Barrios MA, Ramos JA, Aguado de Cea G

Semantic labels and genus: improving specialized domain definitions

8th International Conference on Terminology and Artificial Intelligence (TIA09)
November 18th-20th, 2009
Toulouse, France.
Pages: 6
Presented as poster.

Semantic labels and genus: improving specialized domain definitions

María A. Barrios, José A. Ramos and Guadalupe Aguado

Ontology Engineering Group, Fac. Informática, Univ. Politécnica de Madrid
Avda. Montepríncipe s/n, 28660 Boadilla del Monte, Spain
mabarrios@delicias.dia.fi.upm.es;{jarg,lupe}@fi.upm.es

Abstract: Enriching linguistic resources with domain information has been considered one important target in natural language applications. This domain information can be undoubtedly provided by subject specific definitions. However, definition extraction of this domain information from specialized resources has revealed certain methodological problems in definition construction. This paper presents some problems that are mainly related to inconsistencies in definitions. To face these problems some Meaning-Text Theory tools have been used: semantic labels, lexical functions and actantial structure as a solution for inferring domain knowledge and a way of providing coherence to definitions.

Keywords: Definition extraction, linguistic resource enrichment, definition reuse, Meaning Text Theory.

1 Introduction

Specialized domain definitions have deserved the attention of many scholars from a wide range of fields, though with different purposes: terminology (Wüster, 1979/1998; Sager, 1990; Cabré, 1992), specialized languages (Swales, 1985; Trimble, 1985), corpus linguistics (Pearson, 1998), natural language processing (Malaisé *et al*., 2005; Sierra *et al*., 2008) and standardization institutions (ISO/DIS 704, 2008; ISO 1087-1, 2000). However, although definitions are of utmost importance nowadays in technical documentation, for example, when building an ontology, preparing a knowledge-based model or writing product or software specifications, to mention just a few, the problem of achieving a consensus on good practices in terminological definitions is far from being reached, even in standardized documents (Pozzi, 2000).

Some of these definitional problems refer to inconsistencies in the same lexicographic work, such as term variants, use of different genus to define lexical units belonging to the same lexical field, etc. In this paper, we present a use case in which, taking three domain resources as basis for manual extracting definitions, consistency and sistematicity problems have appeared. The three domain definition resources used have been provided by the National Geographic Institute of Spain (IGN-E), and they correspond to several knowledge bases: the Concise Gazetteer

(NC), the Numerical Cartographic Database (BCN25), and the National Topographic Database (BTN25).

In an attempt to solve these problems, we have applied the Meaning-Text Theory (MTT) principles (Mel'čuk, 1996 & 1997) to propose some systematic solutions for defining specialized domain terminology to be used in a knowledge-based linguistic resource, *BADELE.3000*[1] (Barrios & Bernardos, 2007; Bernardos & Barrios, 2008) that follows the MTT principles. Our aim is to enrich *BADELE.3000,* which is a general purpose linguistic resource containing the 3,000 Spanish nouns more used in Spain, with domain terms in order to convert it into a highly exploitable flexible lexicographic resource. The resulting resource called simply *BADELE* can later be used in ontology development, ontology localization, multimedia tagging and other web semantic tasks.

This paper describes briefly the main concepts of MTT that have been applied here and its lexicographic tools, and then deals with a use case and the solutions proposed to the problems that have been encountered. Finally, some conclusions and future work are also presented.

2 MTT and its lexicographic tools

As for the lexicographic tools applied to BADELE.3000, we have resorted to three concepts proposed by the Meaning-Text Theory (MTT).

The first one is the *lexical function* (LF) (Mel'čuk, 1996): a LF associates a given lexical expression L (such as *sound*), which is the argument or keyword of F, with a set of lexical expressions – the value of F (such as *loud*, *strong*, *heavy*, *deafening*, etc.) – expressing a specific meaning associated with F (for instance, 'intense' for the examples just mentioned which correspond to the LF known as **Magn**).

