HOW CAN SOCIOLINGUISTIC DATA BE USED?*

Celeste Rodrigues
celesterodrigues@campus.ul.pt**

Deolinda Simões
Deolinda.Reis.Simoes@at.gov.pt***

Our paper addresses the ethical and legal issues related to the display of the speech data included in sociolinguistic data bases, considering in detail the case of CPE-Var, a corpus of European Portuguese collected by the first author. Privacy, consent, data integrity, anonymisation, authorship, copyright and related topics are addressed. We provide general background of sociolinguistic data bases, discussion of the possible uses of the data, discussion of the consequences of misuse of the data, besides the relevant support to appropriate scientific use of CPE-Var data. Most ethical and legal questions arise when data are to be used outside the main scope of the initial research project. Even though the legal framework covers the first objective of the research, it is questionable if some other research proposals are also covered by the consent obtained. Our conclusion is that some further uses of CPE-Var data are legitimate, but others must be cautiously avoided or discarded.
Key words: ethic and legal rights, sociolinguistic databases, privacy, informed consent, authorship, copyright

O presente artigo trata os problemas éticos e legais relacionados com a divulgação de dados de fala incluídos em bases de dados sociolinguísticos, tomando para o efeito o caso do CPE-Var, um corpus recolhido pelo primeiro autor. Os tópicos discutidos incluem: privacidade, consentimento, integridade dos dados, anonimização, autoria e copyright, entre outros. É fornecido um background das recolhas de dados sociolinguísticos, uma discussão dos usos possíveis desses dados e das

---

* We thank M. E. Cabrita for a previous linguistic review of the paper.
** University of Lisbon, Arts Faculty (DLGR), Centro de Linguística da Universidade de Lisboa (CLUL), Lisbon, Portugal, under «PEst-OE/LIN/UI0214/2011» Project (Funded by National Fundings delivered by Fundação para a Ciência e a Tecnologia (FCT)).
*** Law and Forensic Sciences expert, Senior Officer of Customs and Tax Authority of the Finance Ministery, Lisbon, Portugal.

consequências dos possíveis usos indevidos, para além de argumentação de suporte dos usos científicos dos dados deste corpus. A maior parte das questões éticas surgem quando se pretende usar os dados fora do âmbito inicial da pesquisa. Embora os fins iniciais da pesquisa estejam legalmente cobertos, é discutível se outros usos serão legítimos. A nossa conclusão é a de que alguns novos usos podem legitimamente ser dados aos materiais, mas outros devem ser avaliados com muito cuidado ou rejeitados.
Palavras-chave: direitos éticos e legais, bases de dados sociolinguísticos, privacidade, consentimento informado, autoria, copyright.

*

1. Introduction

Among ethical problems involved in sociolinguistic research, there are: privacy, anonymisation, copyright and informed consent.

Privacy is one of the major ethical issues discussed here, since private information is mentioned in several interviews. The high spontaneity level of some interviews creates the illusion of confidentiality, allowing private matters to be addressed. Sociolinguistic interviews often share the following characteristics: they are intended to provide useful manageable linguistic data in a format that allows linguistic description, statistical treatment of the data and both social and linguistic profile creation. The researcher tries to get all the information he can to better understand the data afterwards, although he may not show an explicit interest in those particular details. Data, after transcription, are usually included in data bases. Data consist of sound recordings of single individual interviews in pre-established appointments in an environment familiar to the interviewee. CPE-Var interviews have formal discourse, reading tasks and semi-informal discourse. Since the interviews had to be orthographically transcribed by students, privacy of the information concerning the speakers was challenged. In addition, some data is now available in a database, which was created to treat quantitatively the phonetic variants of interviews. This database has personal data of the speakers, although anonymisation requirements are generally obeyed.

Apart from that, nowadays, voices are recognized with high probability in very short samples, if they are submitted to careful acoustic scrutiny. These interviews vary from 60 to 75m long. As a consequence, is it ethically fair to assembly even small parts of the interviews and to diffuse them in public speech databases? Can we use data to study the acoustic charac-

teristics of voice? Is it possible only to reproduce some samples in scientific environments? Which are the limits? Which are the laws that we must observe in this environment?

Speakers own their voices. The researcher who collected those voices is the owner of the interviews. Does the researcher have the right to freely disclose all the content of the interviews? Copyright matters of the researcher seem to interfere with speaker's right to their own voices.

Another ethical question arises when we think about consent. Nowadays, consent is always asked, but not in the 90's. CPE-Var doesn't have a written consent, even though the situation in the recordings was only possible with the voluntary consent of the speaker. What can a lawyer present for/against the use of this kind of data in public databases and in other domains?

Furthermore, how can we assure that the integrity of the data is kept if we allow that data are used by other people? Can we change the data in any way? Can we, for instance, copy silence intervals to the parts of the interaction where speaker proper name appears? Or is it considered manipulation? Which are the rights of the author, as the author of the recordings and of the database where transcriptions appear? Several of these issues are developed below.

Section 2 provides some general sociolinguistic background. Section 3 summarises methodological issues concerning data collection and interview format of CPE-Var. Section 4 mentions major limitations (both ethical and legal) to the use of data that emerged during the research. Section 5 further expands on the solutions implemented to avoid undesired ethical results. Section 6 focuses on the legal background of ethical issues. Finally, in Section 7, we present some concluding remarks and refer to issues that must be taken into consideration in sociolinguistic research.

