

***MadSex*: collecting a spoken corpus of indirectly elicited sexual concepts**

Andrea Pizarro Pedraza¹ 

Published online: 24 November 2018
© Springer Nature B.V. 2018

Abstract This paper introduces *MadSex*, a spoken corpus of 54 sociolinguistic interviews in Spanish based on the topic of sexuality. It was collected in order to study the cognitive sociolinguistic variation of sexual concepts. The paper presents and justifies methodological decisions taken during design, collection and transcription stages. Informants were selected in Madrid, based on a pre-stratified sample divided by sex, age and level of education. The interview methodology relied on an opinion questionnaire designed for the indirect elicitation of sexual concepts, which overcame successfully the limitations imposed by the low frequencies of semantic variables in discourse and the impact of sexual taboo in interaction. Relevant aspects of fieldwork, empathy and ethical protocols are also detailed in the paper. Transcription and markup are explained. Finally, an overview of the corpus is given, as well as some research papers based on it. Examples of the questionnaire are also provided.

Keywords Spoken corpus · Data collection · Interview methodology · Semantic variables · Sexual concepts

1 Introduction

MadSex is the acronym for *Corpus madrileño oral de la sexualidad* (*Madrilenean spoken corpus of sexuality*), a corpus of 54 transcribed interviews of speakers from Madrid (Spain) who answered a questionnaire about sexuality in 2011–2012. The

✉ Andrea Pizarro Pedraza
andrea.pizarro@uclouvain.be

¹ Faculté de Philosophie, Arts et Lettres, Université catholique de Louvain, Place Blaise Pascal, 1 bte L3.03.33, 1348 Louvain-la-Neuve, Belgium

corpus was collected as part of a PhD project¹ with the goal of analyzing the semantic variation of taboo concepts in spoken discourse and its potential relation to context, social variables of the speakers and concept features, therefore bringing together sociolinguistic and cognitive linguistic interests (Pizarro Pedraza 2013, 2016b). This goal imposed restrictions that determined the methodological choices that will be described in this paper.

Semantic variation is probably the most problematic object of study within sociolinguistics, a sort of “non-entity” (Hasan 2009: 41) which is almost totally absent from variation studies. This absence is partly due to the original definition of the sociolinguistic variable, only applicable unproblematically to phonological variation but also to the theoretical model of language assumed by sociolinguistics (Geeraerts and Kristiansen 2014; Hasan 2009). First, the variable is defined as a unit of the linguistic system whose variants are sensitive to linguistic and extralinguistic contexts (Labov 1972: 8). This has been interpreted as a need for the variable to be discrete and definable,² which doesn’t correspond to the fuzziness of semantic meanings (Geeraerts and Kristiansen 2014; Geeraerts 2006). Besides, the variants have to be referentially or functionally equivalent (Labov 1972). At the lexical level, this requirement was equated with a need for strict synonymy (López Morales 2009: 25–26) reducing variational studies to formal or lexical variation. Moreover, the variants need to be very frequent. These criteria are difficult to meet by variants at the semantic level, which more often than not carry differences in meaning—at the semasiological (Pütz et al. 2012), but also at the conceptual onomasiological level—and occur very infrequently in discourse (Arroyo and Luis 2005: 79). As a consequence, semantic variation was often relegated to the periphery of sociolinguistics.

In the past decade, the surge of cognitive sociolinguistic studies (Janicki 2006; Kristiansen and Dirven 2008; Geeraerts et al. 2010; Pütz et al. 2012, 2014; Geeraerts and Kristiansen 2014; a.o.) has contributed considerably to solving this problem. Cognitive Sociolinguistics (CSL) proposes a method for the study of semantic variation -semasiological, formal onomasiological and conceptual onomasiological variation (Geeraerts et al. 1994)—building on the experientialist and usage-based theory of meaning from Cognitive Linguistics (mostly) through corpus-linguistics methods (Geeraerts and Kristiansen 2014).

Although CSL is unquestionably useful for unblocking the study of semantic variation, up until now, the studies have mainly focused on available corpora. These sources have some problems, such as only providing offline data -blind to the communicative situation—and the impossibility of controlling external variables (Heylen et al. 2008). These are not minor problems for sociolinguistics, which wants to understand society as much as language. Thus, such corpora—despite being large and available—are sometimes unsuited or unsatisfactory for solving particular

¹ The project was financed by the CSIC (JAE-predoctoral scholarship).

² I would like to thank the anonymous reviewer for pointing out Labov’s own study of continuous variation in sociophonetics. In fact, the limitations in the analysis of semantic variation come from the interpretation of the definition of the variable, more than from the later operationalization and results related to non-discrete variables in other levels of linguistic analysis.

research questions, such as the ones we will present in this paper.³ Arguably, in some cases, collecting corpora ad hoc is still the most adequate solution, despite it being time-consuming and requiring a high degree of personal involvement.

On top of the well-known problems of corpus collection, the case presented here posed a number of specific challenges related to the study of the semantic variation of sexual concepts. This paper will explain those problems and focus on the concrete solutions adopted for each of them, with the hope of being useful for scholars willing to collect corpora for semantic variation, or for linguistic taboo in particular. First, Sect. 2 will justify the motivations for such a project in order to understand the goals that led to the particular methodological decisions; then, Sect. 3 will explain the specific corpus design, and finally Sect. 4 will describe the fieldwork and data collection and Sect. 5 the transcription phases. A description of the final corpus and some research derived from it will be provided (Sect. 6), followed by some conclusions (Sect. 7). The appendices show the target sexual concepts and a sample of the interview questions.

