions defined within a repesentation format that specifies: (i) segmentation into elementary units; (ii) their organization within a document; and (iii) a reversible means of inserting metalinguistic information that can be exploited automatically.

Generally, the term *tag* tends to be used to refer to technical aspects of the data processing, whilst *annotation* tends to refer to the researcher's scientific options.

6.2 Encoding and content

Annotation can be carried out in XML/TEI, a practice which is widely adopted today both for the *document* (file) and for the *schema* (see *validation* below) with a *header* that introduces either the corpus itself or one of its files.

The hierarchy comprises different levels: (i) the file as a whole; (ii) the structure of the full document; and (iii) the division of the document into paragraphs. These levels give information about four items: (i) the description of the file, equivalent to a bibliographic reference; (ii) the indication of origin that stipulates the relation to the original text; (iii) the characterization of the text containing information relevant to its use (starting with the language in which the text is written); and (iv) the versioning.

Example 8: Metadata
TEI metadata extracted from CLAPI (http://clapi.ish-lyon.cnrs.fr/V3_TEI.php)

```
<teiHeader xml:lang="fr">
    <fileDesc>
        <titleStmt>
            <title> Réunion de conception en architecture – mosaic ~ Mosaic –
            architecture ~ Mosaic – architecture – xml </title>
            <principal>Detienne Françoise</principal>
            <principal>Traverso Véronique</principal>[...]
            <respStmt>
```

```
            <resp>collecté par</resp>
            <name>Detienne Françoise</name>
            <name>Visser Willemien</name>
        </respStmt>[...]
        <respStmt>
            <resp>préparé et balisé par</resp>
            <name>CLAPI – Equipe Médiathèque</name>
        </respStmt>
    </titleStmt>
    <publicationStmt>
        <publisher>Groupe ICOR/ Plateforme CLAPI</publisher>
        <pubPlace>http://clapi.univ-lyon2.fr</pubPlace>
        <availability status="restricted">
            <licence target="http://clapi.univ-lyon2.fr/V3_CGU.php">
                <p>Conditions générales d'accès pour ce document</p>
                <p>Copyright © ICAR. Tous droits réservés.</p>
            <p> Enregistrement vidéo d'une durée de 1h18m45s téléchargeable sous convention de recherche </p>
                <p> Transcription mosaic – architecture – adaptée CLAPI au format doc en téléchargement libre </p>
                <p> Transcription mosaic – architecture – clan au format clan – ca ou cha en téléchargement libre </p>
                <p>Transcription requêtable par les outils librement</p>
                <p>Agrément CNIL de Clapi numéro : 2-12064</p>
            </licence>
        </availability>
    </publicationStmt>
```

6.3 Formatting, tagging and validation

Formatting covers two operations: (i) pre-processing that aims to reduce as far as possible the risk of noise or silence, in the sense that these terms have in information theory, by cleaning and normalizing the data; and (ii) formulating all the instructions governing the annotation.

Alongside the tagging used in processing the text, tagging that aims to label strings of characters adds information (phonetic, grammatical, semantic or sociolinguistic) and provides details that are automatically recoverable about these linguistic properties.

Transcription and tagging of the corpus are linearly connected. It is not possible at this stage to insert a hierarchical annotation in tree form, which means that such a diagram must be executed at a later stage in the data processing.

If the annotation is encoded in XML, a first level of syntactic validation can be carried out using standard validation tools that allow the document to be evaluated with respect to a model defining rules of grammar and vocabulary. Within the range of languages used to define these models, some examples include XML-Schema, RELAX NG Schema, Schematron and DTD.

6.4 Annotation and instrumentation

As annotations respect the linearity of the information sequence, they have the advantage of allowing systematic exploration of data on all or part of one or more files by specific queries adapted to the file content.

To extend the study, other forms need to be used, such as those that can be produced by tree banks, concordancers or by inference or unification (Semantic Web).

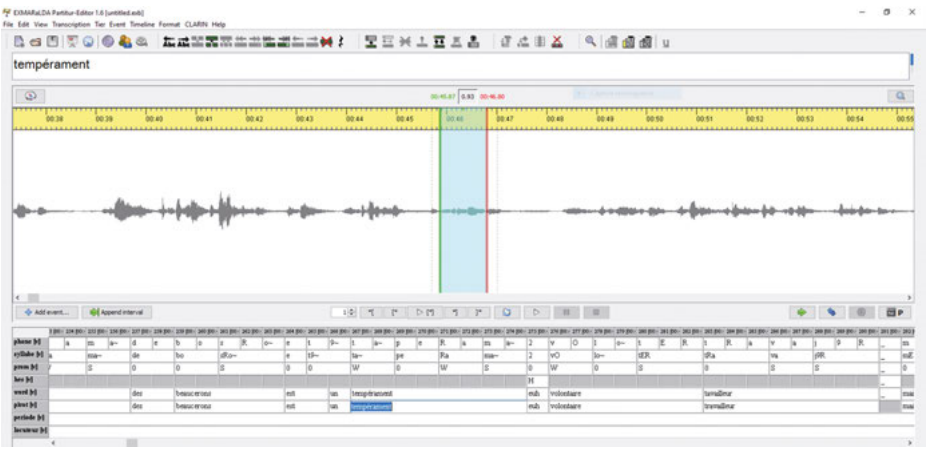
Example 9: Concordancer
Concordance of the word *poêle* – CoCoON http://cocoon.huma-num.fr/exist/crdo/

▶	je crois qu'on on met du beurre dans une	poêle	et on casse des oeufs dessus et on remue
▶	...ien je crois qu'on on met du beurre dans une	poêle	et on casse des oeufs dessus et on remue [ri...
▶	...s oeufs on sale et on met du beurre dans une	poêle	et on met les oeufs d(e)dans et on la fait c...
▶	mett(r)e les oeufs dans la	poêle	et p(u)is laisser cuire euh [pf:noise:instan...
▶	...de terre vous les faites assez cuire dans la	poêle	et p(u)is vous mettez euh les oeufs que vous...
▶	...à-dessus je remue le tout même dans dans la	poêle	et puis alors des fois j'essaie de retourn...
▶	...revenir les euh hm les cèpes dans une autre	poêle	et puis après on les met sur la sur la sur...
▶	...s oignons des patates et des lardons dans un	poêle	et puis après tu verses euh bah quand quan...
▶	...chaud je verse euh les oeufs battus dans la	poêle	et puis c'est tout
▶	...ire cuire des champignons auparavant dans la	poêle	et puis casser votre omelette dessus ou des...
▶	...fondu bon bah je je verse les oeufs dans la	poêle	et puis euh
▶	voilà un petit peu de beurre dans la	poêle	et puis euh voilà après je mets ça à cuire...
▶	...is on enfin on verse tout ça dans la dans la	poêle	et puis on tourne jusqu'à temps que ça soit
▶	... enfin on on verse tout ça dans la dans la	poêle	et puis on tourne jusqu'à temps que ça soit...
▶	...its tu verses euh tes oeufs dedans dans la	poêle	et puis tu mets un peu de persil euh et c'...
▶	... et puis après je prends du beurre dans une	poêle	et puis vas-y et
▶	... vous les f() vous mettez du beurre dans une	poêle	et puis vous faites

Previous 1 2 3 4 5 ... 8 Next

To process corpora, it is necessary to have recourse to instrumentation. Once the type of document has been defined and the model chosen (EAD, TEI, CES, etc.), and depending on the type of transcription to be carried out (phonetic with Praat, for instance, or aligned with the speech signal with Transcriber), different annotation tools can be selected. Among most widely-used tools are: (i) ANVIL for video annotation; (ii) ELAN (*Eudico Linguistic Annotator*) for multimedia files; (iii) EXMARaLDA (*Extensible Markup Language for Discourse Annotation*) for oral corpora. EXMARaLDA is a tool that covers transcription, annotation, management, searching and analysis.

