Mid-Term Report
Education and language in Memories of Labour
Projecto IPG 118

Coordinator: Nelma Moreira

February 2008

This report concerns work developed by the research team until January 2008. The research members of this project involve different social sciences domains: Linguistics (Joaquim Barbosa - FLUP) , Education Sciences (Maria Teresa Medina, João Caramelo - FPCEUP), (Cristina Nogueira , Silvestre Lacerda - CIIE), History (Manuel Loff - FLUP) and Computer Science (Nelma Moreira (coordinator), Rogério Reis, Norberto Lopes - FCUP). Several students are also part of the team: Mara Matias (Mestrado em Ciência de Computadores, FCUP), Bruno Monteiro (Mestrado em Sociologia, FLUP), Adriana Marcela Veiga Pinho Ferreira, Carla Patrícia Figueiredo, Marlene Sousa Carvalho, Pedro Afonso Lebre da Rocha, Raquel Rodrigues Monteiro (Licenciatura de Ciências da Educação, FPCEUP), Carla Susana Fernandes Silva, Carlos Manuel Rodrigues da Silva, Ivânia Magalhães da Silva Ribeiro, João Emanuel Leitão Ramalho, Joana Marcela Ribeiro de Sousa Barbosa da Rocha, Juliana Silva da Cunha (Licenciatura de Língua e Literaturas Modernas, FLUP).

The project aims to contribute, in a multi-referential perspective, to understand the processes of interaction between work, personal formation and professional identity of workers, in Porto, during the second half of the 20th century, through the analysis of biographical narratives that are being collected by the "Centro de Documentação e Informação sobre o Movimento Operário e Popular do Porto" (CDI) of Universidade Popular do Porto. The project has two main research axes that aim to understand the processes of interaction between work, personal formation and professional identity. The first of these axes focus on a linguistic and social-linguistic analysis, the second one focus on the education/formation processes, both of them interconnected with research in computational linguistics and semi-structured data processing.

The work of the students of Linguistics and of Computational Processing of Portuguese developed in this period correspond to the tasks I.1.1, I.1.2, I.2.1, I.2.2 , I.3.1 and I.3.2.

The computer science student analysed some of the existing software tools for natural language processing and specially for Portuguese. For corpora tagging, a tree tagger with special Portuguese rules was chosen for this work. Several

lexical statistics where implemented based on an compact storage of the tagged text: word occurrences, syntactic categories occurrences and occurrences of syntactic patterns identified in this project by the linguistics students (and reported below). As the corpus of this project are XML documents different specifications (abstracts, interviewers information, transcripts, etc), XML APIs were also studied and used to extract the relevant information and to built new XML documents.

The linguistics students analysed several syntactic patterns, namely: coordination, cases of subordination (temporal, concessive and completive clauses) and passive constructions. Based on the usual grammatical definitions and on examples they extracted several simple rules (patterns) that allowed (some) automation. This work was not trivial as these students do not have any formal or computational training during their degree, and so it was also innovative and a challenge.

The work of the students of Education Sciences and of computational manipulation of semi-structured documents in this period correspond to the tasks II.1 - II.3.

The CDI has a corpus of about 80 interviews with workers of different professions and with different social experiences. All biographical narratives have been recorded in audio and video, and a transcription of the interview was produced. The students edited and printed the transcriptions (in a total of more than 140 hours of audio recordings and 3500 pages) and a digital copy of the interviews (audio and video) was organised in order to constitute a permanent archive for the project.

Through the first analysis of the transcriptions of the interviews, a synthesis was elaborated that establishes, for each interviewed, sociographical data, labour pathways, political and unionism experiences and the participation in associative structures, and sketches the formal and informal relationship nets, where they had been involved. In the basis of the crossed mentions that emerged of the transcribed interviews, a list was established identifying potentials future interviewees.

In order to annotate and to classify the biographic narratives, all transcriptions were converted from simple text (Microsoft Word) to XML documents conforming a specific language developed by the project team. A first version of a classifier of transcriptions was developed that allows the creation of categories hierarchies and the association of text segments. The produced documents can be exported to HTML formats. This publisher will be used for the content analysis of the interviews.