2 Previous research and motivations of the corpus

Linguistic taboo refers to the phenomena related to offensive meanings and words, in its referential functions, such as dysphemism or euphemism, or non-referential functions, such as swearing. Research on referential taboo has developed considerably in the last decades (Crespo Fernández 2015; Casas Gómez 2009; Pizarro Pedraza 2018b; Allan and Burridge 2006; Allan 2018; a.o.), which indicates that researchers have overcome to some extent the limitations related to the investigation of topics such as sexuality, death, disease or scatology.⁴ This is very positive, since taboos can help shed light on various aspects of linguistics, such as conceptualization, language acquisition, perception, the connections with social norms or traits of character (Jay 2000, 2009). In particular, sexuality, one of the most global taboos (Crespo Fernández 2015; Allan and Burridge 2006), is considered a central aspect of human life that is related to many facets of existence (WHO 2006), an embodied and social experience, which is inherently discursive (Foucault 1976).

The goal of the present study was to analyze empirically a commonly-accepted idea that had not been statistically tested yet: the fact that linguistic taboo varies according to cultural and social aspects (Casas Gómez 1986; Galli de Paratesi 1964; Crespo Fernández 2007; Jay 2000; Uría Varela 1997; a.o.), which implies that not every taboo concept is (equally) taboo in every group or even within groups (Allan and Burridge 1991). If this were true, a sociolinguistic study should be able to reveal a meaningful correlation between the variation in the expression of sexual concepts in discourse and different social aspects of the speakers. Moreover, based on the

³ Moreover, there is a current interest in more “artificial” (performed) styles which are also extremely revealing of the social meanings of linguistic variants (Eckert 2005).

⁴ Despite having reached some level of visibility and acceptance, taboo researchers still report reactions of discomfort or rejection from their peers when talking about their studies at conferences, for instance.

cognitive sociolinguistic assumption that language reflects the conceptualization of particular realities, which vary with social experience (Kristiansen and Dirven 2008), it should be possible to analyze how different speakers conceptualize sexuality by observing differences in the conceptual onomasiological variation⁵ of the expressed sexual concepts. In order to perform this analysis, the study needed accessible social information and large amounts of sexual concepts with some particular characteristics.

2.1 Spoken data

Previous linguistic studies about taboo in Spanish had generally been based on the study of written texts, such as literature (Montero Cartelle 1996, 2010), or written questionnaires (López Morales 2005; Martínez Valdueza 1995; Calvo Shadid 2009).⁶ This research wanted to contrast their results with the functioning of taboo in spoken discourse. This novelty in the focus was supported by methodological interest: face-to-face situations force the speakers to deal with the *ad-lib* expression of taboo, without having the production time that they do have while writing. Such an online problem-solving environment seemed particularly interesting to observe speakers' behavior towards taboo. Other methodological options to collect spoken data, such as focus groups (see Litosselitti 2010: 166 and sec., for an overview), were also considered but discarded, since for this particular corpus, the interest was to control the context (participants, topics, cf. infra, "Controlled structure") as much as possible, in order to perform a more reliable quantitative analysis, and not so much to explore at length particular topics or attitudes. Certainly, it would be very interesting for the future to use other data collection methods in order to analyze qualitatively the functioning of linguistic taboo in other interactive contexts.

2.2 Maximum variability granted

Sexual concepts are known to be the source of extremely varied equivalent expressions, as is shown by the bulkiness of specialized dictionaries (Cela 1989; Rodríguez González 2011; a.o.). These utterances can be classified along an x-phemistic continuum -euphemism, orthophemism, dysphemism—(Allan and Burridge 2006), depending on the intentions of the speaker. They can be expressed through construal strategies that may rely on every linguistic level (Uría Varela 1997: 10–15), sometimes in idiosyncratic ways that lead to interesting *hapax legomena*, like novel metaphors (Crespo Fernández 2015: 55). Restricting the study to a pre-defined selection of lexical variants would ignore a central aspect of linguistic taboo, which is its extreme variability. This study wanted to guarantee a maximum level of freedom for the speakers to express the concepts they wanted in the way they preferred, from very lexicalized forms to less predictable utterances, such as very vague or very idiosyncratic expressions.

⁵ Conceptual onomasiological variation occurs when "a referent or type of referent may be named by means of conceptually distinct lexical categories" (Geeraerts et al. 1994).

⁶ The only exception is Danbolt Drange's (1997) master's thesis.

2.3 Controlled structure

Nonetheless, so as to facilitate the analysis, it was important to allow that freedom within a structure that would help for the bottom-up extraction of the data and the determination of the reference. As a matter of fact, context is sometimes the only way to understand sexual reference in the case of vague expressions or *hapax legomena*, and it seemed suitable to keep it under control.

In light of the previous features, the corpus needed to have: 1. sufficient tokens of the variable (sexual concepts). 2. Social information about the speakers. 3. Spoken data collected in similar situations, and 4. Maximally varied concepts within a comparable structure. Criteria 1, 2 and often 3 are basic premises for any sociolinguistic study, while criterion 4 is specific of this study, but potentially applicable to any study on semantic variation.

With this in mind, the available Spanish corpora were discarded for two reasons. First, sexuality is only very marginally present, which made it impossible to fit the criteria of sufficient data (Labov 1972). Second, an extraction of data based on predefined lexical elements goes against the bottom-up intention to observe the maximum variability of taboo, with a focus on what is possible and not only on what is common, which is a limitation of corpus studies (Glynn 2014: 27). Moreover, the offline nature would complicate the analysis of production conditions. Other sources, such as TV or radio programs about sexuality had the problem that social information of the speakers was not always available. After discarding these less time consuming options, a choice was made for the collection of an ad hoc corpus. The next sections will explain how the above-mentioned criteria were included in the corpus design and the different phases of our corpus collection.