Example 10: Representation under EXMARaLDA



The range of formats for corpora and the fact that they are more often produced by research teams rather than by institutions, raises essential questions about their compatibility and durability that go beyond questions of accessibility, availability and free use. A recurring question concerns the processes used to codify the metalanguage in such a way that elements with effects across different levels can be identified and used.

7 Annotation structure

7.1 Annotation, annotability, meta-annotation

The choice of annotation depends on the types of analysis that it enables – or prohibits. The occurrences corresponding to the criteria defined (phonological, morphological, syntactic, sociolinguistic, etc.) are first identified. Starting from the data, and respecting pre-established conventions, the selected units are noted in a form that allows the data to be searched. Two operations are carried out at the same time:

(i) demarcation, i.e. segmentation of the elements to be annotated (a linear string of symbols); and (ii) selection of a property of one of these elements, one of its *attributes*, i.e. a name/value pair. There are two methods, i.e. using a rule-based system or using a supervised learning system.

We refer to the open set of properties that are available for this type of operation as *annotability*, also including under this term types of processing that make implementation easier, for instance segmenting an oral corpus into phrases. Annotability is, from a computer science perspective, equivalent to *observability* from a linguistic point of view. In the case of annotation of annotations, we speak of *meta-annotation*.

7.2 Operations

Elements can be segmented at different levels, ranging from small constitutive units (letters, numbers, symbols, punctuation, spaces), to larger units such as morphemes, words, syntactic phrases, propositions and sentences, to extensive thematic units (*topics*). Nonetheless, the central unit remains the word, at least in non-agglutinative languages.

We can distinguish three stages: (i) breakdown into character strings, i.e. *tokenization*. In the case of oral corpora, this involves repeating a task carried out by the transcriber; (ii) *lemmatization*, e.g. is *avions* the 1st person plural indicative imperfect form of the verb *avoir* or the plural of the noun *avion*? Unifying forms (the lemma *aller* for *vont*, *aille* or *irions*) and their ordering into parts of speech (POS) implicitly sketches out the morphological structure. The morphology of Romance languages, and in particular French, is more restricted in speech than in writing, as they inherited Latin writing conventions and are limited in their use of prosodic distinctions; and (iii) *processing*, i.e. starting from the linguistic representation of units, it is possible to determine coreferences, to identify named entities (Eshkol-Taravella et al. 2011) and to carry out parsing and semantic analysis, to structure the units into themes, etc.

7.3 Difficulties

Amongst the possible difficulties are the compatibility of computer tools and of linguistic theories, in particular those theories that are less formalized such as sociolinguistics, and non-scriptural data, a recurring problem in NLP. Furthermore, if the researcher him/herself does not directly carry out the operations, further issues can arise that require supervision and checking.

No consensus has been found on sociolinguistic annotations because of the controversial nature of sociological concepts. The use of sociolinguistic concepts to either criticize or condone the nature of a particular society (Bourdieu 1994) does not allow agreement to be established beyond statistical classifications by age, gender,

income or educational level. A typical example of one such disagreement lies in the different definitions of professions and socio-economic categories across countries, making it difficult to adopt a unified classification in the metadata.

Moreover, social judgments violate the requirement for objectivity. Therefore, assigning a form such as *avoir le seum* to *urban youth* goes beyond a characterization by age or place of residence. Since, in corpus instrumentation, an attribute is assigned a unique descriptor, a convention that defines social attributes in a consensual manner cannot be produced, as it will only reproduce bureaucratic classifications or validate a unilateral vision of the social world.

8 Types of annotation and levels of analysis

Three elements influence the type of annotation: (i) the type of software used; (ii) the data input; and (iii) the conceptual framework which, depending on the type of annotation that has been chosen, can determine the quality of the search results.

In linguistics, the type of query depends on the level of analysis chosen (Benveniste 1966) and the initial form of the data, e.g. text, speech, images or multimodal data. In this section, we first provide an example for written data, and then focus on the annotation of different units.

8.1 Written forms

Work on manuscripts drew attention to the importance of orthographic forms and variants, abbreviations, ligatures and omissions. Initially, the aim was to produce a clean text without defects or interpolations, and to trace the different versions of an original state, whether known or unknown, that scholars tried to reconstruct beyond its various incarnations. For economic and social reasons, such as the prevalence of a particular religious stance or possible links with the Greco-Latin tradition, textual transmission in the Middle Ages was limited to reproducing a small number of works with heterogeneous written practices (copyist workshops).

Written forms changed through the expansion of literacy and the decrease in cost of paper, as well as a change in the way information was exchanged, e.g. through the arrival of postal services, and general changes in behaviour. Studies moved away from codices towards corpora of texts written by less literate writers such as the 1789 register of grievances (Branca-Rosoff/Schneider 1994) or letters from infantrymen during the First World War (*poilus*, Steuckardt 2015). Preparing these sources required reconstructing the text in line with a standardized spelling and in some cases in line with Standard French.

8.2 Phonetics, prosody, phonology

IPA notation is often used on a small scale. Few large-scale corpora use IPA for all the transcriptions. The level of technicality required for the operation takes time, each choice can be debated (e.g. whether *meuf* has been produced with /œ/ or with /ø/), and the result is not as legible as a file using standard spelling. Whilst in the past semi-conventional notations were used, e.g. *ch'crois* for *je crois*, today priority is given to normalized transcriptions that are aligned with the signal, and where one is certain of being able to retrieve all the forms of a word. In the case of *ch'crois* for *je crois*, although the lemma *je* has only one variation, i.e. *j'*, a third option for *ch'* should be added to avoid empty results to search queries for first person subject pronouns.

Phonetic notation, whether relating to pronunciation, e.g. the realization of a schwa, or to prosody, e.g. in the Rhapsodie project, is included in the annotation whenever it is not predictable. The whole of the Perseval corpus (Gómez Molina et al. 2007) is segmented into prosodic groups and other examples include C-Oral Rom and Prieto/Roseano (2010).

Phonology is a mental competence of speakers, unlike phonetics, which is directly accessible from the signal and objectively measurable through instrumentation, and unlike morphology, which remains on a metalinguistic level. Most uses of phonology therefore primarily concern surface processing linked to morphology, either by the realization of a phonetic marker (e.g. liaison in the *Phonologie du français contemporain* (PFC) or the placement of the tonic accent), or by collecting different pronunciations corresponding to the underlying form of a lexical unit (Bergounioux 2016).

Annotation carried out directly on a transcription starts from the conventions that were used to produce the transcription, for instance in (i) the use or otherwise of punctuation marks, capital letters, italics; (ii) indicating vocal sounds, e.g. laughter, coughs; and (iii) noting pauses, disfluencies, etc.

Some examples of the annotation of phonetic or phonological phenomena in sociolinguistics include: (i) identifying consonantal cluster reduction phenomena in varieties of Italian (Vallone/Caniparoli/Savy 2002); (ii) prosodic characterization of discourse genres in Italian (Giordano/Savy 2003) and in French (Beliao/Lacheret/Kahane 2014); (iii) analysis of diatopic variation in European and South-American Spanish prosody (Prieto/Roseano 2010); and (iv) analysis of diastratic and diatopic variation in liaison in French (Durand et al. 2011).

8.3 Morphology, lexicon

In NLP, the main object of analysis is the word. One of the first concrete applications of computer science to natural language was in quantitative linguistics, building on intuitions in Zipf (1935) to produce lexical statistics (frequency lists):

- TLF (Imbs/Quemada 1971–1994; Guiraud 1954),
- LIP (De Mauro et al. 1993),
- CoLFis (Bertinetto et al. 2005),
- NVDB (Chiari/De Mauro 2014),
- CLUL,
- Frecuencias del español CREA (Almela et al. 2005).

