SOLC: An Ortological Database Server to Teach Catalan and to Improve its Knowledge

Lluís de Yzaguirre i Maura, Francesca Salvà Cerdà

Universitat Pompeu Fabra, Barcelona

E-mail: **de_yza@upf.es**, **francesca.salva@iula.upf.es**

ABSTRACT

This is the presentation of the Catalan automated orthoepical dictionary destined to provide an answer to the needs of many linguistic and non-linguistic professionals facing a very plural (and, in many cases, not well defined in relation to the normatives) oral standard. It is mostly focused on the mass media –on acount of its major role in the dissemination of a standard language model– since it provides solutions for radio and television announcers, orthology specialists, dubbing actors, multimedia producers, teachers, public speakers or linguistic engineers. The objective is to turn into rules the formulations of the Proposta d'Estàndard Oral de la Llengua Catalana de l'Institut d'Estudis Catalans [the Catalan Language Oral Standard Proposal of the Institute for Catalan Studies, IEC], which is the proper and recognized authority by all universities and professionals. These rules will make this proposition more systematic, although with some interpretative gaps for the time being.

1. CONTEXTUALITZATION

We start from a situation of centuries of linguistic persecution, especially during Franco's dictatorship when Spanish became an imposed language everywhere. After the end of the dictatorship, and beginning in 1978, the objective of the Catalan administrations operating within the strict legal boundaries of the Autonomy Statute and the Spanish Constitution, was to achieve that Catalan language could recover the status of public language it deserved. Persecution and a strong affluence of immigrants, especially in the 1960s, marked both the language as well as its use in society.

There is no doubt that education and mass media are the fundamental axes of the normalization process. In the early 1980s, Catalan became the language of radio and television. In 1983, the Corporació Catalana de Ràdio i Televisió (the Catalan Radio and Television Corporation, CCRTV) was founded, marking the birth of Catalan language public radio and television: television channels TV3 and Canal 33, Catalunya Ràdio and several stations with music, news and different programs, that became the main transmission sources of the standard, common identifying language. This phenomenon of the advent of public mass media being conducted in the Princedom of Catalonia, was joined by the initiatives of private entities in the rest of the linguistic domain, where these media were arriving, like l'Obra Cultural Balear (Associació Voltor) and Acció Cultural del País Valencià.

We must also add to all this, and as a need and complement of the massive use of Catalan in our mass media, the presence of dubbing enterprises that, also guided by the linguistic model of the media, were to adapt the language model to the variation, and, most of all, the registration, that film dubbing supposed.

2. ORTHOEPIC MODEL OF THE PHILOLOGY SECTION OF THE IEC

From this moment on, the debates around the model of the oral language that was to be circulated began, in a confrontation, on one side, of those who favored a more controlled, genuine and general language, and, on the other side, of those who favored a model more familiar with the language used in the streets. It was not until 1990, that the Proposta d'Estàndard Oral de la Llengua Catalana of the

ISBN 1-876346-48-5

IEC appeared, marking a milestone in the issue of authority, yet still unfinished, since it is only a proposal, not systematic enough, very flexible, and with interpretation gaps that are difficult to be applied by professionals (no linguistics).

Social reality has obviously influenced the writing of that proposal and the model circulated by mass media, to which it is almost essentially directed. The distribution of the Catalan territory under different authorities and administrations has fragmented the standard, and the dialect modalities have become difficult to unify, mostly in what is related to the oral language.

3. THE SOLC PROJECT: OBJECTIVES

Today's formulation of the flexible oral standard as a synthesis of the necessary unity of the language and the plurality of the models and situations (geographic, social or stylistic models) leaves a forced decision margin for the professionals of the language (announcers and language editors as well) in which they feel unassisted. They need a consultation tool to clarify many of the doubts that they are faced with.

Furthermore, we do not have a pronunciation model in the normative dictionary (the Diccionari de la Llengua Catalana of the IEC), surely for the same reason that we do not a have a standard phonetic dictionary: the difficulty to cover the whole variation of the oral standard. We only have at this moment the Proposta d'Estàndard Oral of the IEC (Fonètica i Morfologia [Phonetics and Morphology]), which is the base from which we will make our dictionary.

The SOLC project, Servidor Ortològic de la Llengua Catalana [Orthoepic server for the catalan language], is going to be a prescriptive dictionary, since it tries to offer a tool to guide language professionals facing possible doubts on issues of normative phonetics; and that will include terminology and encyclopedic information, since professional doubts are not only limited to the vocabulary included in a lexicographic tool; and open, since one of the characteristics of neology is the inestability of its pronunciation.

The objective is to turn into rules the implicit formulations of the Proposta d'Estàndard Oral of the IEC. It is fundamentally an on-line database that includes the orthographic representation, the phonematic representation, the grammatical category and the phonetical variations coming from the different environments present in the IEC's proposal (on the one hand, proper and formal general forms, and restricted, admissible and informal forms, on the other), and the five large dialect groups from the Roussillon, the Northwest, Valencia, the Balearic Islands and the Central region (rossellonès, nord-occidental, valencià, insular and central).

Starting from the initial phonematic transcription, rules will be applied to allow the generation of all the different forms of the same lemma. This will be applied to the known vocabulary and, when requested by professionals, to other words.

