LLUÍS DE YZAGUIRRE (1), de_yza@upf.es
ORIOL CAMPS (2), ocamps.n@catradio.com
ANTONI J. FARRIOLS (1), afarriol@pie.xtec.es

(1)
Institute for Applied Linguistics (IULA)
Universitat Pompeu Fabra

(2)
Unitat d'Assessorament Lingüístic
Catalunya Ràdio

1. Freiburger Arbeitstagung zur romanistischen Korpuslinguistik
October 2000

RETOC: a hypermedia compilation of oral texts

1. Introduction

RETOC (Electronic Repertoire of Catalan Oral Texts) is aimed at providing several types of users with oral material. When RETOC was started there was a lake of an oral reference corpus regarding the Catalan language. At that moment the only current experience (i.e. the Corpus from the Universitat de Barcelona, CUB) showed enough particularities to allow for further new initiatives.

Previous experiences, especially that of CUB, have influenced the way in which RETOC is designed and oriented toward the following goals:

- to gather material according to their interest to several groups, their technical quality, their availability and their reusability;
- to deal with this material with the less human cost;
- to spread this material in a way as general as possible
- to attain a procedure of inclusion and treatment of material in a feasible way so that new researchers can be attracted by offering their own material to the corpus.
- to reuse this material in as many ways as possible

2. Fields of research

RETOC is an initiative from the Language Engineering Research Group at the Institute for Applied Linguistics (Universitat Pompeu Fabra). It is a part of larger research from the same group, which includes two additional projects: SINCRO and DOPO. SINCRO (Sistema Informàtic de Normalització de Corpus per a la Recerca Oral = Computer System of Corpora Standardisation for Oral Research) is an application intended to match oral corpora with the corresponding written text and transform them into hypertexts; this tool is the natural outcome of those software tools produced by Dr. Lluís de Yzaguirre for the CUB when he was one of its members (1990-1998). DOPO (Dignòstic Ortològic Per Ordinador = Orthological Diagnosis by Computer) was started by the IULA-UPF together with UAL (Unitat d'Assessorament Lingüístic = Unit for Linguistic Advice) from the radio station Catalunya Radio. DOPO's main goal is to apply new technologies to orthological analysis, which will be presented in this meeting. Therefore SINCRO (already presented in Brest, 2000) is the tool with which RETOC has been produced and DOPO is the first experience of RETOC exploitation with practical goals.

However these three working fields will foster their own entity in due course, at the moment they all form and indissoluble unity, especially due to the contribution of CR to RETOC: all three projects merge when IULA-UPF and CR sign a collaboration program in which

a) CR provides as many oral materials as reasonably required by RETOC in the best technical conditions and transfer facilities.
b) The UAL is getting involved in the DOPO design
c) The UAL becomes a DOPO user

It has to be emphasised that, although CR contribution to RETOC is likely to be reduced percentually in the future, its original relevance cannot be only measured with regard to the quantity and quality of the set of oral texts being provided. Further, these oral texts assume the RETOC's original philosophy in the best way: general interest, economy of treatment, diffusion and reusability.

3. Users

RETOC is intended to reach the following groups:

- Catalan learners, especially those who come from non speaking Catalan territories;
- Catalan scholars in a philological perspective, especially with regard to dialectology;
- Broadcasters and actors/actresses
- Oral correctors
- Phonotechnologists (from telecommunication engineers to producers of systems of voice synthesis and recognition)
- Phonetists and phonologists

With regard to the two last groups, the quality of voice files with free access, is believed to be poor (22kHz, 8 bites, monoaural). However there will not be any technical drawback (only financial costs) so as to gain access to maximum quality versions (44kHz, 16 bites, monoaural) of matched data.

4. Current materials

Today, in October 2000, RETOC contains more than 14 processing hours (already digitalised) from radio news. Two of these 14 hours have already been segmentally matched and can be accessed to the following URL:

http://retoc.iula.upf.es/entrada.htm.

There are about 20 minutes that are matched at a lexical level and the experimental part from CR contains DOPO hypertexts.