As the resources were developed, morphological and lexical annotations were enriched, particularly in the automatic construction of the lexicon and concordancers. One of the first instances of data exploitation was the creation of lemmatized lists organized by decreasing order of frequency and used for research in lexical statistics, to develop learning resources and to produce dictionaries for use in NLP as well as in sociolinguistic analyses.

This type of access to corpora raises questions concerning both the value of units that are considered relevant and the elements of analysis. The status of the word as a scientific concept remains problematic. It is defined primarily in terms of writing, and therefore not very compatible with the idea of language as a verbal stream without fixed divisions. The scientific literature mentions: (i) floating morphemes, e.g. the prefix in *repolir* or *non-agréer*, or the suffix in *ordinatouille* (derived from *ordinateur*; attested on the internet); (ii) chunking into fixed phrases such as *pour autant que* and *condition nécessaire et suffisante*, and into expressions and into proverbs; and (iii) grammaticalization phenomena, for instance where *je sais pas* is equivalent to *à peu près* in *y avait je sais pas moi sept ou huit personnes*.

Moreover, adapting to NLP the linguistic categories based on texts written in an alphabet within a logical tradition rooted in western Indo-European languages limits the extent to which annotations can be generalized and are relevant.

An example of a task frequently performed in data mining and in automatic documentation for the purposes of constructing ontologies and contributing to the Semantic Web is the recognition of named entities, on the border between the lexicon and syntax. This application is particularly important in sociolinguistics where certain names make identification more difficult by providing indications about the relation between the speaker and the content of his/her discourse, e.g. *le Président des riches, ma fifille à moi*.

Annotation of morphological and lexical phenomena allowed sociolinguistic analyses to be carried out that led to: (i) relativizing prejudice about the quantity of dialectal and regional forms in spoken Italian, which are much less frequent than previously thought (De Mauro et al. 1993); (ii) measuring the extent of grammatical variation in popular Spanish (Fernández-Ordóñez 2011); (iii) exploring lexical code-switching phenomena between Spanish and Catalan in Spanish Catalonia (Martínez Díaz 2009).

Other recent examples of morphological and lexical phenomena annotated from a sociolinguistic point of view are: (i) the distribution of colloquial words, of informal

forms like *tu* and of polite forms of address in contemporary French (Beeching 2012); and (ii) lexical composition with diatopic and diaphasic differentiation in contemporary Italian (Chiari/De Mauro 2014).

8.4 Syntax

Determining Part-of-Speech tags taking morphological relations into account makes a syntactic annotation of the propositions following a linear representation possible, a process similar to Hockett's bracketing (Hockett 1954).

When the data include units larger than the word (for instance speech turns), the segmentation is primarily based on projections from syntactic (and semantic) analysis. Treebanks are currently the most common means of representation.

Some examples of Treebanks:

Ancient Greek/Latin	Ancient Greek and Latin Dependency Treebank (AGLDT)
Catalan	Cat3LB
Spanish	Cast3LB
French	French Treebank, Rhapsodie
Italian	ISST (Italian Syntactic-Semantic Treebank)
Portuguese	Projecto Floresta Sintá(c)tica
Romanian	RDT (Romanian Dependency Treebank)

These resources and their websites can be found on the internet, e.g. on the ELRA catalogue: http://catalog.elra.info/index.php?language=fr. In addition to these syntactic treebanks, there is also the C-Oral Rom corpus (Cresti/Moneglia 2005) and the Rhapsodie corpus (Lacheret/Kahane/Pietrandrea, forthcoming) that include annotation of the macro-syntactic oral structures of French, Italian, Spanish and Portuguese (C-Oral Rom) and of French (Rhapsodie).

While the word can be considered as the basic unit for computer scientists, linguists tend to prioritize parsers. In addition to the macro-syntactic annotation in C-Oral Rom and Rhapsodie, already mentioned, examples of linguistic and sociolinguistic uses of syntactic annotations can be found in the following work: (i) syntactic variation in the dialects of Romance languages (Dagnac/Sauzet/Sportiche 2015); (ii) analysis of diaphasic variation of dependency structures in spoken French (Pietrandrea, forthcoming; Kahane/Gerdes/Fleury, forthcoming); (iii) study of diaphasic variation in macro-syntactic structures in French, English, Spanish and Portuguese (Cresti/Moneglia 2005; Pietrandrea, forthcoming); and (iv) analysis of diastratic variation in certain syntactic structures, e.g. the *déqueismo* in Valencian Spanish (Gómez Molina/Gómez Devís 1995).

8.5 Semantics

Beyond what is *de facto* established, deliberately or not, by the transcription (the choice between *en dix ans tout ça va mieux* vs *en disant tout ça va mieux*), the first semantic intervention undertaken by annotations concerns the disambiguation, using POS, of homographic terms within the same category. Is *vers*, for instance, a preposition or a noun, and if a noun, does it refer to a metric unit or is it the plural of the lemma *ver*? Part of this task is made easier by indications, in the metadata, of the domain of specialization concerned or by a link established between the data and a terminology dictionary.

Semantic annotations are also used for: (i) analysis of thematic roles, and verbal and nominal classes; (ii) organization of the temporal dimension; (iii) processing of modality or metaphors; (iv) study of argument structure; and (v) exploitation of information structure.

Examples for modality include MODAL (Pietrandrea, forthcoming) in Italian and French, Avila (2015) for Portuguese, and the SenSem Corpus for Spanish and Catalan (Fernández-Montraveta/Vázquez 2014). Examples of semantic annotation in sociolinguistics include the analysis of methods of constructing shared epistemological knowledge in spontaneous conversations (Pietrandrea, forthcoming).

8.6 Discourse

At the level of discourse, a hierarchical reorganization is required when it comes to annotating coreferential elements and associative anaphors. The discontinuity of the sound sequence that characterizes these elements and the indicators that allow them to be retrieved are problematic for automatic annotation. Adjacent units can no longer be grouped together, meaning that other conventions are required, for example coreference in the ANCOR-Centre (Schang et al. 2011).

Whether the focus is on tagging speech turns or establishing a typology of language acts, annotation comes into play to characterize: (i) forms of dialogue and conversation (CID – Corpus of Interactional Data for French, PraTid for Italian); (ii) information structure (IPIC – Information Structure Database for Italian and Brazilian Portuguese); (iii) discourse relations (Annodis, French) and discourse markers (for an application to the Valibel, Clapi and Corpage spoken French corpora, see Bolly et al. 2015).

The possibility of transposing annotation formats from one Romance language to another raises the question of an intermediary stage between the Indo-European languages and languages that operate differently on the morphophonological level, (e.g. agglutinative languages, tonal languages), or use other ways of expressing semantic relations (e.g. languages with derivational classes or vowel reduction); and within Indo-European languages, between the languages of the western group. With-

in the stratum composed of Romance languages, certain differences can be observed, for instance the use of the neuter gender in Romanian or conjugation tables.

9 Criticism and perspectives

Sociolinguistic annotation suffers from an absence of consensus – inherent to the discipline and its critical function – with regard to the categories to be used as tags. The question arises whether these categories are necessary, i.e. what sort of query results do they provide?, and whether they are relevant, i.e. what contribution do they make to linguistic description? Generic information, e.g. age, gender or profession, is in general placed in the metadata so that annotation is required only when it comes to a potentially repetitive phenomenon or variation between occurrences that may be due to differential *distinctive* uses.

9.1 Sociolinguistic annotations

The observation of phenomena relevant to this category focuses on: (i) lexical units (professional terminology, youth speech or archaic forms, slang and language games such as *verlan*); (ii) certain syntactic turns of phrase, in particular weaker forms in the system that are used in a distinctive way (see Blanche-Benveniste/Martin 2010), e.g. relative constructions; (iii) collective representations (and self-representations) of agents and of their environment.

