Criteria for the Construction of a Corpus for a Mexican Spanish Dictionary of Sexuality

Alfonso Medina, Gerardo Sierra
Instituto de Ingeniería
Universidad Nacional Autónoma de México
Apartado Postal 70-472
04510 Delegación Coyoacán, DF
MEXICO
{amedinau, gsierram}@iingen.unam.mx

Abstract
This paper describes our criteria to build an electronic Corpus for a Mexican Spanish Dictionary of Sexuality. The Corpus will be used, among other things, for the extraction of the field's terminology. Needless to say, automatic extraction of terms presupposes a representative Corpus. Thus, we sketch our strategies to assure such representativity. In short, we address the need for a balanced Corpus; that is, one containing texts belonging to all thematic areas of the field. Then we describe the sort of texts that must be included. Also, we discuss our approach to deal with the important sociolinguistic dimension. Finally, we examine the need of an articulated set of concepts; that is, the conceptual system used by sexologists, sexual educators and other specialists (physicians, psychologists, social workers, etc.). In this way, we expect our Dictionary of Sexuality to have a set of entries which will guarantee the broadest possible coverage of this specialty field's knowledge.

1. Introduction

Researchers and educators rely upon knowledge representation tools that provide consensus, integration and clarity of concepts and definitions within their specialty fields. One very old type of this kind of tool is the dictionary. And one field of distinct complexity, especially in countries like Mexico, is that of Sexuality Studies and Sexual Education. Perhaps the main problem for a dictionary for this field is the dispersion of terminology. Although some scattered glossaries and vocabularies do exist, there is no single reference work in Spanish which provides the terminological consensus needed for Sexual Education; in fact, not even for communication among Sexology specialists. Dictionaries in book form do exist, but they are not meant either for Sexual Education or for sexological research consensus. Some are marketing products intended to promote the image of media celebrities.[1] Others are encyclopedic works portraying distant realities and therefore failing to fulfill the needs of a Mexican audience.[2]

A dictionary tailored to the needs of such an audience presupposes, on the one hand, some articulated set of concepts and definitions representing the up to date sexological knowledge related to important educational issues like Health and Human Rights: the conceptual system deemed by experts to be most relevant in the field of Sexology. On the other hand, the set of terms used by regular Mexican people to refer to their sexual experiences should also be considered. Whether linguistic communities are informed or not about sexological facts, the reality is that most people do use their language to talk, secretly

or not, about sexuality. Indeed, the sociolinguistic dimension is a very complex issue because regional and social variation is inevitable. Nevertheless, its impact on Sexuality Studies and Sexual Education is important. Indeed, new word documentation would help educators monitor student attitudes toward sexual issues and favor overall communication among specialists.

The present paper describes our attempt to deal with the lack of such dictionary. Namely, we describe our criteria to build an electronic Corpus for a Mexican Spanish Dictionary of Sexuality. Once built, the Corpus will be used for terminology extraction. However, identifying the terminology of a field is one very specific task that fits within a larger program of compiling, publishing and distributing a dictionary. For the purpose of education the interest in Mexico for carrying this out is considerable. Mexico's Sexual Education community is willing to offer their expertise and validate the resulting material. And other prestigious research and educational institutions are committed to the advancement and accomplishment of the larger project. Thus, El Colegio de México, a research institute and graduate school which has pioneered lexicographical and terminographical investigations in Mexico (Lara et al., 1979; Lara, 1990; Pozzi, 1996), has an experienced team of terminologists committed to the appropriate formulation of the definitions of the extracted terms. Also, the Institute of Engineering of our National Autonomous University is developing an electronic dictionary[3] capable of performing complex multimedia functions and suitable for massive distribution (Sierra et al., 2002).

2. Criteria for Building a Corpus

Ideally, terminologists rely on the opinion of experts to determine terms and to define them. However, expert participation in a terminological project can be very expensive. Hence, electronic corpora (seen as written representations of expert opinion) are important alternatives to intensive participation by the field's specialists. Nevertheless, expert opinion remains for this project an important secondary source of information.

Terminology extraction from corpora can be accomplished automatically by means of statistical methods. For example, one method of identifying a specialized field's terminology is to compare the distribution of words from a general purpose corpus with that of a specialized corpus. Selection of those word types with higher frequencies in the latter yields a set of potential terms for the field.

However, research proper begins with the selection of the sample texts that should belong to the Corpus. First, the relevant thematic areas of the field in question must be identified. Given the different perspectives from which Sexuality can be viewed, there is a lack of consensus among specialists. Some are medical specialists, some are psychologists, some are social workers, some are Human Rights militants, etc. The thematic structure of the field varies from one perspective to the other. We solved this problem by referring to an international institution considered a scientific authority by all. In this manner, we obtained from the Kinsey Institute the thematic areas of their Library classification system.

