Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 212 (2015) 159 – 165

Procedia
Social and Behavioral Sciences

MULTIMODAL COMMUNICATION IN THE 21ST CENTURY: PROFESSIONAL AND ACADEMIC CHALLENGES. 33rd Conference of the Spanish Association of Applied Linguistics (AESLA), XXXIII AESLA CONFERENCE, 16-18 April 2015, Madrid, Spain

Creating corpus-based ontologies: a proposal for preparatory work

María Rosario Bautista-Zambrana[a]*

[a]*Universidad de Málaga, calle León Tolstoi s/n, 29071 Málaga, Spain*

Abstract

Constructing an ontology for terminological purposes involves several steps: among them, to acquire the knowledge necessary to create the ontology, to conceptualize the domain and to implement the ontology itself. This paper focuses on the first activity and proposes a protocol to work from a specialized corpus, extract terms, detect linguistic equivalents and extract conceptual relations, with a view to creating an ontology, which in turn can be the basis for a multilingual ontoterminological dictionary.

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE

Keywords: ontology; terminology; knowledge acquisition; corpus; term extraction

1. Introduction

It can be considered that the study of ontologies has carved out a significant place in recent years in the field of terminology. Thus ontologies have aroused the interest of quite a few researchers, who have studied how to build on their characteristics to create terminological resources. Several research groups have dealt with this area, among others the Centrum voor Vaktaal en Communicatie in Brussels, the Equipe Condillac from the University of Savoie, the Lexicon group from the University of Granada, the TecnoLeTTra group from the University Jaume I, or the Lexytrad group from the University of Málaga (see Durán-Muñoz & Bautista-Zambrana, 2013).

Ontology (in singular) has its origins in the field of philosophy, where it is the part of metaphysics that deals with the being in general and its transcendental properties (DRAE, 2001). Some decades ago, ontologies were imported

* Corresponding author. Tel.: +34-952133260; fax: +34-952131843.
E-mail address: mrbautista@uma.es

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE
doi:10.1016/j.sbspro.2015.11.314

from ontology engineering, branch of artificial intelligence, which deals with the creation of knowledge-based systems, so that knowledge can be codified and processed by computer. Within this discipline, an ontology is "a database describing the concepts in the world or some domain, some of their properties, and how the concepts relate to each other" (Weigand, 1997: 138). From this discipline, ontologies became adopted by some currents in terminology seeking a greater formalization in the conceptual structuration of the domain. Moreno Ortiz's (2008: 2) definition falls within this approach: ontologies are conceptual and terminological descriptions of a shared understanding about a specific domain.

Ontologies share some components with terminological resources (concepts, conceptual relations, concept denominations, definitions) and at the same time constitute a formalized, explicit, standardized and consensual representation system (Moreno Ortiz, 2008: 3), which can provide a greater degree of formality to terminology. That is why quite a few authors have studied the connection existing between both resources and how the ontological approach can improve terminological resources; apart from Moreno Ortiz (2008), we can cite, for instance, Temmerman & Kerremans (2003), Faber et al. (2009) and Leonardi (2012).

Building an ontology for terminological purposes requires several steps: to specify what type of resources we want to elaborate, with what aim and for which users, to acquire the necessary knowledge to create the ontology, to conceptualize the domain and to implement the ontology itself (see Gómez Pérez et al., 2004). This paper focuses on the second activity and proposes a work protocol for, based on a specialized corpus, extracting terms, detecting linguistic equivalents and extracting conceptual relations, with a view to creating an ontology, which in turn will be the basis for a multilingual ontoterminological dictionary. In this way, the described work guidelines will allow us to detect the concepts (with their corresponding denominations in several languages) as well as the conceptual relations that will afterwards lay the foundation for an ontology. We will provide some examples of the activities that we performed in Bautista-Zambrana (2013) using a corpus on package travel: the extracted concepts and conceptual relations served as a basis for creating an ontology, which later turned into a trilingual ontoterminological dictionary (Spanish-English-German) about the aforementioned domain.

Given that our paper will be illustrated by a case of the legal-tourist domain (package travel in Spain, United Kingdom and Germany, as defined by the Council Directive 90/314/EEC of 13 June 1990 on package travel, package holidays and package tours), therefore a culturally-dependent domain, our protocol will take into account what to do with the terms and relations that are common to the languages and cultures studied, and those that are specific to each one of them.

The paper is divided as follows: Section 2 describes our protocol: it is particularly concerned with corpus compilation, term extraction and detection of translation equivalents, as well as the procedure to extract conceptual relations. Finally Section 3 presents the conclusions of the paper.

2. Protocol

This protocol is based on the work carried out in Bautista-Zambrana (2013) and aims to serve as a model for those researchers who wish to build a multilingual ontology for terminological purposes.

2.1. Corpus compilation

Firstly, it is advisable to compile a corpus about the domain in question and to study its representativeness. When compiling a specialized corpus, it is essential to determine the criteria that will guide its design; in this regard, we need to take into account aspects such as corpus size, length of the texts, number of documents, medium of publication, topic, text type, authorship, languages (and whether the corpus will be comparable or parallel), date of publication, and geographical area (Bautista-Zambrana, 2013). If the purpose is to construct a multilingual ontology, it is advisable to compile comparable subcorpora for each one of the involved languages, as well as a parallel corpus containing translated texts in the same languages.

Once the corpus is compiled, we recommend to study its representativeness; we have carried out this task using the computer application *ReCor* (Seghiri, 2006; Corpas Pastor & Seghiri, 2010), which determines *a posteriori*, in an objective and quantifiable way, the minimum size a corpus should reach in order to be considered representative in statistical terms, irrespective of the language or text type.