In contrast, few sociolinguistic annotations focus on phonology and prosody, except those arising through other types of considerations as in the case of liaison (Encrevé 1988), where morphology and social variation come together. Regardless, sociolinguistic annotation provides additional information to grammatical annotation. The alignment of the transcription with the signal, designed to enable a more reliable analysis, has allowed the homogeneity created by standard spelling to be overcome, but the reader cannot help but add his or her judgments to the voices and speakers.

9.2 Effect of discipline-specific fields and annotating variation

Amongst the obstacles to a unified approach are expectations regarding applications and the structure of research communities. The initial purpose of NLP was not to process language-internal variation but rather variation between languages based on constants, e.g. lexical count, phrase structure. Even when the wide variety of forms made it clear that there were great differences in production, notably in voice recognition, these differences were attributed mainly to inter-individual variation or, to a

lesser extent, to geographical origins (Boula de Mareüil/Woehrling/Adda-Decker 2013). The academic training of computer scientists did not prepare them for collaborating with sociologists and vice versa. The result is that very few studies were carried out in this area until the availability of large quantities of data in oral corpora made it necessary to take variation into account.

The instability of resources, of theories and also of transcription practices means that the variation found in the data tends also to be found, transposed, in the annotations. The result is a series of differences that are transmitted to the processing stage and give rise to difficulties in use and interoperability, i.e. to competing sets of solutions. It is therefore necessary to define evaluation criteria.

At another level, sociolinguistic annotations can involve the types of difficulties encountered. These annotations can be made up of comments added by the transcribers, or, more often, of disagreements between annotators. In turn, these differences can be used as indications of hearer competence.

Example 11: Hearer variation (speaker ESLO1/109)

Transcription A
euh en euh j'ai **dû la regagner oui** cette année c'est impressionnant les progrès quand même
Transcription B
euh en euh j'ai **vu l'an dernier puis** cette année c'est impressionnant les progrès quand même
Transcription C
euh en euh j'ai vu l'an dernier **j'ai vu** cette année c'est impressionnant les progrès **qu'on a faits**

Whilst the earliest works in NLP saw variation more as a difficulty than a surplus of information, certain areas of linguistics, where data are based on comparison, were eagerly awaiting solutions on this issue. This is the case for studies in language acquisition (see CHAT, CHILDES), learner corpora or clinical linguistics. Other problems concern multimodal corpora, for instance in the annotation of video corpora that are indispensable for the study of sign languages.

10 Conclusion

Annotation is the result of transformation in the practices of linguists when faced with the exponential increase in available resources through developments in computer science. To exploit these resources, it was necessary to master tools that, without being specific to linguistics (most of these tools are shared throughout digital humanities), nonetheless needed to be adapted, both when it came to data

collection, especially for oral data, and when it came to processing and storing the data.

A consensus in methods and a certain degree of standardization was achieved on the periphery of the field. For instance, all corpora adopt a process approach, use the same means of inserting elements of analysis into a file and share tasks between transcription, annotation and metadata. Differences between theoretical schools, the nature of the data and the size of the units (from the phoneme to the discourse) require a certain level of exchange. The flexibility of the corpus is limited by the orientation of the work, i.e. whether the approach is primarily linguistic or computational, and by differences in approaches between subdisciplines.

Beyond these distinctions are those that result from the academic context in which the corpus is produced and on its linguistic specificities. The distance between oral and written practices, between the standard language and dialects, but also between dominant forms (diglossia), or competition between overseas varieties, have repercussions on the way that annotation is conceived and used. These consequences are more apparent in oral data that are, by nature, less homogeneous than written data.

11 Bibliography

11.1 Electronic corpora (selected; all last accessed 18.02.2018)

Ancient Greek and Latin Dependency Treebank: http://www.dh.uni-leipzig.de/wo/projects/ancient-greek-and-latin-dependency-treebank-2-0
British National Corpus (BNC): http://www.natcorp.ox.ac.uk
Brown Corpus (English): http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM
CORIS/CODIS (Italian): http://corpora.dslo.unibo.it/coris_ita.html
Corpus ANCOR-Centre (French): http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html
Corpora for Spoken Languages: https://benjamins.com/#catalog/books/scl.15/main
Corpus de Català Contemporani: http://www.ub.edu/cccub
Corpus de Langue Parlée en Interaction (CLAPI): http://clapi.ish-lyon.cnrs.fr
Corpus de Referencia del Español Actual (CREA): http://corpus.rae.es/creanet.html
Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS): http://linguistica.sns.it/CoLFIS/Home.htm
Corpus Sensem (Spanish): http://grial.uab.es/sensem/corpus
Enquête Sociolinguistique à Orléans (ESLO): http://eslo.huma-num.fr
European Language Resources Association: http://www.elra.info/en
Frantext: http://www.frantext.fr
French Treebank: http://ftb.linguist.univ-paris-diderot.fr
frTenTen Corpus (French): https://www.sketchengine.co.uk/frtenten-french-corpus
IPIC (Information Structure Database for Italian): https://benjamins.com/#catalog/books/scl.61.05pan/details
Italian Syntactic-Semantic Treebank (ISST): http://catalog.elra.info/product_info.php?products_id=887
Lessico dell'Italiano Parlato (LIP): http://badip.uni-graz.at/it

London-Lund Corpus of Spoken English: http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM
Modèles de l'Annotation de la Modalité à l'Oral: http://modal.msh-vdl.fr/index.php/le-corpus
Nuovo vocabolario di base della lingua italiana (NVDB): https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana
Phonologie du Français Contemporain (PFC): http://www.projet-pfc.net
Projecto Floresta Sintá(c)tica: http://www.linguateca.pt/floresta
Reference Corpus of Contemporary Portuguese: http://www.clul.ulisboa.pt/en/10-research/713-crpc-reference-corpus-of-contemporary-portuguese
Rhapsodie (French): http://www.projet-rhapsodie.fr
Syntactic Microvariations of the Romance Languages of France (SyMiLa): http://blogs.univ-tlse2.fr/symila
Thesaurus Occitan (Thesoc): http://thesaurus.unice.fr
Treebank 3LB (Catalan, Spanish): http://www.dlsi.ua.es/projectes/3lb
Universal Dependencies Treebank Romanian (UDTR): https://github.com/UniversalDependencies/UD_Romanian
Valibel (French – Belgium): https://benjamins.com/#catalog/books/scl.61.05pan/details
Vienna-Oxford International Corpus of English (VOICE): http://ota.ox.ac.uk/desc/2542
Wortschatz/French (Leipzig): http://corpora.uni-leipzig.de/de?corpusId=fra_mixed_2012

11.2 Printed sources

Almela, Ramón, et al. (edd.) (2005), *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*, Madrid, Universitas.
Avila, Luciana Beatriz (2015), *MASS – Modal Annotation in Spontaneous Speech: Semantic Annotation Scheme for Modality in a Spontaneous Speech Brazilian Portuguese Corpus*, Veredas 19(2), 1–13.
Beeching, Kate (2012), *Sociolinguistic Aspects of Lexical Variation in French*, in: Tim Pooley/Dominique Lagorgette (edd.), *On Linguistic Change in French: Socio-Historical Approaches (Le Changement linguistique en français),* Chambéry, Presses Universitaires de Savoie, 37–54.
Beliao, Julie/Lacheret, Anne/Kahane, Sylvain (2014), *Discourse and Prosody in Spoken French: Why, What and How Should One Count?*, https://halshs.archives-ouvertes.fr/halshs-01066796 (last access 23.06.2017).
Benveniste, Émile (1966), *Problèmes de linguistique générale*, Paris, Gallimard.
Bergounioux, Gabriel (ed.) (2016), *Linguistique de corpus. Une étude de cas*, Paris, Champion.
Bertinetto, Pier Marco, et al. (2005), *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*, http://linguistica.sns.it/CoLFIS/Home.htm (last access 23.06.2017).
Blanche-Benveniste, Claire/Martin, Philippe (2010), *Le Français. Usages de la langue parlée*, Leuven/Paris, Peeters.
Boas, Franz (1911), *Handbook of American Indian Languages. Part 1*, Washington, Government Printing Office.
Bolly, Catherine T., et al. (edd.) (2015), *MDMA. Un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en contexte,* Discours 16, http://discours.revues.org/9009?lang=en (last access 23.06.2017).
Boula de Mareüil, Philippe/Woehrling, Cécile/Adda-Decker, Martine (2013), *Contribution of Automatic Speech Processing to the Study of Northern/Southern French*, Language Sciences 39, 75–82.
Bourdieu, Pierre (1994), *Stratégie de reproduction et mode de domination*, Actes de la recherche en sciences sociales 105(1), 3–12.