3 Corpus design for *MadSex*: sexual concepts as target variable

The first step in collecting a corpus ad hoc is to map out the desired features of the outcome and delineate possible solutions for each of them (Leech et al. 1995: 1). In this particular study, the projected spoken nature of *MadSex* guided the choice of a method that is well established in sociolinguistics, the one-to-one interview. This method is relatively malleable and it could be designed in a way that would include features 1–4 mentioned above. Nonetheless, despite the methodological development of this technique (Labov 1972; Labov 1984; Tagliamonte 2006; Schilling-Estes 2008; Wolfram 2011), it also involved two problems in our case: those derived from working with semantic variables in general (low frequency) and those derived specifically from its tabooed nature (uncomfortable topic).

As opposed to other variables that occur abundantly in discourse, such as those at the phonetic level, semantic variables need to be carefully elicited for them to appear sufficiently, or even at all. Traditional dialectal studies tackled this problem with onomasiological questioning (*How do you call the tool that serves as...?*) or visual priming (Pop 1950; Rabanales 2004; García Mouton 2015). The problems of these techniques are, on the one hand, that they decontextualize the concepts, which makes them unsuitable for observing how linguistic taboo—or any semantic field—

happens in (spoken) discourse, and on the other hand, that they could boost the effect of taboo by asking for sexual concepts directly and even more by offering visual priming. With that in mind, an opinion questionnaire was designed: it focused on the topic of sexuality as a base for a semi-directed interview, with the purpose of triggering abundant sexual concepts without asking directly about them.

Initially, a basic list of desired target concepts was elaborated based on the conceptual division of the recent *Diccionario del sexo y el erotismo* (Rodríguez González 2011), with the following categories (and a number of subcategories): body parts, physiology, personal qualities (e.g. sexually attractive), sexual identities, sexual practices, pleasure and desire, effects of sexual relations (e.g. pregnancy), contraceptive methods and prostitution, and other activities related to sex. This rough list was used for drafting questions that would potentially include those concepts in their answers.⁷ In that way, the speakers would elaborate their answers and produce sexual concepts within fragments of discourse. In order to motivate interest and longer answers, it had the appearance of an opinion survey about aspects of sexual life (cf. Table of target concepts per question, in “Appendix 1”).

The structure of the questionnaire was essentially the same as in other sociolinguistic interviews: it starts with questions about the social and linguistic background of the speaker that are followed by the body of the interview (Tagliamonte 2006: 39). The personal questions focused on the family of the speaker, religiosity, cultural practices, way of life and approximate salary of the household, among others. The body of the interview is composed of thirty-five questions divided in five “modules” (Labov 1984) that move progressively from the more general to the more personal topics (p. 44): sexuality in the Spanish media (module 1), social changes related to sexuality in Spain (module 2), sexuality in daily life in Spain (module 3) and sexual education (module 4, cf. “Appendix 2”), which would focus on the speaker’s experience. A final module about Spanish sexual laws (module 5)—zooming out of the personal experience—was added in order to elicit more argumentative answers and collect previously elicited concepts (abortion, homosexual marriage, prostitution, etc.) from a different perspective, which would potentially motivate different construals.

None of the questions asks primarily about the sexual life of the speakers or their intimate experiences, but in order to enhance their involvement, they were contextualized with regard to Spanish or Madrilénian society. This was typically done by asking their opinions about pieces of information present in the media at the time, or open debates in Spanish society (such as the presence of sexual contents on TV), but also situations witnessed in Madrid or in their own neighborhoods, hearsay or hypothetical cases (see doc. 1). For that reason, replication of this study in different locations would require a cultural adaptation.

Doc. 1. Examples of questions present in the questionnaire.

Hace poco se publicó un estudio que decía que el 70% de los programas de televisión que vemos contienen escenas de sexo; hay gente a quien le parece

⁷ Each question targeted potentially more than one concept and the same concept was targeted by more than one question, so as to guarantee its appearance in the corpus.

que se abusa de estas escenas. ¿A ti qué te parece? ¿Y en los programas que tú sueles ver? ¿Puedes dar algún ejemplo?

‘A recently published study stated that 70% of the TV programs we watch contain sexual scenes. There are people who think this is too much. What do you think? What about the programs you usually watch? Could you give an example?’

¿Has pasado alguna vez por la Casa de Campo o por Montera de noche? ¿Qué has visto? ¿Qué te parece?

‘Have you ever been to Casa de Campo or Montera [prostitution areas in Madrid] at night? What have you seen? What do you think about it?’

Structurally, the 35 questions were composed of an initial question complemented by follow-ups. The initial questions were usually designed as polar questions, which impose less presuppositions than wh-questions (Steensig and Heinemann 2013) and allow the speakers to engage in a topic or not. Most of the speakers elaborate long answers (Pizarro Pedraza 2016a). Moreover, the questions were formulated in order to avoid lexical priming: the main topic (sexuality) is introduced at the beginning of the interview but the questions themselves try not to include the target concept, relying on vague language or common ground.

This approach was carefully designed to indirectly trigger abundant and varied sexual concepts, without being intrusive. The questionnaire was tested through a pilot study and modified until its final version, which was designed to last about an hour.