The second concept is the *semantic label* (Polguère, 2003): for our defining purposes, a semantic label is the equivalent to the genus in traditional definitions by genus and differentia. For instance, *whale* could be defined as a 'sea mammal that breathes air through a hole at the top of its head and is hunted for meat and for other purposes, as a source of other materials'. The first part of this definition, 'sea mammal', the genus, is known in MTT approach as semantic label; the second part of this definition, the differentia, can be sometimes attached to some LFs.

The third concept is the *actant* (Mel'čuk, 2004a, 2004b) and its derivate, the actantial structure. Actants correspond to beings or things that participate in the process expressed by a predicate: MTT approach considers that there is a sort of argument structure in all kinds of predicative words, which means that not only do the verbs have actants but also the adjectives, adverbs and the predicative nouns. The actantial structure reflects the syntactic expression of the actants, as shown in the

[1] *BAse de Datos del Español como Lengua Extranjera de los 3000 sustantivos más usados del español de España* (Database of Spanish as a Foreign Language which contains the 3.000 Spanish nouns more used in Spain)

example of *fleuve* (river) of Dicouèbe[2]: QUI COMMENCE AU lieu X, PASSE PAR LES lieux Z ET SE TERMINE DANS L'étendue d'eau Y (WHICH STARTS AT THE X place, FLOWS THROUGH THE Z places AND FINISHES AT THE Y area).

3 Enriching BADELE: problems and solutions

When we decided to reuse and enrich *BADELE.3000* with subject domain knowledge we assumed that for our goal (defining terms) we could follow the same criteria that had been adopted when defining common nouns in that linguistic resource. However, the systematic criteria followed in *BADELE.3000* were not present in the definitions of the geographical domain resources. The biggest problem found when reusing technical definitions proposed by experts was the lack of coherence due to the fact that definitions were written following an intuitive way. This lack of methodology in the knowledge sources implies inconsistencies in the definitions of terms belonging to the same lexical field; for instance, *autonomous region* and *province* were defined by experts as 'administrative unit', whereas *municipality* (in the same lexical field) was defined as 'a set of people and the territory where they live'.

So, in order to achieve coherence in the global set of definitions of the database, we resorted mainly to one of the MTT tools, the semantic label. Thus, we created one semantic label for every set of terms attached to the same semantic field (for instance, *municipality, autonomous region* and *province* were labeled as 'administrative unit'). These semantic labels helped to "normalize" the definitions proposed by experts, which have to be included in *BADELE*.

Then, we define the actantial structure of this semantic label: every 'administrative unit' has a group of other administrative units or people who live at this administrative unit under a potential specific government ('administrative unit' grouping a set of X governed by Y). Following this criteria, we redefine all these terms. Only the term *comarca* (region) was recognized as a specific 'administrative unit' which is a natural division of territory and has not a specific government.

We found many other cases of different genus in the definitions of terms belonging to the same lexical field, sometimes attached to the fact that the superordinate concept is not clear or that is not related to one specific lexical expression. For example, a *yard* is defined as a 'construction', whereas a *dovecot* is defined as a 'building', a *silo* as a 'container' and a *tenting arena* as a 'yard', as Table 1 shows.

Table 1 Variability in genus

Term	Definition
corral (yard)	*Construcción o pequeña estructura más o menos estable, creada especialmente para cobijarse los pastores o recoger el ganado* (Constructed area, more or

[2] *Dictionnaire en Ligne de Combinatoire du Français*, developed within the MTT framework and accessible via web by http://olst.ling.umontreal.ca/dicouebe/

Term	Definition
	less permanent, for shepherds or cattle)
palomar (dovecot)	*Edificio donde se recogen y crían palomas* (Building for sheltering and breeding doves)
silo (silo)	*Recipiente cubierto destinado al almacenaje de productos sólidos (grano, forraje, áridos, etc.*) (Container used for storing solid products (such as grain, fodder, sand, etc.)
tentadero (tenting arena)	*Corral o sitio cerrado donde se realiza la tienta de los becerros para comprobar su bravura* (Yard or enclosed place where bull calves are tested for their bravery)

Thus, we made use of the semantic label to solve this problem and in a first phase we used 'industrial installation' (see Barrios, Aguado and Ramos, 2009). However, in our linguistic resource, *BADELE.3000,* we found *yard* and *dovecot*, defined (as general terms) as 'animal sheltering', that is understood as 'construction where (an) animal(s) (of the same or similar species) live(s)'. The question that arose was: should we maintain this semantic label? When looking for an answer to this question we referred to lexical functions (LF) and found the LF $\mathbf{S_1}$, which is the name of the first actant of a word. Then we listed all the nouns that are related to this LF whose value is the noun denoting a type of animal, and found nine cases, listed in Table 2.