2. Sociolinguistic background

Among Social Sciences, Sociolinguistics is a sub-field of Linguistics concerned with the correlation of society and language use. Sociolinguistics then combines technical knowledge from Linguistics and several other fields: Sociology, Psychology and behavioural sciences in general, etc. Following this reasoning, sociolinguistic work supposes a wide knowledge of the social tissue and of the behaviour of speakers according to different settings of language use, from the start.

Familiarity with Portugal's social reality and its culture as regards language use, on the one hand, and the linguistic knowledge which rendered such work possible, on the other hand, have led Rodrigues to build CPE-Var corpus (Corpus de Português Europeu – Variação), since such a collection of adequately stratified linguistic data was not available for European Portuguese (EP) in the 1990s. This collection of linguistic data includes native speakers from both major EP dialects and, consequently, may exemplify the behaviour of speakers of two dialects, one closer than the other to standard EP.

EP in mainland Portugal though is acknowledged to have small linguistic variation since speakers from all parts of the country understand one another. Several variation features, namely in the phonetic shape of words and in the prosodic curves that characterise single dialectal varieties can, however, be found. Variation in syntax and in the lexicon is also found, although it is generally considered to be smaller. Several variation phenomena are associated to sociocultural differentiation.

There are two main dialectal and geographical areas in the Portugal's mainland: center and south area is mainly flat and the north is more mountainous. Rodrigues chose two targets: Lisbon area, within the geographical area where linguists have already acknowledged to be found a language variety that generally avoids several linguistic features, known to identify non-standard varieties of EP (in the center and south area), and Braga, which belongs to the northern area where the language has several different (and sometimes stygmatised) linguistic features. The former is often considered to be more innovative and the latter more conservative, as a consequence of language formation and history (Castro 2006, Cintra 1971).

Lisbon, as the capital city, has a considerable part of the country's population (considering the territorial area covered) with sociocultural elite, nearly all governmental bodies, the most important media and road infrastructures. It is also the political and financial center of the whole country. As a consequence, educated speakers from Lisbon or from Coimbra (in center area) tend to use unmarked linguistic features, not only in writing, but also in spoken language (a certain kind of linguistic standard). Like any other dialect, Lisbon dialect has sociolinguistic variation. Only the linguistic features present in the spoken language of the educated speakers tend to spread all over the territory. Some of its features are now found in geographical areas where they were not expected (for instance, the centralization of stressed pre-palatal /e/ is nowadays found in the Alentejo). For all these reasons, the importance of the linguistic features of the dialect spoken in Lisbon is naturally high as such features may spread, even if the speakers

are unaware of this diffusion process. Adoption of standard language features by speakers of other dialects (standardisation) is particularly visible when it is quantitatively relevant. This often happens when these speakers interact with someone from outside their region (be it from the standard dialect or not), as it is the case of the CPE-Var.

Lisbon dialect is different from non-urban areas within the dialect of the center and south. Unlike the rest of the geographical area, it shows a tendency to close initial unstressed front vowels and a tendency towards centralisation in general (either in stressed or in unstressed position). These processes lead to neutralisation of several phonological vowels in certain highly productive contexts. Moreover, the Lisbon dialect has a high level of vowel weakening, culminating very often in vowel deletion. As far as the consonantal system is concerned, the Lisbon area shows an innovative variant related to sonorant /ɾ/. The phonological system also has two contextually motivated variants, a tap in Coda position (*carta, letter*) and in onset intervocalic position (*cara, expensive*) and a trill in onset (*rapaz, boy, carro, car*) after a coda segment (*melro, blackbird*) (see Mateus and Andrade 2000). The trill is realised most commonly in Lisbon as [R], unlike the alveolar pronunciation it shows in some speakers of other dialects (the segments are in free variation). [R] is considered to be an innovative feature of Lisbon speech that is now spreading to the other varieties of the language (it may be noted incidentally that this feature keeps its anterior nature in the nearby regions of the south).

The main goal of Rodrigues's PhD doctoral dissertation was to determine how the features of the standard dialect spread in the dialect used in Braga, as an exponent example of this northerner speech, which means that she had to collect a corpus of Braga speech comparable to the one collected in Lisbon.

Braga is an ancient town, highly rejuvenated by industry and the establishment of university facilities in the past three decades, among other factors. It represents here the north region where Portuguese developed and where some ancient linguistic features are still present today. Salient features of Braga speech include the absence of nasal vowel closing and the pronunciation of a bilabial stop or a fricative for the phonological fricative /v/ (*luvas, gloves*) and a small degree of vowel deletion phenomena. Nasal codas are still present in Braga dialect, unlike in the other dialects. Braga also exhibits a typical prosody, quite distinct from the Lisbon speech (Vigário e Frota (2003), among others). Variation can be found among speakers, according to the degree of attention to speech, the kind of interaction established and the gender and the sociocultural profile of the speakers.

3. Description of methodological issues concerning CPE-Var

CPE-Var interviews follow the general lines of sociolinguistic interviews, described by Labov (1981), as will be shown below. The investigation was initially meant to identify the phonological and phonetic features that were subject to variation in the cities of Lisbon and Braga. It involved, then, a sample of interviews of different speakers characterised by specific sociocultural profiles. Speakers were classified in four education levels and in five age groups. Even with all these differentiation factors, since no income information was collected, the sociocultural profiles were only approached. Several linguistic indices/markers and stereotypes of speaker voice were identified (see Rodrigues, 2003).