4 Fieldwork and interviews

The target population was from Madrid,⁸ the capital city of Spain, which had then slightly more than 3 million inhabitants. It is administratively divided in 21 districts. Two were selected for the study: Chamberí and Villa de Vallecas. They were chosen because they were adjacent to other districts previously targeted by sociolinguistic studies –such as Salamanca and Puente de Vallecas (Cestero Mancera et al. 2015)—allowing potentially for some degree of comparability. Besides, they are socially and politically very different from each other, and they could bring about varied people with different experiences and opinions about sexuality.

A selective sampling method was applied and stratified by age (three groups: 20–34, 35–54, 54 and more),⁹ sex (women, men), and level of education (primary, secondary and university studies),¹⁰ with two informants per cell per district, which

⁸ The study was restricted to speakers who were born in Madrid or who had arrived in their childhood.

⁹ For the age group delimitation, this study followed the division of PRESEEA (2003), largely used in Spanish sociolinguistics.

¹⁰ The level of education was used instead of the complex variable ‘social class’, since it seems that the latter can explain linguistic variation on its own and is therefore a simpler approach, according to sociolinguists of Spanish (Arroyo and Luis 2005: 227 et seq.).

add up to 72 informants in total. With such a method, the purpose was not to claim representativity,¹¹ in the sense of portraying a miniature of the population (Sankoff 2004: 999), but to search for a variety of Madrilenian people. As will be explained later, some demographic subgroups were much easier to find and more participative than others, which suggests that if the sampling method had been based exclusively on willingness to participate, the final sample of speakers would have been quite homogeneous.

Fieldwork was carried out approximately during a year and a half, with interruptions for other research tasks. During that time, informants were searched for through personal network and institutional channels, such as local associations, libraries or cultural centers. The researcher's personal network was contacted, mainly through email, by asking acquaintances to extend the call to potential participants. Direct friends and family were discarded in order to avoid the effect of previous personal relation on the data and to ensure the same circumstances during the interview. At the institutional level, most of the contacts were made in person, by visiting different centers and recruiting volunteers. In order to maximize the visits, all contacts were kept, and only during the phone call sessions the suitable ones were selected for interviewing. The major inconvenience of this technique is that the people who participated in those institutional activities were already a rather homogeneous selection of the population. While older women were relatively easy to find through these channels, younger speakers, especially men and those with lower levels of study were totally absent. An attempt to target them specifically through a sexual health center rendered a few contacts with younger women (not all successful, eventually), but none with younger men. In the end, 54 interviews were collected, as some of the social subcategories were impossible to find in the time that we had for data collection (Tables 1 and 2). The distribution of the speakers reflects social aspects of each of the districts, for instance, in the fact that no male speakers of the lower education level for any of the age groups were found. Moreover, the snowball effect worked very well with female speakers in both districts (through an association and a school), and also with men in Villa de Vallecas, through a group of friends. In general, speakers of the highest education level were the easiest to find, certainly due to the bias of the researchers' own network.

The interviews were carried out by three different persons in an isolated environment. While the responsible researcher, a female, worked with the women, two male colleagues¹² collected the interviews from men, in order to control the possible effect of heterosexual communication on sexual taboo. Although the initial idea was to work with just one male interviewer, none of the collaborators was able to assume the workload involved by the project at the moment. As an attempt towards the highest homogeneity in the procedure, the same protocol was followed

¹¹ At the time of the study (2010–2011), Madrid had around 3.200.000 inhabitants. Taking Labov's (1971) 0.025% as a measure, a representative sample would have been 809 speakers, which is obviously unachievable for this kind of project.

¹² Enrique Menéndez Alba and Mario Martínez Zauner, without whom this project would have been impossible.

Table 1 Final distribution of speakers according to sex, age group and level of education in Chamberí district

Sex	Men			Women			Total
	Age group						
	1	2	3	1	2	3	
Education level							
Primary	–	–	–	–	2	2	4
Secondary	2	1	2	1	1	2	9
University	2	2	2	2	2	2	12
Total	4	3	4	3	5	6	25

Table 2 Final distribution of speakers according to sex, age group and level of education in Villa de Vallecas district

Sex	Men			Women			Total
	Age group						
	1	2	3	1	2	3	
Education level							
Primary	–	2	2	–	2	2	8
Secondary	2	2	–	2	2	2	10
University	2	2	1	2	2	2	11
Total	4	6	3	4	6	6	29

by all three researchers and feedback was given so as to ensure a coherent methodology.

One of the notions that guided the interviewing ethics was empathy. In this kind of interviews, establishing an empathetic link with the speakers is key for comfortable communication and for a successful interview (Wolfram 2011: 301). After all, the speakers are voluntarily giving access to personal information, and a satisfactory experience is probably the most basic step in *giving* them something *back* (Wolfram 1993; Labov 1982). Some manuals give special tips or techniques based on personal experience in the field (Tagliamonte 2006; Briggs 1986; Campoy et al. 2005; a.o.). Reading them is highly recommended before being confronted with the first interviews, since not many linguists are formally trained in the acquisition of those skills. On a practical level, reducing the effect of the physical elements that are part of the interview might help some speakers to feel more at ease, such as dissimulating the recorder or learning the questionnaire by heart in order to avoid reading it. In that line, active listening is a key skill to adapt oneself to the speaker while guiding him/her through the interview structure, to change the topic when needed but also to learn from them (Labov 1984: 40). In this study, abundant personal narratives arose during the interviews and some of them became quite intimate, despite the non-intrusive approach in the questionnaire design. Since sexuality is connected with emotional life, the couple, the family, and personal

desires, the most important was to remain aware of how the speakers received the different questions, in order to lead the interview sensitively and ethically, without forcing them into disagreeable topics just for the sake of the data. Consequently, this led to an uneven amount of data or even to the total absence of some concepts in a few interviews, which is also very revealing.