Branca-Rosoff, Sonia/Schneider, Nathalie (1994), *L'Écriture des citoyens. Une analyse linguistique de l'écriture des peu-lettrés pendant la période révolutionnaire*, Paris, Klincksieck.
Chiari, Isabella/De Mauro, Tullio (2014), *Nuovo vocabolario di base della lingua italiana*, Milano/Roma, Sapienza Mondadori Education.
Cresti, Emanuela/Moneglia, Massimo (edd.) (2005), *C-ORAL ROM. Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphia, Benjamins.
Crystal, David/Quirk, Randolph (1964), *Systems of Prosodic and Paralinguistic Features in English*, The Hague, Mouton.
Dagnac, Anne/Sauzet, Patric/Sportiche, Dominique (2015), *Microvariation syntaxique dans les langues romanes de France. Projet SyMiLa, résultats, prospectives*, paper presented at the SyMiLa workshop "La Microvariation syntaxique dans les langues romanes de France", Toulouse, 11–12 June.
De Mauro, Tullio, et al. (edd.) (1993), *Lessico di frequenza dell'italiano parlato,* Milano, Etas Libri.
Durand, Jacques, et al. (2011), *Que savons-nous de la liaison aujourd'hui?*, Langue française 169, 103–135.
Encrevé, Pierre (1988), *La Liaison avec et sans enchaînement*, Paris, Seuil.
Eshkol-Taravella, Iris, et al. (2011), *Un grand corpus oral "disponible": Le corpus d'Orléans 1968–2012*, Traitement automatique des langues 3(2), 17–46.
Fernández-Montraveta, Ana/Vázquez, Gloria (2014), *The SenSem Corpus: An Annotated Corpus for Spanish and Catalan with Information about Aspectuality, Modality, Polarity and Factuality*, Corpus Linguistics and Linguistic Theory 10(2), 273–288.
Fernández-Ordóñez, Inés (2011), *Nuevos horizontes en el estudio de la variación gramatical del español: el Corpus Oral y sonoro del Español rural*, in: Germà Colón i Domènech/Lluís Gimeno Betí (edd.), *Noves tendències en la dialectologia contemporània*, Castelló de la Plana, Universitat Jaume I, 173–203.
Fort, Karën (2012), *Les Ressources annotées, un enjeu pour l'analyse de contenu: Vers une méthodologie de l'annotation manuelle de corpus*, doctoral thesis, Paris, Université Paris 13, available at https://tel.archives-ouvertes.fr/tel-00797760v2 [last accessed 23.06.17].
Giordano, Rosa/Savy, Renata (2003), *The Intonation of "Instruct" and "Explain" in Neapolitan Italian*, International Congress of Phonetic Sciences 15, 603–606.
Gómez Molina, José Ramón/Gómez Devís, M. Begoña (1995), *Dequeísmo y queísmo en el español hablado de Valencia: Factores lingüísticos y sociales*, Anuario de lingüística hispánica 11, 193–220.
Gómez Molina, José Ramón (2007), *El español hablado de Valencia. Materiales para su estudio. Nivel sociocultural bajo*, Valencia, Universitat de València.
Gougenheim, Georges, et al. (edd.) ([2]1964, [1]1956), *L'Élaboration du français fondamental: Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.
Greene, Barbara B./Rubin, Gerald M. (1971), *Automatic Grammatical Tagging of English*, Providence, Brown University Press.
Guiraud, Pierre (1954), *Les Caractères statistiques du vocabulaire*, Paris, PUF.
Habert, Benoît (2005), *Portrait de linguiste(s) à l'instrument*, http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html (last access 23.06.2017).
Hockett, Charles F. (1954), *Two Models of Grammatical Description*, Word 10, 210–234.
Imbs, Paul/Quemada, Bernard (1971–1994), *Trésor de la langue française. Dictionnaire de la langue du XIX[e] et du XX[e] siècle (1789–1960)*, Paris, Gallimard/CNRS.
Kahane, Sylvain/Gerdes, Kim/Fleury, Serge (forthcoming), *Statistical Analyses of Spoken French Syntactic Structures*, in: Anne Lacheret/Sylvain Kahane/Paola Pietrandrea (edd.), *Rhapsodie: A Prosodic Syntactic Treebank of Spoken French,* Amsterdam/Philadelphia, Benjamins.
Kučera, Henry/Nelson, Francis W./Carroll, John B. (1967), *Computational Analysis of Present-Day American English*, Providence, Brown University Press.

Lacheret, Anne/Kahane, Sylvain/Pietrandrea, Paola (edd.) (forthcoming), *Rhapsodie: A Prosodic Syntactic Treebank of Spoken French,* Amsterdam/Philadelphia, Benjamins.

Leech, Geoffrey N. (2005), *Adding Linguistic Annotation*, in: Martin Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford, Oxbrow Books, 17–29.

Léon, Jacqueline (2015), *Histoire de l'automatisation des sciences du langage*, Lyon, ENS Éditions.

Martínez Díaz, Eva (2009), *Las motivaciones del cambio de código: Del español a la lengua catalana,* Revista electrónica de estudios filológicos 18.

Ogden, Charles Kay (1930), *Basic English: A General Introduction with Rules and Grammar,* London, Paul & Co.

Pietrandrea, Paola (forthcoming), *Epistemicity at Work. A Corpus Study on Italian Dialogues,* Journal of Pragmatics.

Pietrandrea, Paola/Delsart, Aline (forthcoming), *Macrosyntax at Work*, in: Anne Lacharet/Sylvain Kahane/Paola Pietrandrea (edd.), *Rhapsodie: A Prosodic Syntactic Treebank of Spoken French,* Amsterdam/Philadelphia, Benjamins.

Prieto, Pilar/Roseano, Paolo (edd.) (2010), *Transcription of Intonation of the Spanish Language*, München, Lincom.

Schang, Emmanuel, et al. (2011), *Coreference and Anaphoric Annotations for Spontaneous Speech Corpora in French*, in: Iris Hendrickx et al. (edd.), *Proceedings of DAARC 2011. The 8th Discourse Anaphora and Anaphora Resolution Colloquium*, Lisboa, Colibri Edições, 182–190.

Steuckardt, Agnès (ed.) (2015), *Entre village et tranchées. L'Écriture de poilus ordinaires*, Uzès, Inclinaison.

Stührenberg, Maik (2012), *The TEI and Current Standards for Structuring Linguistic Data*, Journal of the Text Encoding Initiative 3, https://jtei.revues.org/523 (last access 13.02.2018).