Recording quality had to be high to pursue such objectives: very small properties of the acoustic signal. Recordings had therefore to be done in a controlled environment with the speakers' previous agreement. These characteristics rend it possible for us to use the interviews for several other purposes nowadays. The interviews are linguistically rich enough to allow the study of several linguistic topics from distinct perspectives: phonetic, phonological, morphological, syntactic, reading and spontaneous discourse analysis, etc. That is why we have to deal with ethical issues in this new phase of data use.(1)

3.1. Data collection

CPE-Var includes 180 single sociolinguistic interviews among speakers of Lisbon and Braga collected from 1996 to the end of 1998. These interviews often include personal information raising a number of ethical and legal problems to the researcher. Sensitive topics present in some interviews and the use of voice materials for identification of acoustic parameters relevant to the identification of speaker's voices raise many ethical concerns: namely information privacy and confidentiality, transcription and recording anonymisation, author's rights, among the most important. Most of the problems may be expected from the beginning, others come around only in the course of the investigation or afterwards.

(1) For example, some orthographic transcriptions of CPE-Var data were useful for papers on EP syntax properties (Duarte, Inês, Maria João Freitas, Anabela Gonçalves, Matilde Miguel and Celeste Rodrigues (2002)) and a part of CPE-Var is currently being used to identify robust acoustic parameters of speaker's voices.

Sociolinguistic interviews often share the following characteristics: they are intended to provide useful manageable linguistic data in a format that allows linguistic description, statistical treatment of the data and both social and linguistic profile creation. Researchers try to get all the information they can to better understand the data afterwards, although they may not show an explicit interest in those particular details. Data, after transcription, are usually included in databases. Data consist of sound recordings of single individual interviews conducted in pre-established appointments in an environment familiar to the interviewee. CPE-Var interviews have formal discourse, reading tasks and semi-informal discourse.

Informal speech samples sometimes include personal information. It would be un-ethical to divulge this information (for instance, original explanations/ideas, appreciation on the moral conduct of public or otherwise recognisable figures, on the speaker's health, contents of an unexpected phone call, etc.) for it was not meant to be revealed though it was recorded. Some speakers avoid sensitive subjects, others do not. Moreover, these statements/opinions combined with personal information given at the beginning of the interaction and the quality of the voice recorded clearly identify the speaker, even though the researcher omits the speaker's name in the transcription or the database. Nowadays, accurate voice recognition occurs in very short samples if they are submitted to careful acoustic scrutiny. These interviews vary from 60 to 75m in length. Therefore, some questions may be raised: is it ethically fair to assembly even small parts of the interviews and to divulge them in public speech databases? Is it possible to reproduce only some samples in scientific environments? Where do we set the limits? Which laws must we comply with in this environment? Must all data collections have informed consent files signed beforehand? Most of them do not have such files, especially the first ones. CPE-Var does not either, even though the situation in the recordings was only possible with the voluntary consent of the speaker. On what grounds can a lawyer argue for or against the use of this kind of data in public databases and in other domains?

3.2. Speakers

Target speakers were born either in Lisbon or Braga and have lived in their cities most of their lives. They have been selected according to their social profile: they are either male or female, from five age groups and four education levels.

Some interviewees had some previous knowledge of the interviewer's work, others did not. Speakers were asked to collaborate in a research that would lead in the first place to the interviewer's PhD dissertation. Speakers were informed of the general research goal, that is, dialectal comparison. Speaker anonymity was assured. Speakers agreed to participate in all the tasks the interview comprised, including reading of a long word-list, reading of a list of sentences and a text. They also agreed to entertain a non-oriented dialogue on topics of their choice. No consent document was signed at the time (1996-1998) since speaker's agreement was regarded as tacit and at that time such document was not required. Their consent is obvious considering their willingness to be interviewed and the fact that several refer to the recording procedure in the course of the interview.

Most speakers relaxed after the reading tasks, allowing sometimes for the linguistic interaction to go on and on, and achieving a degree of attention to discourse next to the vernacular. Other speakers, though, always kept in mind that they were speaking to someone that did not belong to their linguistic variety (as the results of Rodrigues (2003) show). Even though speakers pay different degrees of attention to the way they talk in this last part of the interview, the spontaneous speech collected is of high standard quality. It is informal enough to describe the linguistic differences of the two dialects under analysis, namely regarding both phonological and phonetic properties of the language in its current style (in the above described situation). Linguistic variation phenomena were the main focus of research.

3.3. Interview

Interviews were all made by the same interviewer, keeping as far as possible the same linguistic situation. Interviews had no observers and were made in a quiet room; most recordings were made in an environment familiar to the interviewee.

A PMD Marantz portable recorder equipped with a unidirectional microphone placed in front of the interviewee (at 20 cm distance) was used to capture the recordings. Analogical recordings are now adequately preserved in digital format.