Taking into account the homogenous features of the texts provided by CR, which are initially designed to contribute to DOPO, we allow its detailed presentation in the following communication.

It can also be found radio samples previously provided by CR over the test stage prior to the RETOC definition.

There are also two conferences held at IULA, which proved to be fruitful to search for the possibilities of exploiting these types of texts.

Finally, to offer the chance to contrast dialectal varieties, it has been included, thanks to the *Associació Cultural del Matarranya*[1], one tale (lasting 9 minutes) included in a CD named "Qüento va, qüento vingue".

Current materials can be looked up according to both segment textual and, occasionally, word document ordering. In an oncoming future, a general index will allow to find any word occurrence within RETOC.

As far as the core of current RETOC is concerned, that is, all materials from CR, it has to be recognised that we were too optimistic about dealing with all its files, as CR had provided us with both text and voice in an magnetic form. Nevertheless, facts have showed that only a fifth part of these materials was directly exploited; the remaining part is fraught with some of these problems:

- there is sound music, which is disrupting
- we lack of the text
- material containing advertisements
- material containing music
- material in a foreign language (statements, interviews, etc.)

The problem encountered by the sound music is twofold: first, it makes that the chances of a particular excerpt to be reused for further research projects decrease; second, it goes against the principles on which SINCRO is based to reach an automatic segmentation. Therefore the main reason to dismiss all excerpts with music is that they all are heading news that are developed in detail and so they constitute to some extent already-treated material.

5. Future perspectives

Two important grants have been asked which, if given, there will be room for being optimistic about the future:

- An action of the PROFIT 2000 program, financially supported by the Ministry of Science and Technology from the Spanish government. This will enforce the DOPO project in that the RETOC material coming from CR would be treated in further detail; besides, the SINCRO general performances would be improved. In

[1] The Matarranya is a Western Catalan speaking territory placed at the border of the Aragon autonomous community where Castilian is also an official language.

short, this action will facilitate the collection of new material together with a more detailed processing of the current material.

- An integrated action performed by all Catalan language Services from the universities belonging to the *Xarxa Joan Lluís Vives* is aimed at providing RETOC with 250 conferences at a university level covering 25 specialised domains.

Furthermore, considering the collaboration with the CUB group it has to be expected that, when their corpus is public, the RETOC hypertextual indexes would permit to compare its own texts with those of the CUB group.

6. Further materials

Apart from the already digitalised materials, a number of new ones will be included once they are properly treated. Thus what is being elaborated includes what follows:

- A sample of radio stations from the Balearic Islands in the year 2000, compiled for RETOC by the General Directorate of Language Policy. It amounts to 69'.
- Materials from the CIO, which is a part of the CUB group. It includes about 10 hours from several radio stations in the Barcelona area in the year 1993.
- Several samples of about 90 minutes from the radio station *Arrels* of Perpinyà.

Also materials from thesis dissertations are likely to be included. In this case, doctoral candidates will be given technical support in the treatment of their materials which implies a CD editon of their work in the RETOC form. In return, once their theses are completed, they will become members of RETOC.

7. Ways of diffusion

The main way of RETOC diffusion is and has to be Internet; we have chosen a multimedia combination of voice files and HTML hypertexts due to their multiplatform nature. Thus it is only required a HTML viewer in a multimedia computer (see http://retoc.iula.upf.es/entrada.htm). However, there is no technical hindrance to use alternative ways, namely CD and DVD. Further text compilation can be published in a paper form. Finally, for those research projects requiring high quality versions, adapted versions can be made subject to agreement.

8. Future procedures of interrogation

If the working stages presented so far are developed within the allowed schedule, then we believe that access to the RETOC stages will be improved by adding a broadcaster classification according to anthropophonic features. This would enrich texts with lemmatisation or word categorisation with regard to the tools and the forms and tagsets of the IULA’s Technical Corpus, with phonetic transcription.