Vallone, Marianna/Caniparoli, Valentina/Savy, Renata (2002), *Una prima indagine su fenomeni di riduzione consonantica nella varietà napoletana nel corpus AVIP*, in: Agostino Regnicoli (ed.), *La fonetica acustica come strumento di analisi della variazione linguistica in Italia, Atti delle XII giornate di studio del Gruppo di Fonetica Sperimentale (GFS)*, Macerata, Calamo, 83–90.

Zipf, George (1935), *The Psychobiology of Language: An Introduction to Dynamic Philology*, Cambridge, MA, M.I.T. Press.

Damien Mooney

2 Quantitative approaches for modelling variation and change: a case study of sociophonetic data from Occitan

Abstract: This chapter presents and evaluates a variety of different statistical modelling techniques that have been used in variationist sociolinguistics to determine the linguistic and social factors that condition language variation and change, with the aim of operationalizing the central theoretical construct of the "variable rule". Both continuous and categorical sociophonetic data from Occitan are analysed: formant measurements for the mid-vowels, and rhotic consonants. Beginning with a traditional VARBRUL analysis, the chapter presents a series of increasingly complex statistical models for the Occitan variables, illustrating the evolution of statistical practice in sociolinguistics over the past 30 years. The analyses presented highlight the primacy of mixed-effect (regression) models in the field, as the results of these analyses can be more reliably generalized to the larger population from which speakers have been sampled.

Keywords: sociophonetics, Occitan, regression, variable rule, statistical analysis

1 Introduction

The "variable rule" has been a central theoretical construct in variationist sociolinguistics since Labov (1969) first introduced it in his analysis of African-American Vernacular English copula contraction and deletion (see also ↗6 Speaker variables in Romance; ↗8 Variation and grammaticalization in Romance). This construct has as its basis the notion of "orderly heterogeneity" (Weinreich/Labov/Herzog 1968, 100), or the postulate that language variation and language change are constrained by a combination of (potentially interacting) social and linguistic factors. Variable rules are "abstract optional rules" which form an integral part of a language variety's structural description (Cedergren/Sankoff 1974, 333–334). The extent to which variable rules reflect actual linguistic competence at the level of the individual and of the "speech community" has been a matter of some theoretical debate (see, for example, Sankoff/Labov 1979); from the linguist's perspective, however, the variable rule can be considered as "the probabilistic modelling and statistical treatment of discrete choices and their conditioning" (Sankoff 1988, 984). This chapter presents and evaluates different techniques, in the variationist sociolinguist's toolkit, that can be used to undertake this statistical treatment, illustrating these quantitative methods with sociophonetic data from Occitan.

https://doi.org/10.1515/9783110365955-003

During the 1970s, variable rule analysis was further developed in studies of language variation and change and this included the development of the "variable rule program" (Cedergren/Sankoff 1974; Rousseau/Sankoff 1978), a statistical modelling package for sociolinguistic analysis which provided a means of estimating the parameters of variable rules (Johnson 2009, 359). The variable rule program was created as a response to the fact that existing statistical modelling techniques, such as "analysis of variance" ANOVA, were largely unsuitable for analysing spontaneous speech data, which are notoriously unbalanced in their distribution (Tagliamonte 2006, 130). The variable rule program has existed under many guises since its initial development (Tagliamonte/Baayen 2012, 136): *Varbrul* (Cedergren/Sankoff 1974); *Goldvarb* 2.0 (Rand/Sankoff 1990); *Goldvarb X* (Sankoff 2005); *Goldvarb Lion* (Sankoff/Tagliamonte/Smith 2005). These packages, often collectively referred to as VARBRUL, have allowed the sociolinguist to model statistically the distribution of two discrete linguistic variants, as well as the (collective) effect of social and linguistic factors that condition the variation observed. Tagliamonte/Baayen note, however, that the past 30 years have seen the development of more sophisticated statistical modelling techniques, which may be more appropriate for analysing language data (2012, 136). Packages such as *Rvarb* (Paolillo 2002), *Rbrul* (Johnson 2009) and *R* (R Development Core Team 2009) provide the analyst with the opportunity to implement a more advanced version of variable rule analysis; these approaches have the primary advantage of facilitating a higher level of generalizability to the wider population with respect to the results obtained.

Traditionally, structured variability in Romance varieties has received relatively little attention when compared with the large body of variationist sociolinguistic literature on variable rule analysis in varieties of English; studies of the Romance languages have tended to examine low-level phonetic transfer and change, rather than investigating the social and linguistic constraints that govern these developments. There are some variationist studies, however, which have presented applications of the (traditional) variable rule program to varieties of French[1] such as, for example, Ashby (1981) and Van Compernolle (2008) on negative particle deletion, Ashby (1982; 1988) on left- and right-dislocations, Ashby (1992) and Williams/van Compernolle (2009) on forms of address; Regan (1996) on second language acquisition, Moisset (2000) on variable liaison, and Temple (2000a; 2000b) on plosive devoicing. Variationist studies of Canadian varieties of French has been particularly progressive in using the variable rule program to analyse spontaneous speech data; for example, Paradis/Deshaies (1990) on stress alignment in Quebec, Poplack (1992) on the subjunctive, Nagy/Blondeau (1999) on double subject marking in Montreal, King/Nadasdi (2003) on future temporal reference in Acadia, Sankoff/Blondeau (2007) on rhotics in Montreal, King/

1 Similar approaches have been applied to other Romance languages such as Spanish (see, for example, Sessarego/Tejedo-Herrero 2016), and Catalan (see, for example, Simonet 2010; 2011).

Martineau/Mougeon (2011) on first-person plural pronouns, and Comeau/King/Butler (2012) on past-tense aspectual distinctions in Acadia. More recently, other researchers have taken advantage of the advanced modelling techniques offered by the R environment such as, for example, Roberts (2012) on future temporal reference in Martinique, Burnett/Tremblay/Blondeau (2015) on negative concord in Montreal, and Mooney (2016a; 2016b; 2016c) on dialect levelling in the phonological system of southwestern metropolitan French. To my knowledge, no studies of language variation and change have used the variable rule program on data from minority languages in the francophone context, the so-called *langues de France*, or regional languages. Some researchers have performed ANOVAs on France's regional languages, such as Villeneuve/Auger (2013) on subject-doubling and negative particle deletion in Picard, Sichel-Bazin/Buthke/Meisenburg (2012) on Occitan prosody, and Kennard/Lahiri (2015; 2017) on mutation in Breton; with the exception of Villeneuve/Auger (2013), the data presented in these studies were largely experimental and therefore more suited to ANOVA than studies of spontaneous speech (Tagliamonte 2006, 130).

Figure 1: Gallo-Romance languages (Mooney 2016b, 9)

Figure 2: Gallo-Romance dialects (Mooney 2016b, 9)

The statistical analyses presented in this chapter model linguistic variation and change in the consonantal and vocalic systems of a local variety of the Occitan language. The most significant division within Gallo-Romance is between the dialect area in the north, the *langue d'oïl*, and the dialect area in the south, known as the *langue d'oc* (see Figure 1). The modern *langue d'oc* area is commonly divided into six main dialectal areas (see Figure 2): *gascon* in the southwest, including the *Béarnais* and *Aranais* sub-dialects; central *Languedocien*; *Limousin* and *Auvergnat* in the north; *Provençal* in the southeast, including the *Nissart* sub-dialect; *Vivaro-Alpine* or *Alpine*

Provençal above the *Provençal* region. In the second half of the twentieth century, the establishment of the *Institut d'Études Occitanes* led to the use of the term "Occitan" to refer to all *langue d'oc* dialects, collectively considered to be a single language. Indeed the term "Occitan" has become a source of ideological conflict in southern France, especially for those who consider local varieties of the *langue d'oc* to be languages in their own right (see Blanchet/Schiffman 2004; Moreux 2004; Mooney 2015 for discussion); nonetheless, I will use the term "Occitan" here for simplicity. The data presented in this chapter come from the *Béarnais* sub-dialect of *Gascon*, spoken in the region of Béarn or the historically Romance-speaking part of the Pyrénées-Atlantiques *département* in southwestern France. Like all Occitan dialects, *Béarnais* has found itself in an increasing state of language obsolescence from the late nineteenth century onwards. In the entire *Gascon* region, the highest concentration of speakers exists in Béarn, making *Béarnais* the principal surviving sub-dialect. Moreux (2004) suggests that, at the beginning of the twentieth century, there were about 40,000 fluent native speakers of Occitan in Béarn, noting that the large majority of these speakers were over the age of 65 and rural dwellers.