Interviews have the following structure: at the beginning, a part of formal discourse (where speakers identified themselves and their social profile – this part of the interview is the most formal of all parts involving sponta-

neous speech), three reading tasks (550 isolated words presented in separate cards, a list of sentences also presented in separate cards, and a one-page text from the weekly press), followed by an semi-informal dialogue where speakers were asked to relax since their test phase had been accomplished. This structure was used to test if a different order of the standard Labovian procedure would capture more casual spontaneous speech at the end. Labovian sociolinguistic interview tries to capture informal speech style at the beginning and proceeds to more formal styles. We anticipated people's relationship would develop all along the interview, leading to a lower degree of formality in the final part, if the interview is long enough. This assumption was verified, since results of quantitative analysis of variation phonological phenomena show an increase in the use of informal variants in this last part of the interview (Rodrigues, 2003).

3.4. Interviewer

The interviewer belongs to a third dialect variety of EP, and thus all interviewees should feel free to use their native dialect alike. She is a native speaker of the southern dialect of continental EP, but not of the Lisbon area. She acquired EP on the west coast of the Alentejo. As she is a non-standard speaker of EP, all the interviewees should feel equally at ease to make use of their native dialect.

4. Ethical problems

Some interactions conveyed personal details of private life that from the beginning create in the researcher a sense of responsibility as regards disclosure of these aspects of the interview. We believe that not all parts of the dialogue present in the interview should be disclosed, even in an academic environment. Due to the singularity of the interview, the intimacy degree of the participants grew along the interview. Most of the interviews show speaker behaviour close to the vernacular, as expected. Note that most speakers had no previous knowledge of each other and met only once before the recording.[2]

(2) CPE-Var allows the identification of different attitudes from speakers towards the interviewer and the interview itself. Most speakers accommodate linguistically to the interview situation, some showing more careful speech than others.

There were speakers who mentioned having witnessed sexual assault committed by children, other speakers had been involved in conflicts with fire guns, others revealed aspects of their health and private life, events which the researcher deemed inappropriate to freely disclose. For this reason, one of the most important ethical problems identified is related to secrecy, in other words, the need for the researcher to keep private life information only to herself, if it was disclosed in the interviews. How can private information be kept when it is present in the recording and the recording is going to be transcribed by undergraduate students? How can the researcher assure a speaker's anonymity if speaker's name is spontaneously given during the interview? How can anonymity be preserved if acoustic study of the voice recorded can reveal a speaker's identity?

Another ethical problem relates to the use of the data for objectives different from the ones initially devised. Speakers were informed of the general aims of the research. They cannot, however, be informed of new research goals since 15 years have elapsed, their phone numbers have changed and they cannot be reached anymore. This problem is sensitive since tacit consent was obtained and the law now states the need for signed consent. If written consent had been obtained, the rights of the researcher would have been assured. Does tacit consent have the same value as written consent? We will discuss how Portuguese law rules over this subject matter.

A third ethical problem is related to the right of speakers to their voice in legal recordings, such as the ones we have in CPE-Var. What can one do with the informant's voice? Can we manipulate the interview, for instance, to erase parts of the speech deemed to be intimate or parts where the name of the speaker is given?

A fourth problem concerns the rights of the researcher to the original research. In this particular case, the researcher created her own model of interview (although based in Labovian interviews), she collected the whole set of recordings, she transcribed most of the interviews and she is also the author of the database where samples of the data are included. Are there any limits to her authorship and copyright due to the kind of data involved (namely, voice recordings)? All the tasks named above are time consuming and a lot of effort was made to bring them about and to keep them completely private.

Does the researcher have the right to combine linguistic information and speakers' identity or can she only deal overtly with linguistic information and the speaker's profiles? Can the researcher use the data for academic

purposes or only for the PhD for which it was initially designed? How long are speakers entitled to their privacy? When can the data be freely used?

These are the major ethical topics which will be discussed presently.

The use of CPE-Var in the course of the PhD dissertation raised no ethical problems to the researcher because all parts of the transcriptions presented avoided private or sensitive topics. Moreover, the academic environment is the one initially anticipated. In spite of that, a first problem concerning the use of the CPE-Var interviews came up when transcriptions of the whole interviews were to be made. Complete orthographic transcription was supposed to be made by undergraduates. This is when the researcher started to worry about privacy issues. How could she continue to assure privacy/confidentiality of the interactions and speakers' anonymity? Furthermore, were there in the interviews any hints about the speaker's consent, so that it would be undisputable that they could legally be used?

4.1. Ethical problems in detail

Let us start with the issue of consent since most ethical questions raised above relate to it.

Written consent is always required nowadays although it was regarded as unnecessary in the 1990s. CPE-Var does not have signed informed consent. Sometimes oral consent was given during the interview, sometimes it could be inferred from the interview itself. Tacit consent was always obtained, however, since all the speakers agreed to willingly participate. Furthermore, they read aloud a large amount of reading materials to be clearly captured by a visible tape recorder placed in front of them. Their performance would be impossible had they been coerced. The interviews involved only previously contacted speakers who showed no a priori problems about speaking with the interviewer. No payment or any kind of feedback was promised since the research itself was not funded.

CPE-Var recorded interviews are then admissible data for language study and, furthermore, the voices captured are useful for several purposes. We believe that we have obtained tacit consent since the interviews were previously scheduled and recorded with a table microphone and a portable PMD tape recorder in front of the speaker. The law excludes evidence in court only if information is obtained through private life intrusion, through home, mail or communication violation without the consent of the holder. This kind of evidence would then be considered null and void. None of the

above-mentioned situations occurred during CPE-Var recordings. These recordings would, therefore, be legitimate even in a court of law.

