



LREC 2012 Tutorial

21 May – Morning Session

Computer Assisted Terminology Processing

Speakers:

Maria Teresa Cabré

Rogelio Nazar

Amor Montané

Audience Profile:

Terminologists, translators and other language and terminology related professionals.

Technical Requirements:

A computer with Internet access for each participant would be helpful, but is not mandatory.

Motivation:

In this talk we will offer an introduction to the foundations of terminology processing within the framework of the Communicative Theory of Terminology (TCT), as well as a sketch of the practice of glossary creation. We will explain the general guidelines to elaborate a terminological product, typically a glossary or vocabulary, following the main phases of the terminology work flow: definition of the work, research in the domain field, concept structure development, creation of a corpus, terminology extraction and database creation and revision. We will present Terminus¹, a web application newly developed by the IULATERM Research Group in Terminology at Pompeu Fabra University (Barcelona), which has been designed specifically to facilitate the elaboration of a terminographic product from beginning to end. Terminus has a graphic and user-friendly interface that integrates tools for corpus constitution and management as well as terminology extraction and the design and customization of terminological records according to the user's needs.

Most current trends in terminology processing rely heavily on the aid of computational tools. Frequently, these tools comprise different systems for terminology database management. The organization of terminology in databases vastly facilitates the retrieval of information stored in terminological records and enables the user to group terms according to common features. Terminological databases of this kind usually allow the user to archive bibliographical references for the sources of the terms included in the database, as well as other informations such as definitions, contexts of occurrence, variants of the term and equivalences in other languages. Other functions that are incorporated into the terminology work flow, and are also an important part of the process, are the utilities for the constitution and exploration of specialized textual corpora, including tools for the extraction of concordances, corpus statistical analyses and automatic terminology extraction. However, and despite the growing interest in terminology processing tools, in our professional experience we have witnessed that the software industry still has not provided integrated technological tools to address the needs of terminology professionals, and this motivated the development of Terminus, which integrates the complete terminology work flow.

¹ The new version of Terminus (2.0) is hosted on a provisional URL (<http://iula05v.upf.edu/>) which can be accessed for demo purposes. This version will soon replace the first release which is currently hosted on the official URL (<http://terminus.iula.upf.edu>).

The development of a terminological project within the framework of the TCT is based upon two basic principles: quality and adequacy. The quality principle is guaranteed by an extensive research on the analyzed domain field and selection of reliable specialized documentation. The adequacy principle, in turn, consists of studying the professional needs of the end-users of the terminographic product. Terminological projects developed within the framework of the TCT are never created in the abstract, since they are conceived from the very start with respect to a particular function in order to satisfy the requirements of a specific group of professionals. This means that each project is adequate and coherent with respect to a communicative situation determined by such parameters as the professional activity, the specific field, the topic and the sociolinguistic context. The adequacy principle affects each one of the different stages of elaboration of a terminological project. In this tutorial we will focus on a typical work in terminology which is the creation of a glossary.

Outline:

We will talk about how to develop a model of a terminological vocabulary and explain the main concepts upon which terminography is based. Participants will be invited to develop with us a fictional terminographic resource using Terminus. They will be asked to think about the needs of a model of the user of such resource, thus elaborating a project adequate for addressing the needs defined for the constitution and exploitation of the corpus. Among the several corpus exploration methods provided by the program, such as key word extraction and n-gram frequency counts, there is an originally designed machine learning algorithm for terminology extraction, which can be trained by the user to be used in different domains simply by submitting lists of examples of terms of the desired domain field.

To sum up, the topics of interest, which will be taught in sequential order, are as follows:

- 1) Definition of the terminological project (title, design of terminographic records, etc.)
- 2) Research in the analyzed domain field and retrieval of the necessary documentation
- 3) Documentation on similar or related resources already developed
- 4) Elaboration of the conceptual structure of the field
- 5) Constitution of a corpus
- 6) Term candidates extraction and validation
- 7) Creation and edition of the database entries (writing definitions, finding equivalences, etc.)

The following figures show different screenshots of instances of the program.