The ethical protocol included information and consent forms granting data protection. They were signed by the speakers and the interviewers and kept by both parties. Moreover the transcriptions were anonymized, which will be explained in the next section.

5 Transcription

The corpus has a total of 60 h of interviews (average 1 h and 5 min per audio file; minimum: 46 min; maximum: 1 h and 27 min). The only reportable incidents were that one of the participants could not finish the interview because she had to leave early, but it was kept as part of the corpus due to the abundance of data it already presented; and two interviews were discarded due to bad audio quality, and to redundant informant selection.

Transcription is a common means for representing spoken language -albeit partially (Thompson 2005: 73; Cook 1995)—and decisions about it are crucial: the output will constitute the corpus on which most types of searches and studies will be applied (Kirk and Andersen 2016: 292). A simplified orthographic method, based on PRESEEA (2003), was used by transcribers, coherent with the goals of the study (semantic variation).¹³ The aim was to achieve a sufficiently informative yet readable transcription method (Thompson 2005: 74; Tagliamonte 2006: 54; Edwards 1995: 24) that could be easily completed with other linguistic layers, if needed. Punctuation signs were used as markup, in order to give information about pauses, intonations and other extralinguistic elements (see Table 3).^{14,15}

Each transcription is divided in speech turns, which are preceded by the code of the speaker or the interviewer. The code is composed of letters and numbers unique to each speaker, which are used as an identifier of all the documents related to the speaker (audio file, transcription, consent forms, and notes).

Since the transcriptions are thought of as a corpus that can be shared for research purposes, the text files have been anonymized for data protection. Names of persons and places and other personal information (dates, events...) that could help identify the speaker are replaced by a generic filler in square brackets (e.g. [name], [place]...).

Various transcribers participated in the first transcription of around half of the corpus, although verification and anonymization was done by the responsible researcher, in order to achieve a more unified result. In total, the transcription files were revised and corrected three times.

¹³ Initially, the transcriptions were not aligned to the audio files, which is a shortcoming that can be solved in the future, by merging both into a software.

¹⁴ Regular Spanish writing uses opening and closing question and exclamation marks.

¹⁵ The transcription uses conventional writing of interjections, onomatopoeias and other “semi-lexical” elements (Andersen 2016: 324) following the RAE recommendations.

Table 3 Markup and transcription norms

Element in the audio file	Transcription norm
Speech turn	New line. Speaker code + semicolon
Interrogative intonation	Closing question mark
exclamatory intonation	Closing exclamation mark
Pause	Slash
Suspended sentence	Suspension points
Overlaps	Square brackets on each turn
Reported speech	Quotation marks
Other sounds	In square brackets [laughter], [click], [clearing throat]...
Onomatopoeias, interjections, etc.	Conventional spelling
Unintelligible fragment	Double brackets

6 Overview and evaluation of the corpus

The collected corpus *MadSex* has more than one million words. As explained in the introduction, the name is an acronym for *Madrileanian spoken corpus of sexuality* (in Spanish, *Corpus madrileño oral de la sexualidad*). Currently, it is available on request, after granting exclusive scientific use of the data, through the website *corpor@louvain*,¹⁶ which gathers corpora collected within the Université catholique de Louvain. In the future, it will be queryable with a password access. Comments and social information useful for sociolinguistic studies (age, level of education, gender, and district) are associated with each text file. As opposed to smaller communities where identification might be easier through the combination of all the social information of the speakers (Childs et al. 2011), in a big city like Madrid each social profile could point to thousands of persons.

In the end, the design of *MadSex* results in a corpus that is different from other corpora in various aspects. It is a sociolinguistic corpus designed for studying semantic variation, which is a field that is yet to be fully exploited due to the methodological problems of the discipline in dealing with variants that carry differences of meaning (Hasan 2009; Robinson 2010; Kristiansen and Dirven 2008; Pizarro Pedraza 2016b). The parallel structure of the interviews, makes it easier to analyze the answers of the 54 speakers to the same questions, and helps in the extraction of variables and variants, by providing contexts for disambiguation. Moreover, it is the first Spanish sociolinguistic corpus dedicated to sexuality, which presents an abundant and varied amount of sexual concepts with the social information of the speakers.

Due to this specificity, some studies have used *MadSex* for its primary goal of studying aspects of linguistic taboo, such as: onomasiological variants of body parts and physiological concepts (Pizarro Pedraza 2018a) metapragmatic euphemistic

¹⁶ <http://corpora.fltr.ucl.ac.be:8080/corpus/madsex> (the URL is provisional).

formulae (Pizarro Pedraza 2018c), profiling of the agents of the verb concept ‘to abort’ (Pizarro Pedraza and De Cock 2018) or the comprehension of sexual metaphors in L2 (De Cock and Suñer 2018). The variety of results found in these different studies shows the versatility of *MadSex* as a corpus and brings about some features worth mentioning. For instance, the corpus is very rich in indirect expressions for sexual concepts (such as metaphors and metonymies), as shown in Pizarro Pedraza (2013) and De Cock and Suñer (2018). Other studies have also used it as a spoken corpus of Madrilénian Spanish for the study of other linguistic phenomena, such as the pragmeme of expanded polar answers (Pizarro Pedraza 2016a) or insubordinate clauses (Gras and Sansiñena Pascual 2017). Therefore, it is also a good resource to study other phenomena besides semantic variation, which was its original purpose, such as syntax and discourse.