The interviews frequently include the speakers' name, as we have mentioned. Furthermore, speakers may often be identified by their knowledge of the topics mentioned, opinions given, names dropped during the talk, etc. As a consequence, privacy / anonymity / confidentiality issues were at stake, if undergraduate students were to have access to the interviews. Students asked to perform the orthographic transcription task had been previously coached and asked not to disclose any information present in the recordings without the researcher's consent. No problem arose in this transcription phase or ever since, even though we have no guarantee that undergraduate students kept their word. We are fully aware that that kind of speech obtained during the interviews was only possible in those particular circumstances: two people alone, face-to-face, in a dialogue setting that creates the illusion of intimacy. This illusionary sense of intimacy was considered fundamental in order to get the sort of data the research set out to analyse. Our target was semi-informal speech data which are only available when the informants trust the interviewer and feel at ease.

Phonetic transcriptions were made by the researcher herself; therefore, no problems occurred during that phase either. For that matter until conclusion of the researcher's PhD no problems came up in the use of the data of CPE-Var.

A database was built to keep all the relevant data used in Rodrigues's PhD. Data include samples of phonetically transcribed occurrences of words in context, classified linguistically. The classification includes the speakers' sociolinguistic profile. The database is available exclusively to the holder; it is considered private and Rodrigues holds all legal rights to it, namely, reproduction, display, manipulation, etc. See below what the Portuguese law states about copyright.

Some ethical issues came up later, when voice of some informants was to be used to analyse their acoustic properties in order to establish criteria / parameters for voice identification. Would it be morally reasonable to use the data for this specific purpose? The main area of research had been mentioned, but the specific topic addressed was not. We could not use a methodology that would constrain the informant's speech in order to collect good quality informal speech data. In this sense, our means were justified by the objectives of data collection. Concessions were made due to the need to collect spontaneous speech data of appropriate quality. Even if it is reasonable, does the law in any way prevent this kind of research with

this kind of data? How can speakers' anonymity be assured if their voice is under acoustic scrutiny? Forensic Phonetics has developed hugely in recent years and, if careful treatment of appropriate voice data is made, it can provide reasonable evidence of the speaker's identity beyond the shadow of a doubt. We participate in a research project that aims at evaluating the relative importance of acoustic parameters towards speaker identification focusing in EP dialectal data. Can we use CPE-Var data for this research? It is worth noting that the recordings would not exist had Rodrigues not made them - they are therefore a product of her technical work and she may argue that she is the owner of this 'artifact'. More than 15 years have elapsed since data were collected: is this enough to allow for disclosure of the data for different research uses?

Since consent for disclosure of the interviews contents was inferred, all parts of the CPE-Var data that are not deemed private nor can they do the speaker any harm can be used. These data are disclosed by means of the orthographic/phonetic transcriptions but not through the all sound recording of the voices, even though some parts of the recording can be disclosed to exemplify several research issues. This means that crimes, intimate affairs, health condition, etc. mentioned by some speakers are not disclosed. What about the confidentiality of new ideas, theories, phone numbers which are revealed, etc? They were not disclosed in the first attempts to show the data. Meanwhile, after 15 years we do not deem that secrecy of these matters is needed. Time has rendered such secrecy useless (for instance, telephone numbers have changed and are therefore no longer accessible).

Data were only disclosed for scientific purposes in small samples (proportionality) enough to verify the hypotheses. Transcribers were asked not to disclose any of the information present in the interviews lest they face prosecution.

What about anonymisation?

Anonymisation of transcriptions was mainly assured by means of coding procedure. For instance, speaker Maria Melo (fictitious name) was codified as informant number 32 with the profile LF32, meaning that it is a Lisbon female graduate speaker in the 2nd age group. Speakers were thus retrieved straightforwardly. What about the recordings? Although the speaker's name does not appear clearly in all recordings, in some it does either at the beginning or somewhere else.

When the interviewee's first name appears, it can be replaced by another name in the transcriptions (normally, a similar one, for prosodic reasons). What should be done within the recording itself? Should we replace the

name by a silence interval? This changes the recording integrity and may even be termed manipulation. What else can we do? The solution we have adopted is simply, for the purposes of the research project concerning the acoustic analysis of voices from the CPE-Var, not to use those parts of the interview which include the speakers' identification or other identifiable people.

Do we have the legal right to use CPE-Var data for other research uses that could not have been anticipated from the beginning? Are there any limits to using it?

To assure anonymisation only the number of the informant, if possible, is given. If that is not possible, then the substitute (the name by which the proper name was replaced) is used. Otherwise the code attributed to the cell of the speaker is mentioned.

Using this procedure guarantees that our speakers cannot suffer any damage for having placed their trust in the interviewer during the interview. A speaker who might potentially be charged with committing a crime cannot be charged since his anonymity was preserved. This procedure also renders it impossible for any absent person referred to in the interview to bring charges against us since we do not disclose his/her name.

Even though we did not promise speakers any feedback of our collection of data, they may receive feedback, if they so wish. Informants may have access to the publications authored by Rodrigues, including books, scientific papers, presentations to symposia, etc. They can also reach Rodrigues at the University.