Table 2 Values of Lexical Function S_1

Noun	S_1
jaula (cage)	$\mathbf{S_1}$(jaula) = pájaro [$\mathbf{S_1}$(cage) = bird]
establo (stable)	$\mathbf{S_1}$(establo) = ganado (toro, caballo, oveja, cabra, cerdo) [$\mathbf{S_1}$(stable) = cattle (bull, horse, sheep, goat, pig)]
corral (yard)	$\mathbf{S_1}$(corral) = ganado, gallina, conejo [$\mathbf{S_1}$(yard) = cattle, hen, rabbit]
pocilga (pigsty)	$\mathbf{S_1}$(pocilga) = cerdo [$\mathbf{S_1}$(pigsty) = pig]
redil (fold)	$\mathbf{S_1}$(redil) = ganado [$\mathbf{S_1}$(fold) = cattle]
nido (nest)	$\mathbf{S_1}$(nido) = ave [$\mathbf{S_1}$(nest) = bird]
caseta de perro (kennel)	$\mathbf{S_1}$(caseta de perro) = perro [$\mathbf{S_1}$(kennel) = dog]
cuadra (stable)	$\mathbf{S_1}$(cuadra) = caballo [$\mathbf{S_1}$(stable) = horse]
palomar (dovecot)	$\mathbf{S_1}$(palomar) = paloma [$\mathbf{S_1}$(dovecot) = pigeon]

This lexical relation proves enough consistency to explain why all the nouns of the first column are labeled in our linguistic resource as 'animal sheltering'. Then we decided to maintain this semantic label for *yard* and *dovecot*, considering that it helps to achieve more coherence for the set of entries of our resource.

What should we do in other cases whose relations are not so clear such as *silo* and *testing arena*? The first term is not related to any noun denoting an animal-type by $\mathbf{S_1}$, and the second one, even if related to *cattle* it is not a place where *cattle* live. *Yard* (and other nouns of the same lexical field) was defined by experts as 'construction'. This word (construction) was considered in our resource as a semantic label, with the meaning of ‘something constructed for any concrete activity or purpose and usually placed in a specific location’. Then we realized that the LF **Hiper** (that is, the hypernym) relates ‘animal sheltering’ to its hypernym ‘construction’, and uses the semantic label ‘construction’ for *silo* and *testing arena*, which were added to our resource, thus sharing the semantic label with other general terms. Consequently, the definitions were modified before including them in our resource, as shown in Table 3.

Table 3 Our proposal

Term	Definition	Semantic Label
corral (yard)	*Habitáculo de animal para el ganado* (Animal sheltering for cattle)	‘Animal sheltering’
palomar (dovecot)	*Habitáculo de animal para palomas* (Animal sheltering for doves)	‘Animal sheltering’
silo (silo)	*Construcción destinada al almacenaje de productos sólidos (grano, forraje, áridos, etc.)* (Construction for storing solid products, such as grain, fodder, sand, etc.)	‘Construction’
tentadero (tenting arena)	*Construcción en la que se realiza la tienta de los becerros para comprobar su bravura* (Construction where bull calves are tested for their bravery)	‘Construction’

4 Conclusions

We have expanded and enriched the linguistic resource *BADELE.3000*, which contains the most used 3000 Spanish words, with 350 terms of the Geographic Phenomena domain. These terms were defined by experts and reused in order to improve a lexicographic resource, called simply *BADELE*, a database with more specific domain information and more comprehensive in comparison to BADELE.3000.

