Identifying ontological mismatches between EcoLexicon and FunGramKB

Antonio San Martín
Pamela Faber
Universidad de Granada

Resumen

EcoLexicon es una base de conocimiento terminológica sobre medio ambiente que está en proceso de ser conectada a FunGramKB, una base de conocimiento multipropósito diseñada para PLN. Ambas bases de conocimiento se conectarán mediante alineamiento, que implica que EcoLexicon se convertirá en una ontología satélite de la ontología nuclear de FunGramKB, aunque ambos recursos continuarán siendo independientes.

La mayoría de las dificultades que surgen durante cualquier alineamiento proviene de la heterogeneidad. Los tipos más importantes de heterogeneidad son la sintáctica, la terminológica, la conceptual y la pragmática. Con el fin de definir la mejor estrategia para llevar a cabo el alineamiento y garantizar un mantenimiento eficiente a posteriori, es necesario estudiar detenidamente las disimilitudes entre ambas ontologías. Por ello, este artículo analiza las heterogeneidades ontológicas entre EcoLexicon y la ontología nuclear de FunGramKB para identificar posibles incompatibilidades ontológicas.

Palabras clave: ontología, alineamiento, incompatibilidad, EcoLexicon, FunGramKB.

Abstract

EcoLexicon is a frame-based terminological knowledge base on the environment that is in the process of being linked to FunGramKB, a multipurpose knowledge base designed for NLP. The approach chosen for the linking is the one known as alignment, which entails that EcoLexicon will become a satellite ontology of the FunGramKB core ontology, though each resource will continue to be independent.

Most of the difficulties in any alignment stem from heterogeneity. The most important types of heterogeneity are syntactic, terminological, conceptual, and pragmatic. In order to define the best strategy for such an alignment and guarantee efficient maintenance afterwards, the dissimilarities between the ontologies need to be carefully studied. For that reason, this paper analyzes the ontological heterogeneities between EcoLexicon and the FunGramKB core ontology in order to identify possible ontological mismatches.

Keywords: ontology, alignment, mismatch, EcoLexicon, FunGramKB.

1. Introduction

1.1. EcoLexicon

EcoLexicon (http://ecolexicon.ugr.es) (Faber *et al.* 2006; Faber *et al.* 2007; Faber 2011) is a multilingual visual thesaurus on the environment in English, Spanish, and German (currently under expansion to French, Russian, and Modern Greek). The targeted user groups are scientific writers, translators, and environmentally-aware sectors of the general public. In EcoLexicon, environmental concepts are codified in terms of both hierarchical and non-hierarchical semantic relations that are visually represented as a dynamic network. This information is complemented with natural language definitions in English and Spanish. It is hosted in a relational database that is being converted into a formal ontology for reasoning techniques and user queries (León and Magaña 2010). EcoLexicon focuses on conceptual organization, the multidimensional and multilingual nature of terminological units, and the extraction of semantic and syntactic information through the use of multilingual corpora.

The information in EcoLexicon is structured in terms of propositions and knowledge frames that are organized in an ontological structure. Its conceptual design is derived from information semi-automatically extracted from specialized texts and the structure of terminological definitions. Its top-level concepts are object, event, and attribute categories. The user interface offers various types of information (Figure 1).

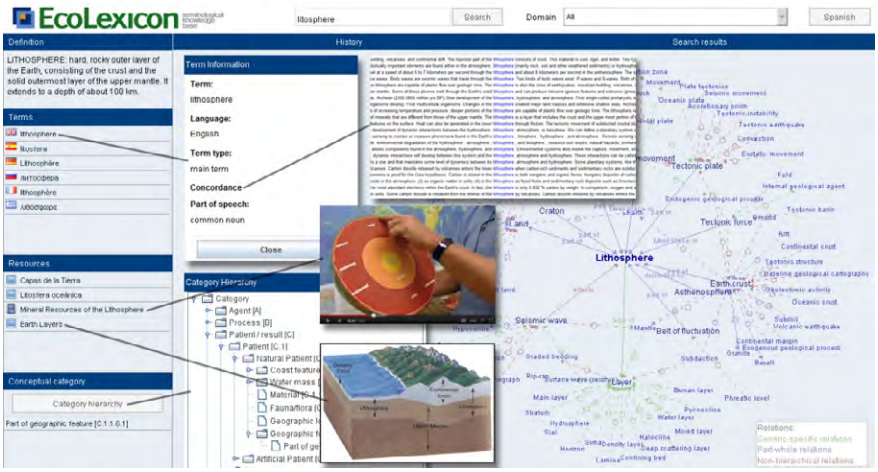

Figure 1. EcoLexicon representation of LITHOSPHERE.

In EcoLexicon, each concept is linked to other concepts by a closed inventory of semantic relations. Apart from the conceptual representation and the definition, concepts are linked to the terms in different languages, graphical resources, and their conceptual role in the environmental event (Figure 2).