5. Ethical and legal rights

Legal limitations to the use of acoustic signal data recorded and their transcription outside the context where they were collected, even if exclusively in academic environment and for research purposes, constitutes a very important and open issue from the Ethics and Law viewpoints.

We will now address the above-mentioned open questions, particularly those that concern the speaker's personal data and such data legal protection – i.e. the human being and his own data -, namely, the right to the reservation of the intimacy of private and family life under Article 26 Nr. 1 of the *Constituição da República Portuguesa*, the Portuguese Republic Constitution.

5.1. Privacy

It is worth noting that the above-mentioned rights have been deemed so important in modern societies that they were afforded constitutional dignity. In addition, according to publications of Professors Gomes Canotilho and Vital Moreira (2007), the normative scope of the fundamental right to the reservation of the intimacy of the private and familiar life must keep in mind the following aspects: respect of each other's behaviour, respect of anonymity and the respect of rules of life in society. These fall under personality rights, the violation of which is punished by law.

Canotilho and Moreira (*idem* p. 182) argue that "the right to the secrecy of human being" (image rights, the right to speak, right to private life) should be intrinsically linked to personality rights: "the constitutional criterion should perhaps start from the concepts of privacy - (Article 26 Nr. 1) and human dignity - (Article 26 Nr. 2 of the Portuguese Constitution) so as to define a concept of privacy of each person, culturally appropriate to contemporary life.".

In line with this, Andrade (2004, p. 498) also contends that the latest revision of constitutional law: "raises to the constellation of Rights, Freedoms and Guarantees in criminal matters (Article 26, Nr. 1) the right of every man – and him only, to decide who can record his voice".

He further emphasised: "the full availability of the human person on the spoken word as a direct expression of his/her own personality and dignity", which was enshrined in the Constitution.

Andrade also considers the right to speak from a double dimension perspective:

a) A positive dimension – meaning the legitimacy to authorise the recording and the audition, freely, with no restrictions;
b) A negative or exclusive dimension – meaning the freedom to refuse the recording and the audition, with no restrictions.

Andrade argues that the law protects the right to speak as personal goods, as a direct expression of the speaker's personality in communication with other members of society.

In this respect it is worth noting that the fundamental provision of said Article 26 Nr. 1 also grants protection under legal and constitutional law to the image and to the voice, more precisely to the word, by listing these rights, together with the above, such as the right to good name, to personal identity, to the personality development and to citizenship. This means that

the word is protected by the supreme law of the nation by making sure that particular caution must be exercised when capturing voice and image to avoid conflicting with other rights, including the right to privacy.

5.2. Database and Copyright

The Portuguese Code of Copyright and Related Rights of 1985 (Decree-Law Nr. 63/85, March, 14th, amended by Law Nr. 65/2012, December, 20th) aims at protecting authors of literary, scientific and artistic works. In its Article 1, Nr. 1, it defines such works as: "intellectual creations of the literary domain, scientific and artistic in any way externalised."

This provision protects works from any misuse, as well as from any economic benefits arising out of such misuse or exploitation. The Copyright Code protects patrimonial rights, but also the rights of a personal nature, called moral rights. Influenced by European Union law, the copyright term was extended and works are now protected for a period of 70 years (when before the term was 50 years) to strengthen the preservation of historical and cultural heritage.

It should be noted that along with the industrial property law, protection of the literary, scientific and artistic works as well as the rights of creators, both from the economic and moral standpoints falls under copyright law, one of the areas of intellectual property.

Regarding the contents of the data on CPE-Var, described above, the legal framework should comprehend all the data recorded as forming a body, which must be understood to mean a *database,* and to that extent its legal protection is conferred. Therefore, in our research there are two databases: one is the collection of recorded interviews and the other the database created to process quantitatively the phonetic transcription of word occurrences in the CPE-Var recordings.

Rodrigues, in her capacity of CPE-Var author, is entitled to act in the management, protection and defense of her work, and may authorise its use by third parties.

Nowadays databases enjoy broad legal protection, both under the Portuguese and EU law. Thus, Decree-Law Nr. 122/2000, July, 4th, which transposes into national law EU Directive Nr. 96/9/EC of the European Parliament and the Council, March,11th, sets out the legal protection regime of databases. The solution now adopted, by overriding EU law, provides double protection: for one, the databases that constitute intellectual creations

are protected by copyright with some special features; and then it provides protection of the investment of the manufacturer of certain databases.

The Portuguese Law then sets out what shall be and is meant by database by defining it in its Article 1, Nr. 2 (subject) as: "a collection of works, data or other materials, arranged in a systematic or methodical way and individually accessible by electronic means or other."

It also establishes that databases are protected by copyright and that such protection is subject to country of origin, considering the author who is qualified as such by the law of the database country of origin.

The European Community, given the importance of fundamental freedoms referred to in the Charter of Fundamental Rights of the European Union, strengthens the protection of fundamental rights, setting out (under Article 7) respect for private and family life: "Everyone has the right to be respected for his private and family life, his home and communications." and (under its Article 8, Nr. 1) entitled protection of personal data it expressly lays down: "Everyone has the right to the protection of personal data", reinforcing and conferring legal dignity to this subject matter.

Portuguese Law assures data integrity and preservation of personal dignity, which is one of the most important legal principles of the Portuguese Constitution. It also assures the right use of the data and, for that reason, good-faith is imperative in all manipulation of the data and in the general use of the data exclusively in academic environment to protect its quality, the author and the speakers' anonymity.