Second, the key texts of the field must be included. Although widely read works are originally written in some language other than Spanish (mostly English) or in some other dialect (mostly Argentinean and Peninsular Spanish), many have greatly influenced the training of researchers and educators in Mexico. Thus, it is not surprising that many of the

Mexican terms for sexual affairs are in fact loans from other languages or other terminologies from other dialectal systems.

Third, we will include the written works of renowned Mexican researchers and educators. The Sexual Education and Sexological Research community in Mexico is small but very productive. The Mexican Federation of Sexual Education and Sexology (Federación Mexicana de Educación Sexual y Sexología, FEMESS) has members of very diverse disciplines, among them: Medicine, Sexual Education, Sexual Health, Culture and Ideology of Sexuality and Politics of Sexuality. Throughout the years its members have built their own frameworks and teaching methods. Much of their written work has not been published, so we have requested that they provide us with their notes and other unpublished materials.

Regarding sociolinguistic research, the Corpus will contain results from a poll designed to obtain the terms used by people to refer to sexual matters (anatomic parts, sexual stereotypes, contraceptives, sexual transmissible diseases, derogatory words with sexual connotations, etc).[4] Due to the complexity of the sociolinguistic dimension, further steps can be added at a later stage to this strategy. A reasonable goal for this proposal is to apply such inquiry to a variety of small groups of society members (workers, university students, adults over 40, etc.).[5]

Needless to say, sociolinguistic research is radically different from the compilation of experts' texts. Our Corpus will contain both sources of information; so, both types of discourse will be subjected to the same statistical investigations. However, some aspects of the poll may not allow us to elicit appropriate contexts of term use (consider, for instance, one word or enumerative answers to certain poll questions). Nevertheless, we can obtain important quantitative data by including poll results (in a specific format) in the Corpus. And we do expect the poll to yield, after all, some data about the terms' uses and typical contexts.

Another strategy we decided to implement was to collect electronic material with sexual connotations available in the Internet (chats, groups, lists). Although questions may be raised about how representative this is of Mexican Speech, it does represent a sample of spontaneous discourse among a segment of young people.

In order to illustrate how the final dictionary may profit from this variety of information sources, we have provided in Table 1 some terms related to the notion of bisexuality. The first two are representative of the colloquial register and the last one stands for the formal register. The last column exhibits some typical contexts in which the terms occur. Notice how the contexts of use of the colloquial terms resemble spoken conversation. We expect our sociolinguistic research to yield this kind of information: colloquial terms and their contexts will come mostly from the polls.

Also, examine the different meanings displayed for the formal term. Different text sources, representing distinct schools of thought, may conceive phenomena differently. Their points of view may be reflected by listing the definitions which best describe them. Thus, whether being bisexual is a matter of preference, attraction or biology will not be a matter of discussion; these points of view will simply be presented as the possible senses of the term. Meanwhile, a truly prescriptive dictionary would present one point of view as the correct one.

Terms	Definitions	Uses and contexts
Bi	(s. y adj.; **coloquial**) →Bisexual (apócope). [..].	*Te dije que aquí sólo viene gente **bi**.*
Bicicleta	(s. y adj.; **coloquial**). →Bisexual. [..].	*No te imaginas, como **bicicleta** que era, bien que le pedaleaba tanto a Rosita como a Jorge.*
Bisexual	(s. y adj.) **a) (formal)** →Preferencia sexual de las personas que no sienten mayor atracción por personas de algún sexo o género en particular [..]. **b) (formal)** Atracción sexual por personas de ambos sexos y/o que tiene relaciones sexuales indistintamente con ellas [..] ⋮ **z) (coloquial, peyorativo)** →Homosexual no asumido [..].	*El **bisexual** masculino propone y representa el goce como una forma de ascesis.* *El paciente se asumió **bisexual** al principio de la consulta.* *Los pinches **bisexuales** son locas de clóset.*

Table 1: Illustration of terms and their contexts

Lastly, technical terms may have become part of people's common vocabulary. And the typical uses of those terms may reflect the people's negative attitude towards certain sexual matters such as bisexuality (see sense 'z)' of third term in table 1). We have in fact noticed in preliminary polls that well over half the terms we have elicited are classified as negative by the very same people polled. The marking of terms as pejorative will constitute a portrait of regular people's attitudes towards sexuality. Hence, it would be desirable to improve the sociolinguistic poll and to apply it periodically in order to reflect changes of attitude, and the emergence of neologisms. This does not exclude the addition of expert materials to the Corpus.

In short, this project can grow indefinitely. Later versions of the dictionary may even include encyclopedic information. But in this paper it would be very premature to sketch more than what has been dealt with. To summarize, the following points are the key ideas mentioned above:

1. In the face of lack of consensus among specialists, regarding the thematic areas of their field, find a neutral authority respected by all specialists.
2. Include in the Corpus translations of widely read books in the field. It is important to select those translations read by the experts; especially, those used in the training of researchers and educators. Also, important works from other Spanish speaking countries should be considered.

