Computer-assisted Terminology Processing

21 May 2012

Presenters:

M. Teresa Cabré, M. Amor Montané, Rogelio Nazar

Tutorial Overview

In this tutorial we will offer an introduction to Corpus Based Terminology both in theory and practice. In the first part, we will present some basic notions about terminology work and in the second part we will concentrate on the practical aspects of Corpus Based Terminology based on a software especially designed for this purpose according to the general guidelines of the Communicative Theory of Terminology (Cabré, 1999), henceforth TCT. The TCT adopts a linguistic approach to terminology taking also cognitive and discursive aspects into account; therefore its object of study are terms conceived in the context of specialized discourse. For the TCT, a linguistic approach to terminology is a corpus-based approach, in the line of advocates of corpus linguistics (Sinclair, 1991; Sager, 1991; McEnery & Wilson, 1996; etc.). This means that the TCT is not only interested in prescriptive terminology, i.e. the terminology established by standards or found in official databases but also (and particularly) in those terms that are actually used in L.S.P. (Language for Special Purposes) corpora by field experts. In other words, the TCT not only adopts an *in vitro* approach but is also interested in terms *in vivo*.

In the second part of the tutorial, we will go from theory to practice, and we are going to explain the various processes of corpus-based terminology work. Here, each where participant will have the possibility to start with us a glossary project, going through the whole workflow of dictionary creation from corpus creation to terminology extraction and management. This practical module will be based on the [Terminus](#) software, which was developed through a collective effort of linguists and programmers of the University Institute for Applied Linguistics of Universitat Pompeu Fabra in Barcelona. Terminus is a web based platform where the user will find tools for the whole terminology work-flow, including the compilation and analysis of corpora, terminology extraction. The software helps the user to create textual corpora and analyse its vocabulary with statistical means, including a supervised-learning algorithm for terminology extraction. The program can be "trained" by the user with lists of examples of validated terms. With these examples, the program develops a mathematical model of the terms which is then used to extract new terms from corpora. The extracted terms are organized in a database which can be fully configured and re-designed by the user.

The motivation behind Terminus is linked to our experience in teaching terminology, especially in the scenario of the elaboration of multilingual terminological projects based on corpora. We noticed that there was a mismatch between the ideas we were transmitting to our students and what they really could do in practice with the tools currently available in the market. By the same token, we also noticed that the execution of such projects involve notions that are too technical or abstract to be explained just with words. Consequently, we started to devise a computational tool that would help students better grasp complex notions by conducting their own experimentation, i.e. *learning by doing*.

Tutorial Contents

1. Foundations of corpus-based terminology

2. Empirical analysis

2.1. Corpus compilation

2.2. Concept structure design

2.3. Corpus exploration and terminology extraction

2.4. Glossary creation

2.5. Term management

3. Final remarks and next steps