4. THE SOLC PROJECT: STAGES

Both the elaboration and the circulation will be carried out, in principle, through the Internet with an on-line database. The creation of the database will follow a group of structured phases in three stages: internal, restricted, and open.

The internal stage has the objective of creating the first nucleus of transcriptions from a 5,000-20,000-word basic dictionary, formulating the rules, and the revision of their application. Revision will be carried out by specialists from each part of the territory with sufficient knowledge of the variation and the standard. It is convenient that they carry out their work completely immersed in their territorial areas. Data will be automatically generated by the five large geographic varieties, and by the more and less formal variations. The source of the main nucleus of the list of lemmas will be Lluís de Yzaguirre's dissertation on syllabic structure, as well as his rule formulation system.

The objective of the restricted stage is the automatic transcription of the whole vocabulary and the manual revision of an approximately 5% randomized control sample. Professionals will have access, since this stage pretends to benefit from the contributions of those who face

ISBN 1-876346-48-5

significant orthoepic doubts every day, pretends to try to solve their doubts, and pretends to add their contributions to the database.

The public stage represents the continuation of the randomized manual revision, with the follow-up of consultations and doubts by the general public. It will start when the revision of the 5% control sample evidences a minimum level of error.

5. DESCRIPTION VERSUS ORTHOLOGY

The difficulty to systematize the Proposta d'Estàndard Oral de la Llengua Catalana also lies in the absence of an oral corpus that covers the type of language that is going to be analyzed, and that does not allow documenting most of the phenomena being commented.

The Laboratori de Tecnologies Lingüístiques, de l'Institut Universitari de Lingüística Aplicada (Linguistic Technologies Laboratory of the Universitary Institute of Applied Linguistics, LATEL) has begun an oral corpus project that covers these needs, and has put it in the service of the SOLC Project in such a way that, in case of an orthoepic doubt, an automatic consultation can be made. (To date, April 2003, the corpus has 136,000 occurrences, and previsions for the end 2004 are 500,000 occurrences). Corpus material comes most of all from mass media (mostly radio and television from the Princedom, Balearic Islands, Valencia, and also some registered from Ràdio Arrels, in Northern Catalonia) and academic speech.

6. FUTURE PERSPECTIVES

Now that the most difficult part has been carried out (the creation of the rule system and the informatic procedure) we must fill professionals from different specialisms with enthusiasm, so they may come to participate both in the analysis and in the transcription. For that reason, this project will be presented at a philologists' forum on September 2002 at the 13è Col·loqui Internacional de Llengua i Literatura Catalanes [13th International Colloquium on Catalan Language and Literature].

7. CONCLUSION

We are presenting an initiative in which the application of new technologies in the normalization of the language of a minority, besides its intrinsic advantages, brings the benefit of allowing to carry out work in a net that covers the whole territory. It will be useful to train of announcers, orthologists, oral editors, linguistics and to teach catalan.

REFERENCES

[1] Cabré, MT.; De Yzaguirre, Ll. i E. Clua (1999) "Diccionari ortològic català", conferència al Congrés Llengua i Mitjans de Comunicació. LXXV anys de ràdio. 1924-1999, Universitat de Lleida, desembre 1999. Publicada a Creus, I.; Julià, J i S. Romero (eds.) Llengua i mitjans de comunicació, Pagès editors.

[2] Camps, O.; De Yzaguirre, Ll. i A. Matamala (2000) "DOPO, un outil d'analyse orthologique", comunicació presentada a 1. Freiburger Arbeitstagung zur Romanistischen Korpuslinguistik, Freiburg, 6-8 octubre i publicada a a PUSCH, Claus D. / RAIBLE, Wolfgang (Hrsg.) 2002: Romanistische Korpuslinguistik - Korpora und gesprochene Sprache / Romance Corpus Linguistics - Corpora and Spoken Language (= ScriptOralia; 126). Tübingen: Narr. 500 pp. ISBN 3-8233-5436-1.

[3] De Yzaguirre, L.; Camps, O. i A. Farriols (2000) "RETOC: a hypermedia compilation of oral texts", comunicació presentada a l. Freiburger Arbeitstagung zur Romanistischen Korpuslinguistik, Freiburg, 6-8 octubre.

[4] De Yzaguirre, Ll.; Clua, E. i A. Farriols (2000) "Les corpus oraux et l'enseignement de la langue", comunicació presentada a GLAT 2000, Multilingual Communication and Interactivity: The Word and Beyond, a Brest, 11-13 juliol. Publicada a Actes de GLAT 2000, École Nationale Supérieure des Télécommunications de Bretagne, Brest, p. 191-199. ISBN: 2-908849-09-7.

[5] Proposta per a un Estàndard Oral de la Llengua Catalana (I. Fonètica). Institut d'Estudis Catalans, 1990.

[6] Proposta per a un Estàndard Oral de la Llengua Catalana (I. Mofologia). Institut d'Estudis Catalans, 1990.

[7] http://retoc.iula.upf.es

[8] http://retoc.iula.upf.es/solc

[9] http://retoc.iula.upf.es/latel

 ISBN 1-876346-48-5

ISBN 1-876346-48-5