5.3. Personal data and consent

In addition, in this context the "Law on Personal Data Protection," Law Nr. 67/98, October, 26th, transposes into the Portuguese legal system Directive 95/46/CE of European Parliament and the Council, October, 24th 1985 on the protection of individuals with regard to the processing of personal data and such data free movement.

Law Nr. 67/98 first sets out the general principle that the processing of personal data must be made in a transparent manner and with respect for private life as well as rights, freedoms and guarantees. In our view, this principle applies to the contents of the interviews and to the personal data reported by respondents during the interview, which were duly protected and complied with such legal requirements.

The Personal Data Protection Law defines, among others, the meaning of the following expressions: personal data, personal data processing and data consent, due to the particular importance of this latter term. Accordingly, the law sets out clearly (Law Nr. 67/98, Article 3 h)) what is meant by consent of the owner of the data: "any expression of intent, free, specific and informed, under which the holder accepts that personal data be processed."

The interviewee who provides access to the interview and simultaneously authorises its recording is thereby granting his free and spontaneous consent. As argued by Andrade (1992, p. 251) consented recording or its use is excluded from the typicality of illegal events for the following reason: if consent given by the author is valid, it can never be alleged to be null or void on grounds of intrusion of privacy, since consent was given by the holder (a right freely available) and as such the interview has the holder's effective intervention.

The National Commission for Data Protection (NCDP) was created in Portugal to examine compliance with the legislation. The NCDP is an independent administrative body with power to control and monitor personal data processing, with strict respect for human rights and fundamental freedoms and guarantees set out in the Portuguese Constitution and Law. The NCDP is the National Authority for Control of Personal Data that liaises with the data protection supervisory authorities in other countries.

6. Concluding Remarks

Our paper described in detail the CPE-Var collection of sociolinguistic data. During the processing of the data, several problems arose which posed some ethical dilemmas. In the above sections we have discussed some of those problems from a legal point of view, namely privacy, confidentiality, anonymity, use of data outside the scope of the former investigation for which they had been collected, copyright issues, personal right to voice and image recording, consent, etc.

As regards privacy/confidentiality, it should be noted that CPE-Var speakers have never expressed their wish to privacy. All the recordings were obtained after a brief appointment where speakers were informed of the general research goal and where they gave their consent verbally to the recording. In the 1990s this was deemed adequate procedure to collect legal speech data in Portugal. No image was captured. Since the recorder and a clearly visible microphone were placed before the speakers, all the recordings had the interviewees' spontaneous collaboration. After the reading

tasks, interviewees spoke freely and for as long as they wanted. Their consent to the recording is obvious.

Orthographic transcriptions of the spoken speech were carried out by undergraduate students under the supervision of the corpus author. Undergraduates were asked not to disclose the names and information present in the interviews. Phonetic transcriptions were carried out by Rodrigues only and therefore no problem arose during that task.

Rodrigues is deemed to be the author of the collection of recordings and she has therefore exclusive power to explore and give access to the data, once data integrity and appropriate use are assured. Recorded voice is considered property of the author and is used exclusively for research purposes. Exemplification is restricted to small excerpts proportional to the needs.

A code was attributed to each speaker in order to keep anonymity in the database created to process the phonetic transcriptions. Speakers have been ascribed an alphanumerical code that combines information about their provenience, age, gender and instruction level.

Rodrigues uses CPE-Var to explore voice quality. For that purpose, she only uses the parts of the talk that do not contain personal data or any other private information.

All the results of scientific exploitation of CPE-Var data are currently available in several publications, providing the social feedback desired in all scientific research.

We believe that speakers collaborated with the researcher in good faith and as a consequence, Rodrigues feels she is under obligation to respect them by making sure that data will only be used in their best interest (that is, by producing no damage to their lives).

Portuguese Law follows EU law in the protection of all fundamental rights of speakers and of the author of the various products of this sociolinguistic research. In Social and Human Sciences most of the above mentioned ethical issues have proved to be particularly sensitive within Medical Sciences. The importance of good care of the human body and the right to receive best treatment are major concerns in this area.

Above all, it is fundamental for each researcher to establish a clear line between what is ethically appropriate/irreproachable behaviour and behaviour which is ethically reproachable. If this line is never crossed and if the researcher acts according to the universal principle of good faith s/he will be able to process ethically all sensitive items presented by the data. The researcher has to reach a compromise between the best interest of the research and the best interest of the speakers in a sociolinguistic research.

Does our methodology pass the ethical tests normally used, say, in Medical Sciences?

3) The impartiality test – Would we want someone to apply to us the methodology we have adopted?
4) The universalisability test - Would we want our methodology to be applied to any other similar cases?
5) The interpersonal justifiability test – Do we have good reasons to justify our methodology options?

Our methodology passes the impartiality test. We would not mind if equal methodology was applied to us, since the interview was previously scheduled and lasted only one hour. Our recordings assume, from the beginning, that a semi-informal dialogue involving an interviewer from a dialect different from the informant's would occur. As it occurs only once, in principle, no damage results to the informant, apart from possible waste of time. Informants could refuse to participate in the study, either at the beginning or during the interview (free will principle). Furthermore, the research's main goal is socially recognisable as good, that is, the acquisition of more scientific knowledge of EP language daily use.