7 Conclusion

This paper has presented the corpus *MadSex* with two main goals: to share the methodological solutions for replication purposes, and to explain the features of the corpus, for possible queries. The methodology developed with *MadSex* based on indirect elicitation can be useful for any sociolinguistic study of semantic variation in speech. In that sense, it provides a step-by-step reasoned method for corpus collection, from the design to the transcription phase. Since the study of semantic variation is yet to be developed in Sociolinguistics, the case of *MadSex* could guide anyone interested in collecting their own corpus on particular topics, which would only require some adaptation of the methods for specific contents.

Based on past requests of advice, it will certainly be useful for those who would like to deal with sensitive topics (sex, death, migration...). In that case, it is worth underlining again that the questionnaire needs to be adapted to the sociocultural norms and context of the speech community under research. Moreover, *MadSex* can be used as a database for studies on conceptualization of sexuality, linguistic taboo or language and sexuality.

Appendix 1: Target concepts per module and question

Conceptual area	Target concepts	Module	Question
Body parts	Sexual organs (male, female)	Media	1, 2, 4, 5
		Education	20, 21, 22, 23, 24, 25
	Other erogenous zones	Media	1, 2, 4, 5
		Education	20, 21, 22, 23, 24, 25
	Nakedness	Media	1, 2, 4, 5
		Social changes	10
		Everyday life	17

Conceptual area	Target concepts	Module	Question
Physiological processes	Menstruation	Education	20, 21, 22, 23, 24, 25
	Ejaculation	Education	20, 21, 22, 23, 24, 25
Sexual/Erotic practices	(Have) sexual intercourse	Media	1, 3, 4, 5
		Social changes	7, 8, 10, 14
		Everyday life	15, 16, 18, 19
		Education	20, 21, 22, 23, 24, 25
		Laws	27, 32
	Oral, anal sex	Social changes	10, 12, 14
		Everyday life	19
		Laws	32
	Masturbation	Education	20, 21, 22, 23, 24, 25
	Other (orgies, three-some, S&M, etc.)	Everyday life	19
	Pleasure	Social changes	12, 13, 14
		Everyday life	15, 16, 19
	Arousal	Education social changes	20, 21, 22, 23, 24, 25
			12, 13, 14
	Desire (lack of desire)	Everyday life	15, 17, 18, 19
		Social changes	12, 13, 14
		Everyday life	15, 18, 19
	Virginity	Social changes	7, 8
		Education	20, 22, 23, 25
Other aspects of sexual relations/effects	Pregnancy	Education	20, 21, 22, 23, 24, 25
		Laws	27, 28, 29
	STD	Education	20, 21, 22, 23, 24, 25
		Laws	27, 28
	Contraceptive methods	Education	20, 21, 22, 23, 24, 25
		Laws	27, 28, 29
	Abortion	Education	20, 21, 22, 23, 24, 25
		Laws	27, 28, 29
Personal qualities	To be sexually attractive (man, woman)	Media	1, 2, 4
		Everyday life	15, 17
	To be sexually active, ardent (man, woman)	Media	4
		Social changes	7, 8, 12, 13, 14
		Everyday life	15
	To be promiscuous (man, woman)	Media	3
		Social changes	7, 8, 12, 13
		Everyday life	15
		Laws	27, 28

Conceptual area	Target concepts	Module	Question
Prostitution	Prostitute (male, female)	Social changes	10, 11, 12, 13, 14
		Laws	31, 32, 33, 34
	To prostitute oneself	Social changes	10, 11, 12, 13, 14
		Laws	31, 32, 33, 34
	Brothel	Social changes	11, 12, 13, 14
		Laws	31, 32, 33, 34
	client	Social changes	10, 11, 12, 13, 14
		Laws	31, 32, 33, 34
Options and sexual categories	Heterosexuality	Social changes	9
		Everyday life	19
		Laws	26
	Homosexuality	Social changes	9
		Everyday life	19
		Laws	26
	Bisexuality	Social changes	9
		Laws	19
	Transsexuality	Social changes	10
		Laws	30

Appendix 2: Interview questions sample. Module 4: questions about sexual education

20. Let's talk about something else now. We have talked before about that moment when children start asking questions when they see sex scenes on television. A little later, they will start to develop, their body will change and they will become teenagers. When do you think you should start explaining these issues to your children?
21. Have you had to explain this to your children/little brothers...? If so, how was it?
22. The body changes for a boy and a girl are not the same. Do you think that they should be explained differently? Is there anything that needs to be explained more to a girl? And to a boy?
23. What are the most important things you think they should know? A boy? A girl?
24. In your case, did they explain anything to you? Who was it? Do you remember?
25. Have you ever encountered a situation in your life for which you would have liked to have more information? For example, with the body changes in adolescence?