This chapter begins by describing the Occitan data set collected in Béarn, and by outlining the linguistic variables to be modelled statistically – the dependent variables (section 2): (i) rhotic consonants; (ii) front mid-vowel contrast. The social and linguistic factors expected to condition variation and change in the Occitan phonological inventory are then presented – the independent variables (section 3). The body of the chapter presents a series of (increasingly complex) statistical modelling techniques for the Occitan linguistic variables under consideration, beginning with a traditional VARBRUL-style analysis of the Occitan rhotics (section 4.1), before discussing interactions between independent variables and some methodological issues involved in including correlated social and/or linguistic factors in the analyses proposed (section 4.2). The front-mid vowels are submitted to an *Rbrul* analysis in section 4.3, using a technique previously unavailable in the VARBRUL suite. The Occitan data are then submitted to a series of the most up-to-date statistical modelling techniques available for variationist data (section 4.4), with the final section discussing some proposed statistical techniques for resolving on-going issues encountered with current modelling methodologies.

2 Dependent variables: Occitan sociophonetic data

In variationist sociolinguistics, the "dependent variable" is the linguistic variable whose distribution we are interested in analysing statistically. Dependent variables in sociolinguistic studies are usually either binary or continuous: binary variables have two discrete variants and are categorical in nature; continuous variables are characterized by having a range of variants on a gradient scale (Hay 2011, 200). The Occitan data presented in this section are sociophonetic in nature, meaning that it was

collected using traditional Labovian sociolinguistic methods and that it was analysed using the acoustic phonetic techniques of laboratory phonetics. The data set contains examples of both binary and continuous dependent variables, the rhotic consonants and the front mid-vowels, respectively. The Occitan corpus contains high quality acoustic data, collected in 2012, for ten bilingual Occitan-French speakers, five male and five female, over the age of 65, and native to the region of Béarn. All informants participated in a wordlist translation task from French into Occitan and were recorded using a sampling rate of 44.1 kHz and a 16-bit PCM sample size on a Marantz PMD661 Solid State Sound Recorder. Subsequent acoustic analyses were performed in Praat version 5.2.21 (Boersma 2001; Boersma/Weenink 2012).

2.1 Categorical variables: Occitan rhotic consonants

There are very few comprehensive analyses of the distribution of rhotic consonants in Occitan varieties. Most commentators agree, however, that *Gascon* has two historically appropriate rhotic consonants, the voiced apical trill [r], and the voiced apical tap [ɾ], which are in contrastive distribution in intervocalic position, e.g. *poret* /puˈret/ ('chicken') ~ *porret* /puˈret/ ('leek') (Bec 1973; Cardaillac Kelly 1973), and not contrastive in other contexts, such that "an archiphoneme could be set up for all other positions" (Cardaillac Kelly 1973, 32). The apical rhotics are not, however, in strictly complementary distribution in non-intervocalic contexts: the distribution of [r] and [ɾ] is somewhat constrained by their position within the syllable and with respect to word boundaries with a tendency for [r] to occur word-initially and as an onset after [n], and [ɾ] to occur in onset clusters and in the syllable coda, but this distribution is by no means categorical (Cardaillac Kelly 1973, 32; Mooney 2014, 345).

Previous analyses of the *Gascon* rhotics have noted the transfer of dorsal rhotic consonants from French due to prolonged language contact. The phonological inventory of modern standard French contains one rhotic consonant phoneme, the voiced uvular fricative /ʁ/, which is often realized as a voiced uvular trill [ʀ] by older rural speakers. Cardaillac Kelly found that dorsal realizations [ʁ ʀ] occurred as variants of /r/ and /ɾ/ "in *all* positions as a consequence of bilingualism" (1973, 32), that when dorsal variants are used in intervocalic position, the phonemic distinction between /ɾ/ and /r/ is neutralized, e.g. *poret ~ porret* [puˈʁet], and that female speakers used more dorsal variants than male speakers.

The analysis of the Occitan rhotic consonants considered contact-induced change of the place of articulation of the categorical dependent variable, (R), with binary variants [apical] and [dorsal], representing the historically appropriate and transferred forms, respectively. 466 tokens of the (R) variable were categorized on the basis of an auditory or impressionistic analysis, which was supplemented by visual inspection of the acoustic spectrogram; an equal number of token counts was extracted for both male and female speakers.

2.2 Continuous variables: Occitan front mid-vowels

Traditionally, Occitan distinguishes between two mid-vowels, /e/ and /ɛ/, in the front of the vowel space (e.g., *peis* /peʃ/ 'fish', *pè* /pɛ/ 'foot'); these vowels are contrastive phonemes in *Gascon*, e.g., *qu'ei* /kej/ 'he is', *qu'ai* /kɛj/ 'I have'. There is some evidence to suggest that this phonemic distinction is not maintained in certain varieties of Occitan (Séguy 1954–1973); the analysis of the front mid-vowels aimed to determine the extent to which this contrast is maintained in Béarn. Figure 3 presents the full Occitan oral vowel system.

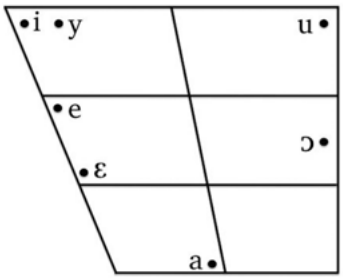


Figure 3: Occitan oral vowels (Mooney 2014, 346)

While the phonemic distinction between the front mid-vowels is theoretically categorical, /e/ or /ɛ/, their phonetic realizations can be analysed as a continuous variable (E) by measuring the first two formant frequencies, F1 and F2, of each vowel token in the corpus; these formant values are located on a gradient scale in the acoustic vowel space. Formant frequencies are commonly held, in acoustic phonetic studies of oral vowels, to have general non-linear articulatory correlates: F1 exhibits an inverse correlation with vowel height; F2 exhibits a positive correlation with vowel frontness/backness. The first and second formants were estimated in Praat using the LPC (Linear Predictive Coding) algorithm. The vowel onset and offset were first labelled in a Praat text grid and a script was used to automatically extract the value of F1 and F2 at the vowel midpoint; 253 tokens of the Occitan front mid-vowels were included in the analysis.

3 Independent variables: social and linguistic factors

Independent variables are generally linguistic or social factors that we expect to influence the distribution of the dependent variable. In traditional variationist studies, independent variables are referred to as "factor groups" and their variants are referred to as "factors". For example, an independent variable for speaker sex may be

included in the analysis as a "sex" factor group with variants [male] and [female]. These terms are not used outside the context of the variable rule program; in general statistical practice, factor groups are referred to as "predictors" and factors are referred to as "levels".