The definitions proposed by experts were very useful, although we found some problems, most of them derived from formal inconsistencies. In order to solve them we propose the use of MTT tools. When we found inconsistencies related to different genus in the definitions of terms belonging to the same lexical field, we used the semantic labels, the actantial structure and the complete set of lexical functions. By doing so, it enables to achieve a set of term definitions which are coherent not only with their own set of terms, but also with the set of all the common nouns contained in our original linguistic resource.

As our long-term goal is to obtain a linguistic resource which is as exploitable as possible, part of the future work will be dedicated to enrich BADELE with information of other domains. We can conclude that reusing definitions of terms proposed by experts is a good way to enrich a general linguistic resource, as long as all the formal inconsistencies are solved. The MTT tools have proved to be a valuable means by which to solve these problems.

5 Acknowledgements

This work has been partially funded by the National Project "GeoBuddies: Anotación semántica colaborativa con dispositivos móviles en el Camino de Santiago" (TSI 2007-65677 C02) and the European Project "NeOn" (FP6-027595).

References

BARRIOS M.A., AGUADO DE CEA, G., RAMOS J.A. (2009) Enriching a lexicographic tool with domain definitions: Problems and solutions. In Sierra *et al* (ed.) *Proceedings of the 1st International Workshop on Definition Extraction*. Incoma. Shoumen (Bulgaria), 14-20.

BARRIOS M.A. & BERNARDOS M.S. (2007) *BaDELE.3000*: An implementation of the lexical inheritance principle. In GERDES *et al*, (eds.) Meaning-Text Theory 2007. *Proceedings of the 3nd International Conference on Meaning-Text Theory*. Wiener Slawistischer Almanach. Sonderband, 69. 97-106.

BERNARDOS M.S. & BARRIOS M.A. (2008) Data model for a lexical resource based on lexical functions. *Research in Computing Science*, vol. 27.

CABRÉ M.T. (1992) *La terminología*. Barcelona: Empuríes.

ISO/DIS 704 (2008) *Terminology work — Principles and Methods*.

ISO 1087-1 (2000) *Terminology work. Vocabulary: Theory & Application*.

ISO 1087-2 (2000) *Terminology work. Computer Applications*.

ISO 19110 (2005) *Geographic Information – Methodology for feature cataloguing*.

MALAISÉ V., ZWIEGENBAUM P & BACHIMONT B (2005) Mining defining contexts to help structuring differential ontologies. *Terminology*, 11(1), 21-54.

MEL'ČUK I. (2004a/b). Actants in semantics and syntax I: Actants in semantics; Actants in semantics and syntax II: Actants in syntax. *Linguistics*, 42:1, 1-66; 42:2, 247-291.

MEL'ČUK I. (1996) Lexical functions: A tool for the description of lexical relations in a lexicon. In WANNER, L. (ed.), *Lexical functions in lexicography and natural language processing*. Amsterdam/ Philadelphia. John Benjamin, 37-102.

OGC (2003) *OpenGIS Reference Model*, Version 0.1.2, OGC Inc. Wayland, MA, USA.

PEARSON J. (1998) Terms in Contexts. *Amsterdam.* Philadelphia: John Benjamins.

POLGUERE A. (2003) Étiquetage sémantique des lexies dans la base de données DiCo. *TAL*. Vol 44, nº 2/2003, 39-68.

POZZI M. (2000) ISO 704 e ISO1087-1: Dos Normas del ISO/TC37 en conflicto. *VII Simposio de RITERM*. http://www.riterm.net/actes/7simposio/pozzi.htm

SAGER J.C. (1990) *Curso práctico sobre el procesamiento de la terminología*. Madrid: Fundación Germán Sánchez Ruipérez.

SIERRA G., ALARCÓN R., AGUILAR C. & BACH C. (2008) Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1), 74-98

SWALES J. (ed.) (1985) *Episodes in ESP*. Oxford: Pergamon Press.

TRIMBLE L. (1985) *English for Science and Technology: A Discourse Approach*. Cambridge: CUP.

WÜSTER E. (1979[1998]) *Introducción a la teoría general de la terminología y a la lexicografía terminológica*. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.