References

- Allan, K. (Ed.). (2018). *The handbook of taboo words and language*. Oxford: Oxford University Press.
- Allan, K., & Burridge, K. (1991). *Euphemism and dysphemism. Language used as shield and weapon*. New York: Oxford University Press.
- Allan, K., & Burridge, K. (2006). *Forbidden words*. New York: Cambridge University Press.
- Andersen, G. (2016). Semi-lexical features in corpus transcription. Consistency, comparability, standardisation. *International Journal of Corpus Linguistics*, 21(3), 323–347.
- Arroyo, B., & Luis, J. (2005). *Sociolingüística del español. Desarrollos y perspectivas en el estudio de la lengua española en contexto social*. Madrid: Cátedra.
- Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.
- Calvo Shadid, A. (2009). *Análisis Sociolingüístico sobre el tabú sexual en el español de Costa Rica*. Unpublished PhD thesis. Bergen: University of Bergen.
- Campoy, H., Manuel, J., & Almeida, M. (2005). *Metodología de la investigación sociolingüística*. Málaga: Comares.
- Casas Gómez, M. (1986). *La Interdicción lingüística. mecanismos del eufemismo y disfemismo*. Cádiz: Servicio de publicaciones de la Universidad de Cádiz.
- Casas Gómez, M. (2009). Towards a new approach to the linguistic definition of euphemism. *Language Sciences*, 31(6), 725–739.
- Cela, C. J. (1989). *Diccionario secreto*. Madrid: Alianza.
- Childs, B., Gerard, V. H., & Jennifer, T. (2011). Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, 7(1), 163–180.
- Cook, G. (1995). Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer* (pp. 35–53). New York: Longman.
- Crespo Fernández, E. (2007). *El Eufemismo y el disfemismo. procesos de manipulación del tabú en el lenguaje literario inglés*. Alicante: Publicaciones Universidad de Alicante.
- Crespo Fernández, E. (2015). *Sex in language. Euphemistic and dysphemistic metaphors in internet forums*. London: Bloomsbury.
- Danbolt Drange, E. -M. (1997). *La Mujer y el tabú. un análisis sociolingüístico del tabú en el lenguaje femenino de Viña Del Mar*. Unpublished M.A. Dissertation. Bergen: University of Bergen.
- De Cock, B., & Suñer, F. (2018). The Influence of conceptual differences on processing taboo metaphors in the foreign Language. In A. Pizarro Pedraza (Ed.), *Linguistic taboo revisited. Novel insights from cognitive perspectives* (pp. 201–222). Berlin: Mouton De Gruyter.
- Eckert, P. (2005). *Variation, convention, and social meaning*. Oakland: Annual Meeting of the Linguistic Society of America.
- Edwards, J. A. (1995). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on computer* (pp. 19–34). New York: Longman Group Limited.
- Foucault, M. (1976). *Histoire de La Sexualité. La Volonté de Savoir* (Vol. 1). Paris: Éditions Gallimard.
- Galli de Paratesi, N. (1964). Semantica dell' eufemismo. l'eufemismo e la repressione verbale con esempi tratti dall' italiano contemporaneo. *Pubblicazioni Della Facoltà Di Lettere E Filosofia* XV(1). Torino: G. Giappichelli Editore.
- García Mouton, P. (2015). "Dialectología y geografía lingüística". *Enciclopedia Lingüística Hispánica*. Abingdon: Routledge.
- Geeraerts, D. (2006). *Words and other wonders*. Berlin: Mouton de Gruyter.
- Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The structure of lexical variation*. Berlin: Mouton De Gruyter.
- Geeraerts, D., & Kristiansen, G. (2014). Cognitive linguistics and language variation. In J. Littlemore & J. Taylor (Eds.), *Bloomsbury companion to cognitive linguistics*. London: Bloomsbury.
- Geeraerts, D., Kristiansen, G., & Peirsman, Y. (Eds.). (2010). *Advances in cognitive sociolinguistics. Advances in cognitive sociolinguistics*. Berlin: De Gruyter Mouton.
- Glynn, D. (2014). Polysemy and synonymy. Cognitive theory and corpus method. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 7–38). Amsterdam: John Benjamins Publishing Company.