For the categorical (R) variable, the analyses included five independent variables or predictors, two social and three linguistic (see Table 1). Speaker sex and the speaker's place of origin were included in the analysis to determine the extent to which the distribution of apical and dorsal variants was influenced by a speaker's gender and/or regional origin. Three linguistic predictors were also included; previous studies of the *Gascon* rhotics have suggested that their distribution is partially constrained by syllable type and by phonological context. For the latter, a general distinction has been drawn between front and back consonants and between front and back vowels for the "preceding phoneme" and "following phoneme" predictors, the hypothesis being that adjacent anterior articulations will favour apical realizations and that posterior articulations will favour dorsal realizations. In the statistical analyses that follow, the models include either "syllable type" alone or "preceding phoneme" and "following phoneme" together (see section 4.2 for discussion).

Table 1: Independent variables included in statistical analyses of dependent variable (R)

Predictor	**Levels**
Speaker sex	Male
	Female
Place	Gan
	Nousty
	Nay
Syllable type	Simple onset
	Complex onset
	Simple coda
	Complex coda
Preceding phoneme	Front vowel
	Back vowel
	Apical consonant
	Dorsal consonant
	Non-lingual consonant
	Pause

Table 1: (continued)

Predictor	Levels
Following phoneme	Front vowel
	Back vowel
	Apical consonant
	Dorsal consonant
	Non-lingual consonant
	Pause

For the continuous (E) variable, the analyses included seven independent variables or predictors: two social and four linguistic. Again, speaker sex and place or origin were included as social predictors. "Phoneme" was included as a predictor to determine the extent to which the historically appropriate phoneme (/e/ or /ɛ/) could predict F1 and F2 values when phonological context had been taken into account. Syllable type was also included in as an independent variable as the distribution of the front mid-vowels is heavily influenced by open and closed syllabic contexts in the local variety of French spoken in the region (Mooney 2016b). In the F1 statistical analyses, F2 was included as a predictor to investigate potential significant correlations between the formant frequencies; F1 was equally included as a predictor in the statistical models containing F2 as a dependent variable.

Table 2: Independent variables included in statistical analyses of dependent variable (E)

Predictor	Levels
Speaker sex	Male
	Female
Place	Gan
	Nousty
	Nay
Phoneme	/e/
	/ɛ/
Syllable type	/Cv#/– Open final
	/CvC#/– Closed final
	/vCV(C)#/– Open medial
Preceding phoneme	Various

Table 2: (continued)

Predictor	Levels
Following phoneme	Various
F1 or F2	Continuous

4 Statistical modelling

Before undertaking a statistical analysis of the formant frequencies (F1 and F2) extracted from vocalic data, it was first necessary to normalize the data set. This is because different speakers exhibit variation in the formant values they produce for a given phonological vowel because of physiological differences in their vocal tracts. Normalization aims to eliminate variation which is caused by anatomical differences while preserving variation that is sociolinguistically significant (see Mooney 2016b for methods used).

Statistical modelling allows the sociolinguist to identify the various components of a variable rule using statistical inference. There are many statistical tests that can be used to examine the effect of an independent variable on the distribution of the variants of a dependent variable, such as a t-test or Spearman's correlation (Hay 2011, 206–207), but these are largely inappropriate for sociolinguistic or sociophonetic data sets. This is because it is not possible to use these tests to consider the effect of multiple (potentially interacting) predictors on a dependent variable, the essence of a variable rule. In order to obtain "an assessment of the significance of each candidate predictor over and above any variation that can be explained by the other potential predictors" (Hay 2011, 207), we must use a statistical modelling technique known as regression:

> "Regression analysis is a statistical tool for the investigation of relationships between variables. [...] To explore such issues, the investigator assembles the data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence" (Sykes 1993, 1).

ANOVA is a special case of regression that has been widely used in experimental linguistic studies, but this method is not suitable for spontaneous speech data as it assumes an even distribution of the data across the cells of the data set and this is almost never the case in sociolinguistic studies (Tagliamonte 2006, 137).

The regression models presented in this chapter have all been carried out in the R environment (version 3.2.3) using the *Rbrul* (version 2.3.2) text-based interface (Johnson 2009) which makes use of existing functions in the R environment. The remainder of this section presents a series of increasingly complex regression analyses for categorical and continuous dependent and independent variables. Where possible,

the regression models have been presented as they appear in the *Rbrul* interface to familiarize the reader with R output. All models distinguish the following levels for statistical significance: $p < .05$ and $p < .01$, for which the probability of observing the effect returned by chance is less than 5 % and 1 %, respectively; $p < .001$ is highly significant; for $p < .0001$, the probability of observing the result returned is considered to be approximately zero ($p \approx 0$), or 100 %.

4.1 Logistic regression models

The statistical modelling technique most widely used in sociolinguistics is logistic regression, a type of "generalized linear model" (Agresti 2007, 67): logistic regression examines the effect of multiple predictors on a binary categorical dependent variable. The variable rule program, or VARBRUL analysis, allows the analyst to model the effect of categorical predictors ("factor groups") on a categorical variable; the versions of VARBRUL that are currently available do not support continuous dependent or independent variables.[2] The Occitan rhotic consonant data set is modelled statistically in this section using (i) a traditional VARBRUL-style binomial stepwise regression analysis, and (ii) a simple main effects one-level binomial regression analysis in *Rbrul*.

The VARBRUL series of applications offers two options for data analysis (Tagliamonte 2006, 139): (i) binomial one-step and (ii) binomial step-up/step-down. The first option provides statistical information on all predictors included in the analysis, including those that are not determined to have a significant effect ($p < .05$) on the dependent variable. The second option, also known as stepwise regression, has been used most often in studies of language variation and change (Tagliamonte 2006, 140):

> "Stepping up, [the program] starts with no predictors and adds the most significant factor group, if there is one, before repeating the procedure. Stepping down, it starts with all possible predictors and removes the one that contributes least to the model, and then repeats this until all remaining predictors are significant" (Johnson 2009, 380).

When the stepwise procedure is complete, VARBRUL returns a logistic regression model that includes all "factor groups", or predictors, that "affect the response variable of interest, in what direction and to what degree" (Johnson 2009, 359).

2 *Varbrul 3* allows continuous variables to be included in the analysis but this version of VARBRUL analysis has not been made available on personal computers (Sankoff 2006, 1157; Johnson 2009, 360).

Table 3: VARBRUL-style stepwise regression analysis of Occitan rhotic consonants, with [dorsal] as application value (Log likelihood = –193.429, degrees of freedom = 12, significance = 0.000, input = 0.107)

Factor group	*Factors*	*Total N*	*% of total*	*Factor weight*
Place of origin	Nay	145	32	0.708
	Gan	231	22	0.615
	Nousty	90	6	0.205
Following phoneme	Pause	66	55	0.817
	Apical consonant	77	27	0.569
	Non-lingual consonant	43	26	0.510
	Front vowel	214	12	0.402
	Dorsal consonant	9	11	0.348
	Back vowel	57	12	0.313
Preceding phoneme	Pause	17	41	0.869
	Front vowel	215	29	0.658
	Back vowel	89	28	0.560
	Dorsal consonant	50	10	0.476
	Non-lingual consonant	61	3	0.211
	Apical consonant	34	3	0.203

Table 3 presents the results of a typical VARBRUL-style[3] stepwise logistic regression, using VARBRUL terminology for the categorical Occitan (R) dependent variable, with variants [dorsal] and [apical]. For binary categorical variables, one variant must be designated as the "application" or "response" value, or the "variant defined as the outcome of the variable rule" (Tagliamonte 2006, 263). The results returned by the variable rule program are relative to the application value. For example, the model presented in Table 3 included [dorsal] as the application value and so any significant effects shown to favour or disfavour the dependent variable are, in fact, shown to favour or disfavour dorsal variants. Four independent variables, or "factor groups" were

3 The VARBRUL-style analysis was actually implemented in *Rbrul*, operating with different settings. Johnson has shown that, operating in a simpler mode, "Rbrul provided nearly identical output to the actual GoldVarb program" (2009, 381).