3. Include in the Corpus the work of renowned experts in the field written in Spanish, published or not. A balanced representation of the local community deserves to appear in the Corpus.
4. Design a strategy for obtaining sociolinguistic data. One way to accomplish this is by designing and applying a poll in order to obtain lexical information. A possible complement is the inclusion of Internet materials in the Corpus.

3. Closing Remarks

The gathering of representative discourse of the field is the first step towards constructing the electronic Corpus. Some materials are already available in electronic form. Most of them are being scanned. The target size for the Corpus is two million tokens. This should be an adequate size for obtaining a terminology of 2,500-5,000 types. The Corpus will be available on the Internet for researchers and educators to use as a reference tool. Term contexts and definitions found there will be consulted for the writing of the target definitions of the final dictionary. A concordancing tool will naturally be provided for this and any general purpose exploration of the Corpus.

Once a set of terms has been extracted, it will be necessary to determine how each term relates to each other. More importantly, grouping them according to their lexical relations and building with them a hierarchical conceptual structure will give an idea of what is missing. Text sampling errors can be repaired and missing terms can be detected by building such a structure. Then, specific research can be conducted to fill the empty spaces in the conceptual system represented by the terms. The set of terms resulting from this will also be available on the Internet. The idea is to build a data base that can host them, their definitions and other relevant lexicographical data.

Finally, given the nature of language, it is important to notice that both the Corpus and the terminology extracted are expected to change with time. That is why the writing of the definitions and of future versions of the dictionary will require that both remain in the Internet for an indefinite length of time.

Endnotes

[1]For example, radio and television commentator Anabel Ochoa's *La palabra común, Diccionario erótico México-España*, Mexico, Colofón, 2002.

[2]For instance, the Chilean Osvaldo Quijada's *Diccionario integrado de sexología*, Madrid, Alhambra, 1983.

[3]Electronic dictionaries offer many advantages. For instance, very sophisticated searching possibilities of semasiological and onomasiological types (Sierra & McNaught, 2000). Also, multimedia and game components make them very attractive.

[4]Labov's conversation modules are a desirable method to be considered for this poll (Labov, 1984). Also, research experience previously conducted in Mexico to collect lexical items from children should be taken into account (López Chávez, 1993).

[5]Naturally, this can only be a first step towards a wider investigation across the regional and social boundaries of the country. Moreover, this does not exclude the possible inclusion in the corpus of other sexual research polls previously conducted in the country.

References

Biber, D, Conrad, S. and Reppen, R. 1998. *Corpus Linguistics*. Cambridge: Cambridge UP.

Cabré, M. T. 1993. *La terminología. Teoría, metodología, aplicaciones.* Barcelona: Antártica.

Cabré, M. T. 2000. 'Elements for a Theory of Terminology: Towards an Alternative Paradigm.' Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 6:1, pp. 35-57.

Labov, W. 1984. 'Field Methods of the Project on Linguistic Change and Variation.' In: Baugh, J. and Sherzer, J. (eds.) *Language in Use: Readings in Sociolinguistics*. Englewood Cliffs, NJ: Prentice-Hall, pp. 28-53.

Lara, L. F., Ham Chande, R. and García Hidalgo, M. I. 1979. *Investigaciones lingüísticas en lexicografía.* Mexico: El Colegio de México.

Lara, L. F. 1990. *Dimensiones de la lexicografía. A propósito del Diccionario del Español de México.* Mexico: El Colegio de México.

López Chávez, J. and Meza Canales, R. M. 1993. *Léxico disponible de preescolares mexicanos*. Mexico: Facultad de Filosofía y Letras, UNAM.

McEnery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh UP.

Medina, A. 2001. 'Diccionario de terminología de la sexualidad en México', IV Congreso Nacional de Educación Sexual y Sexología. Congress organized by the Mexican Federation of Sexual Education. Veracruz, Veracruz, Mexico, March, 2001.

Medina, A. and Sierra, G. 2003. 'Construcción de un sistema lexicográfico como apoyo a la educación sexual', V Congreso Nacional de Educación Sexual y Sexología. Congress organized by the Mexican Federation of Sexual Education. Morelia, Michoacán, Mexico, August, 2003.

Pozzi, M. 1996. 'BTMEX: A flexible Terminological Data-Management System.' Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3:1, pp. 111-124.

Sager, J. C. 1990. *A Practical Course in Terminology Processing*. Amsterdam/ Filadelfia: John Benjamins.

Sierra, G. and McNaught, J. 2000. 'Design of an Onomasiological Search System: A Concept-Oriented Tool for Terminology.' Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 6:1, pp. 1-34.

Sierra, G., Medina, A., Moreno, M., Garduño, G. and Castillo, G. 2002. 'Diccionario básico de sexualidad'. Paper presented by Marlene Moreno in the Congreso Latinoamericano de Multimedieros Universitarios 2002. Congress organized by the University Association for Multimedia, National Autonomous University of Mexico, Mexico City, November, 2002.