- Gras, P., & Sansiñena, M. S. (2017). Exclamatives in the functional typology of insubordination: Evidence from complement insubordinate constructions in Spanish. *Journal of Pragmatics*, 115, 21–36.
- Hasan, R. (2009). *Semantic variation: Meaning in society and in sociolinguistics*. London: Equinox Publishing Ltd.
- Heylen, K., José, T., & Dirk, G. (2008). Methodological issues in corpus-based cognitive linguistics. In G. Kristiansen & R. Dirven (Eds.), *Cognitive sociolinguistics: Language variation, cultural models, social systems* (pp. 91–128). Berlin: Mouton de Gruyter.
- Janicki, K. (2006). *Language misconceived. Arguing for applied cognitive sociolinguistics*. Mahwah: Lawrence Erlbaum Associates.
- Jay, T. (2000). *Why we curse*. Philadelphia: John Benjamins.
- Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2), 153–161.
- Kirk, J. M., & Andersen, G. (2016). Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics*, 21(3), 291–298.
- Kristiansen, G., & Dirven, R. (2008). *Cognitive Sociolinguistics. Language variation, cultural models, social systems*. Berlin: De Gruyter Mouton.
- Labov, W. (1971). Some principles of linguistic methodology. *Language in Society*, 1(1), 97–120.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of ePennsylvania Press.
- Labov, W. (1982). Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language in Society*, 11, 165–202.
- Labov, W. (1984). Field methods of the project on linguistic change and variation. In J. Baugh & J. Sherzer (Eds.), *Language in use: Readings in sociolinguistics*. Prentice Hall: Englewood Cliffs.
- Leech, G., Myers, G., & Thomas, J. (Eds.). (1995). *Spoken English on computer*. New York: Longman Gr.
- Litosseliti, L. (2010). *Research methods in linguistics*. London: Continuum International Publishing Group.
- López Morales, H. (2005). Sociolingüística del tabú. *Interlingüística*, 16, 7–20.
- López Morales, H. (2009). El estudio de la variación lingüística. In M. V. Camacho Taboada, J. J. Rodríguez Toro, & J. Santana Marrero (Eds.), *Estudios en lengua española : descripción, variación y uso. Homenaje a Humberto López Morales* (pp. 9–34). Madrid: Iberoamericana Vervuert.
- Mancera, C., María, A., Martos, I. M., & García, F. P. (2015). *Patrones sociolingüísticos de Madrid*. Oxford: Peter Lang.
- Martínez Valdeza, P. (1995). *El Tabú Lingüístico: Estudio Sociolingüístico de Las Palmas de Gran Canaria*. Unpublished PhD thesis. Las Palmas de Gran Canaria: Universidad de Las Palmas.
- Montero Cartelle, E. (1996). Pene: Eufemismo y disfemismo en el gallego medieval. *Verba*, 23, 307–336.
- Montero Cartelle, E. (2010). La Sexualidad medieval en sus manifestaciones lingüísticas: Pecado, delito y algo más. *Clio & Crimen*, 7, 41–58.
- Pizarro Pedraza, A. (2013). *Tabú y eufemismo en la ciudad de Madrid. Estudio sociolingüístico-cognitivo de los conceptos sexuales*. Unpublished PhD thesis. Madrid: University Complutense of Madrid.
- Pizarro Pedraza, A. (2016a). Pragmemes in the sociolinguistic interview: A case study on expanded polar answers. In K. Allan, A. Capone, & I. Kecskes (Eds.), *Pragmemes and theories of language use* (pp. 351–374). Cham: Springer.
- Pizarro Pedraza, A. (2016b). Variación semántica y significado social: hacia una sociolingüística cognitiva de la Tercera Ola. *Dicenda—Cuadernos de Filología Hispánica*, 34, 311–338.
- Pizarro Pedraza, A. (2018a). Calling things by their name: Exploring the social meanings in the preference for sexual (in)direct construals. In A. Pizarro Pedraza (Ed.), *Linguistic taboo revisited. Novel insights from cognitive perspectives* (pp. 245–268). Berlin: Mouton De Gruyter.
- Pizarro Pedraza, A. (Ed.). (2018b). *Linguistic taboo revisited. Novel insights from cognitive perspectives*. Berlin: Mouton De Gruyter.
- Pizarro Pedraza, A. (2018c). Pardon my Spanish: Attenuation of taboo through metapragmatic euphemistic formulae. In E. C. Fernández (Ed.), *Taboo in discourse: Studies on attenuation and offence in communication* (pp. 181–208). Bern: Peter Lang.
- Pizarro Pedraza, A., & De Cock, B. (2018). Taboo effects at the syntactic level: reducing agentivity as a euphemistic strategy. *Pragmatics: Quarterly Publication of the International Pragmatics Association*, 28(1), 110–135.

- Pop, S. (1950). *La dialectologie : Aperçu historique et méthodes d'enquêtes linguistiques*. Louvain: J. Duculot.
- PRESEEA. (2003). *Metodología del Proyecto para el estudio sociolingüístico del español de España y de América*. <http://preseea.linguas.net/Portals/0/Metodologia/METODOLOG%C3%8DA%20PRESEEA.pdf>. Accessed on April 20, 2018.
- Pütz, M., Justyna, A. R., & Monika, R. (Eds.). (2014). *Cognitive sociolinguistics: social and cultural variation in cognition and language use*. Vol. 59. *Benjamins current topics*. Amsterdam: John Benjamins Publishing Company.
- Pütz, M., Robinson, J. A., & Reif, M. (2012). The emergence of cognitive sociolinguistics. An introduction. *Special Issue of Review of Cognitive Linguistics*, 10(2), 241–263.
- Rabanales, A. (2004). La encuesta como técnica dialectológica. *Onomázein*, 9(1), 75–93.
- Robinson, J. (2010). Awesome insights into semantic variation. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in cognitive sociolinguistics*. Berlin: Mouton de Gruyter.
- Rodríguez González, F. (2011). *Diccionario del sexo y el erotismo*. Madrid: Alianza Editorial.
- Sankoff, D. (2004). Problems of Representativeness. In U. Ammon, N. Dittmar, K. J. Mattheier, & P. Trudgill (Eds.), *Sociolinguistics. Soziolinguistik. An International Handbook of the Science of Language and Society. Ein Internationales Handbuch Zur Wissenschaft von Sprache Und Gesellschaft* (pp. 998–1003). Berlin: Walter de Gruyter.
- Schilling-Estes, N. (2008). Stylistic variation and the sociolinguistic interview: A reconsideration. In R. Monroy & A. Sánchez (Eds.), *25 Años de lingüística aplicada en España: hitos y retos*. Murcia: Servicio de Publicaciones de la Universidad de Murcia.
- Steensig, J., & Trine, H. (2013). When 'yes' is not enough—As an answer to a yes/no question. In B. S. Reed & G. Raymond (Eds.), *Units of talk—Units of action* (pp. 207–242). Amsterdam: John Benjamins Publishing Company.
- Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
- Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 59–70). Oxford: Oxbow book.
- Uría Varela, J. (1997). *Tabú y eufemismo en latín. Classical and Byzantine Monographs*. Amsterdam: A. M. Hakkert.
- Wolfram, W. (1993). Ethical considerations in language awareness programs. *Issues in Applied Linguistics*, 4, 225–255.
- Wolfram, W. (2011). Fieldwork methods in language variation. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE handbook of sociolinguistics* (pp. 296–312). Thousand Oaks: Sage Publications.
- World Health Organization (WHO). (2006). *Defining sexual health*. Report of a technical consultation on sexual health. Geneva: WHO. http://www.who.int/reproductivehealth/publications/sexual_health/defining_sexual_health.pdf. Accessed 21 Nov 2018.