Our methodology also passes the universalisability test. It can be used by other researchers, since it is described in depth in published papers and no damage results from its application to the target speakers (on the contrary, its application enriches society, as more scientific knowledge can be obtained).

It also passes the interpersonal justifiability test since our reasons for choosing this methodology are explainable and easily understandable.

Portuguese literature on ethical issues in sociolinguistic research is scarce. We found no works published on the subject, besides the ones on bioethics, economic and human resources, law or computational engineering, media and philosophical theories were identified. Problems may arise if there is a conflict between two methodologies achieving the same result, one more beneficial than the other. Our methodology is not intrusive (it produces no harm to the speakers) therefore we believe that it is minimally invasive of personal life.

References

Amarakone, Keith and Sukhmeet Panesar (2006): *Ethics and the Human Sciences*, Elsevier Limited, printed in Italy, p. 211.

Andrade, Manuel da Costa (1992): *Sobre as Proibições de Prova em Processo Penal*, Coimbra, Coimbra Editora, p. 121 and ff..

Andrade, Manuel da Costa (2004): *Consentimento e Acordo em Direito Penal,* Coimbra, Coimbra Editora, ISBN 972-32-0438-X, p. 498 and ff.

Britz, J. J. (un-dated), "Technology as a Threat to Privacy: Ethical Challenges to the Information Profession", http://web.simmons.edu~chennitNIT-279696-025-Britz.html.

Canotilho, Gomes and Vital Moreira, (2007): *Constituição da República Portuguesa Anotada,* ISBN 9789723214628, p. 182 and ff.

Castro, Ivo (2006) *Introdução à História do Português,* Edições Colibri, Lisboa.

Cintra, Luís Filipe Lindley (1971) "Nova proposta de classificação dos dialectos galego-portugueses," *Boletim de Filologia*, XXII, 81-116 (repr. in CINTRA, Luís Filipe Lindley (1983), *Estudos de Filologia Portuguesa*, Lisboa).

Código do Direito de Autor e dos Direitos Conexos, 1985 (Decree-Law Nr. 63/85, March, 14th ammended by Law Nr. 65/2012, December, 20th).

Constituição da República Portuguesa, April 2nd, 1976 – 7th, Constitutional Revision Law Nr. 1/2005, August, 12th.

Carta dos Direitos Fundamentais da União Europeia – Jornal Oficial (2010/C 83/02), March, 30th, 2010.

Dattalo, Patrick (2010): "Ethical Dilemmas in Sampling," *Journal of Social Work Values and Ethics,* Volume 7, Number 1, Copyright 2010, White Hat Communications.

Decree-Law Nr 334/97, November, 27th [transposing into Portuguese Law *Directive Nr. 93/98/CEE, of the EU Council, October, 29th* - changing the term of copyright and related rights protection, *Código do Direito de Autor e dos Direitos Conexos*].

Decree-Law Nr 122/2000, July, 4th [transposing into Portuguese Law *Directive Nr. 96/9/CE, of the European Parliament and of the European Council, March, 11th 1996*, relative to electronic/digital databases protection].

Duarte, Inês, Maria João Freitas, Anabela Gonçalves, Matilde Miguel and Celeste Rodrigues (2002): "Geometria de traços e distribuição de pronomes sujeito em PE e em PB", III *Workshop do projecto Português Europeu e Português Brasileiro: Unidade e Diversidade na Passagem do Milénio,* Faculdade de Letras de Lisboa, September, 23-25th.

Iserson, Kenneth, Arthur B. Sanders and Deborah Mathieu (eds.) (2001): *Ethics in Emergency Medicine*, Galen Press, Ltd., 2nd ed.

Labov, William (1981) "Field Methods of the Project on Linguistic Change and Variation", in *Sociolinguistic Working Papers*, Nr. 81, p. 28-53.

Mateus, Maria Helena and Ernesto d'Andrade (2000): *The Phonology of Portuguese*, Linguistics, OUP, Oxford.

Protecção de Dados Pessoais – Law Nr. 67/98, October, 26th [transposing *Directive of EU Nr. 95/46/CE, of the European Parliament and European Council, 1995, October, 24th*, into Portuguese Law relative to the protection of individuals concerning personal data processing and their free circulation].

Protecção Jurídica das Bases de Dados – Decree-Law Nr. 122/2000, July, 4th [transposing into Portuguese Law *Directive Nr. 96/9/CE, of European Parliament and of European Council, March, 11th*].

Rodrigues, Celeste (2003): *Lisboa e Braga: Fonologia e Variação*, FCT-FCG, Lisboa.

Schramm, Fermin Roland (2004): "A moralidade da prática de pesquisa nas Ciências Sociais: aspectos epistemológicos e bioéticos", *Ciência e Saúde Coletiva*, 9 (3), p. 773-784.

Simões, Deolinda (2010(11)): *Medidas Legislativas para Protecção da Cadeia Alimentar no Âmbito da Importação e da Admissão*, Master Thesis FMUL (Mestrado em Medicina Legal e Ciências Forenses), Lisboa, pp. 198 (http://hdl.handle.net/10451/2701).

vigário, Marina & Sónia frota (2003) The intonation of Standard and Northern European Portuguese. *Journal of Portuguese Linguistics* 2-2 (Special Issue on Portuguese Phonology edited by W. L. Wetzels), 115-137.

recensões