9. Ways of co-operation

The research field in Language Engineering at the Institute for Applied Linguistics from the Universitat Pompeu Fabra is open to several ways of collaboration regarding RETOC, the SINCRO software and the DOPO strategy:

- research teams interested in developing a marking-up strategy relevant to their own purposes (including intonation, pragmatic and phonematic transcription) concerning already matched texts or the revision of other teams' task.
- secondary school and university lecturers who, together with their students, could devote workshops to orthographically transcribe those excerpts lacking of text;
- as has been noted before, doctoral candidates who provide materials from their theses according to the RETOC procedures;
- research teams applying this strategy to other languages; here we will pay particular attention to those working with minority or threatened languages;
- partners who translate texts into other languages so as to facilitate their use in distance education strategies;
- or, simply, users who let us know their comments and criticisms. Put your questions in RETOC'S FAQ:

http://www.iula.upf.es:1080/faq/cache/147.html

10. References

CABRÉ, M. T.; DE YZAGUIRRE, LL. and M. LORENTE (1997) «El projecte CECA (Corpus escrit del català)», paper presented at the *Segon Col·loqui Lingüístic de la Universitat de Barcelona (CLUB-2)*. In: BOIX, E.; LORENTE, M.; LLORET, M.R. and LL. PAYRATÓ (eds.) *Corpus corpora. Actes del I i II Col·loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: PPU-Secció de Lingüística Catalana de la Universitat de Barcelona:115-125

http://terminotica.upf.es/ceca.htm

CAMPS, O. DE YZAGUIRRE, LL. and A. MATAMALA, «DOPO: un outil d'analyse orthologique», paper presented at the *1. Freiburger Arbeitstagung zur romanistischen Korpuslinguistik,* Octobre 2000

http://retoc.iula.upf.es/retocdoc/dopo.ps

DE YZAGUIRRE, LL., *Les Phoneticiels du Corpus de l'Université de Barcelone*, VI Journées GLAT

DE YZAGUIRRE, LL. (1996) «Els lingüicials del corpus UB» In: BOIX, E.; LORENTE, M.; LLORET, M. R. and LL. PAYRATÓ (eds.) *Corpus corpora. Actes del I i II Col·loquis*

Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: PPU-Secció de Lingüística Catalana de la Universitat de Barcelona: 297-304

http://terminotica.upf.es/membres/DE_YZA/PUBLI/LING.HTM

De Yzaguirre, Ll.; Vivaldi, J.; Badia, T.; Pujol, M.; Tuells, T. i M. T. Cabré (1998) «IULA's LSP Multilingual Corpus: compilation and processing». Paper presented at the ELRA conference, Granada, May 1998. *Proceedings of First International Conference on Language Resources and Evaluation*.

http://www.iula.upf.es/corpus/corpubca.htm

DE YZAGUIRRE, LL. , A. MATAMALA and M. T. CABRÉ (2000a): «El lematizador "PALIC" del IULA (UPF)». Paper presented at the XVII AESLA Congress, poster on *Corpus and Computational Linguistics*.

http://terminotica.upf.es/CREL/palic-es.ps

DE YZAGUIRRE, LL., A. MATAMALA, C. BACH, N. CASTILLO and E. USTRELL (2000b) «AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF)» Paper presented at the XVII AESLA Congress. Poster on *Corpus and Computational Linguistics*.

http://terminotica.upf.es/CREL/ambi-es.ps

DE YZAGUIRRE, LL.; CLUA, E. and A. J. FARRIOLS (2000) «Les corpus oraux et l'enseignement de la langue», paper presented at the *GLAT 2000, Multilingual Communication and Interactivity: The Word and Beyond*, *Brest*. 11-13 July.

http://terminotica.upf.es/brest/entrez-y.htm

MOREL, J.; TORNER, S.; VIVALDI, J.; DE YZAGUIRRE, LL. and M. T. CABRÉ (1997) «El corpus de l'IULA: etiquetaris» *Papers de l'IULA*, *sèrie informes,* 18. Barcelona: IULA, Universitat Pompeu Fabra.

http://www.iula.upf.es/paperca.htm