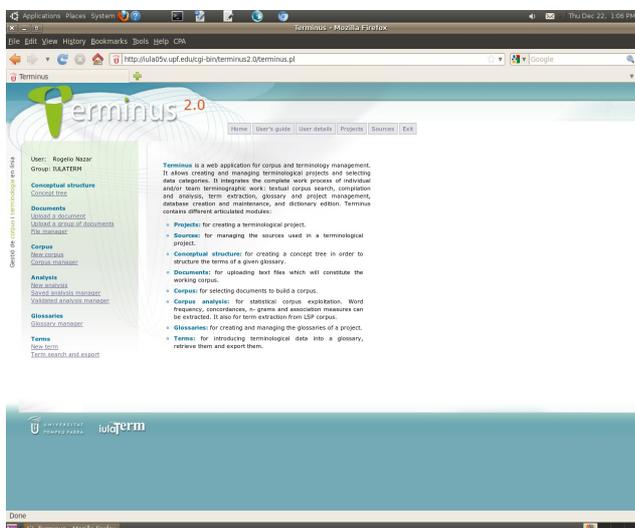


Figure 1: Terminus' home page

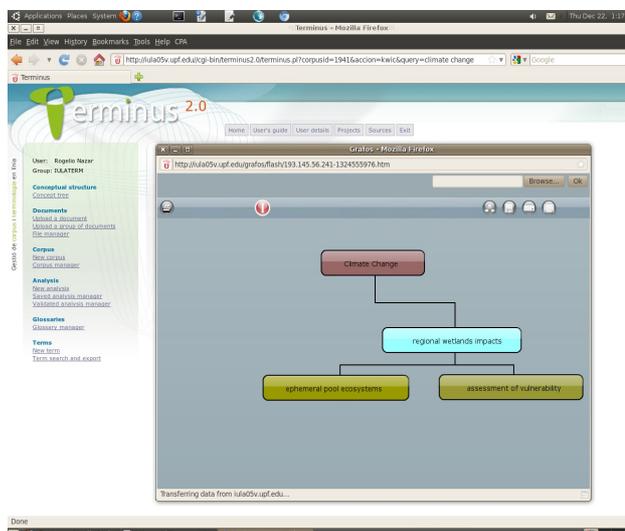


Figure 2: Elaboration of a concept structure

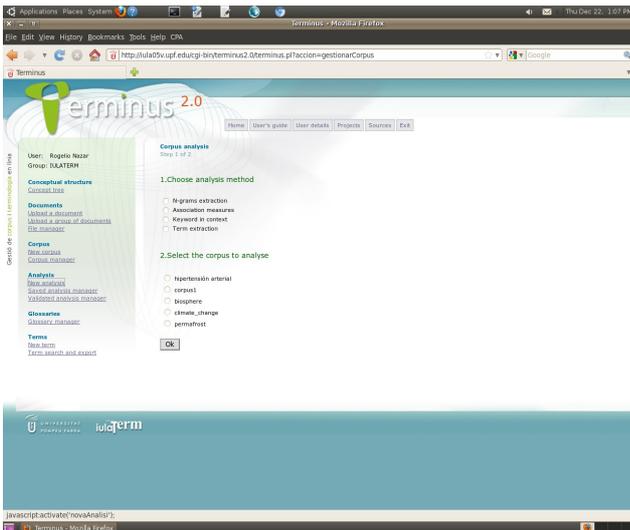


Figure 3: Different tools for corpus exploitation

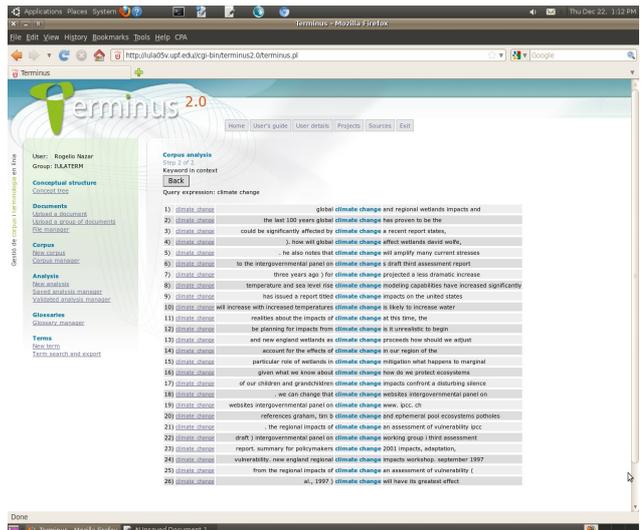


Figure 4: Key Words in Context

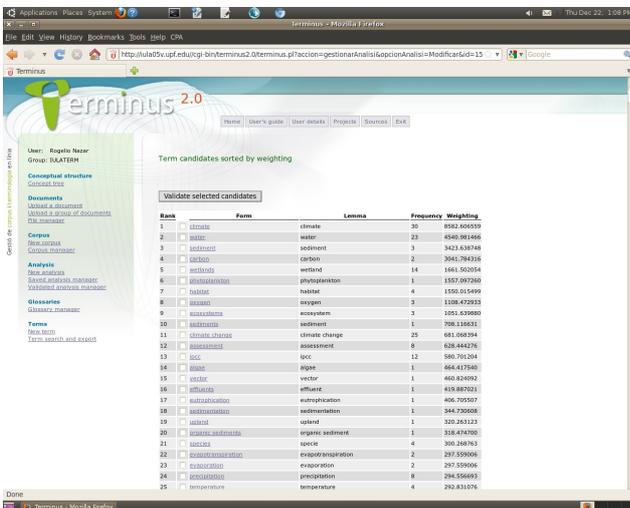


Figure 5: Term extraction

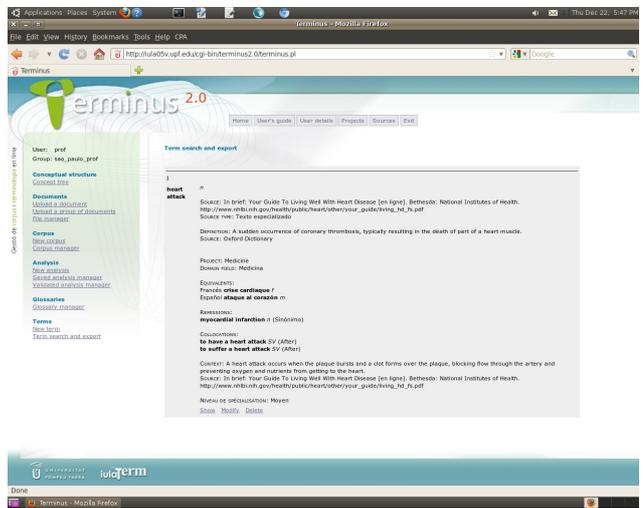


Figure 6: Retrieving database records