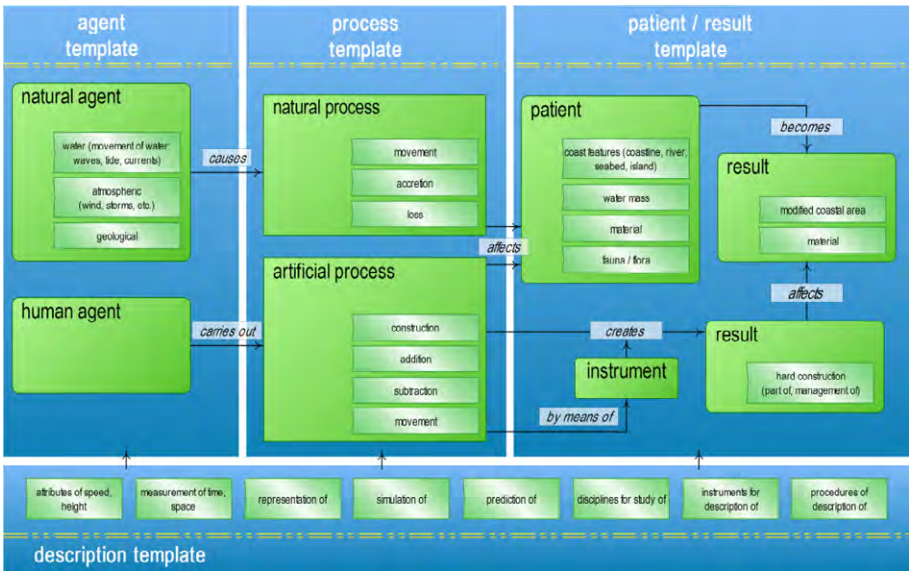

Figure 2. The environmental event.

1.2. FunGramKB

FunGramKB is a multipurpose knowledge base specifically designed for Natural Language Processing (NLP) with modules for lexical, grammatical, and conceptual knowledge (Periñán and Arcas 2010) (Figure 3).

FunGramKB's lexical level and grammatical level are language-specific whilst the conceptual level is not. The conceptual level in FunGramKB is composed of an ontology, a cognicon (which stores procedural knowledge), and an onomasticon (which stores information about instances of entities and events).

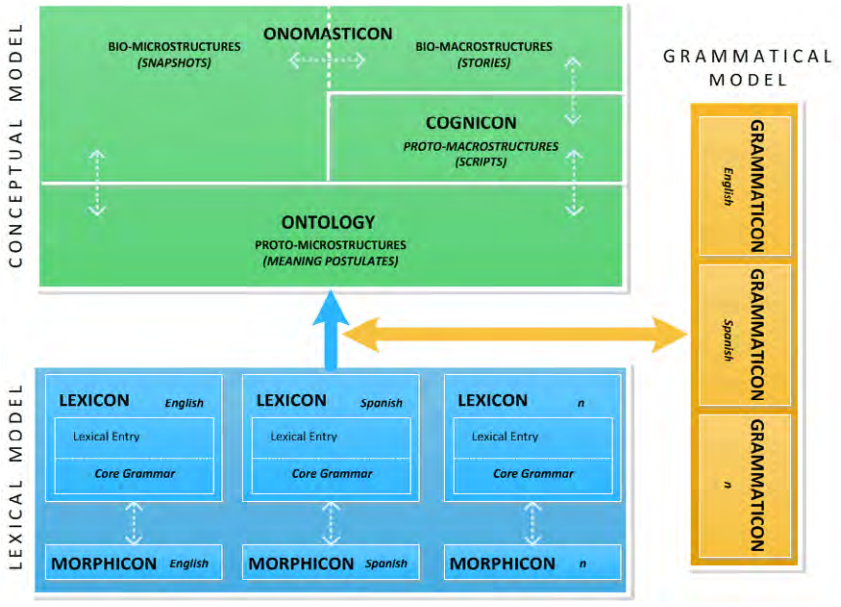

Figure 3. Architecture of FunGramKB (Source: http://www.fungramkb.com/).

1.2.1. The FunGramKB ontology

The FunGramKB ontology is a concept taxonomy, derived from general linguistic concepts, in which interlinguistic differences in syntactic constructions do not involve conceptual differences. It is being extended to include specialized knowledge by establishing links to satellite domain-specific ontologies.

Concepts are described in terms of meaning postulates (MP), written in a conceptual representation language, COREL. They belong to three levels. The upper level is composed of 42 metaconcepts distributed in three subontologies: #ENTITY, #EVENT, and #QUALITY. Basic concepts are at the middle level and are marked by + (e.g. +PENCIL_00). They are used in the MPs, and also encode the selection restrictions in thematic frames. The third level is composed of terminal concepts, marked by $ (e.g. $WHISTLE_00). They are not used to define other concepts in MPs.

1.3. Linking EcoLexicon and FunGramKB

For the purposes of EcoLexicon, a surface-semantic representation was initially considered sufficient because users can thus view concepts within a semantic network. Nevertheless, this kind of representation is not suitable for NLP (León and Reimerink 2011: 138). As a solution, EcoLexicon is evolving towards the status of a formal ontology and is being linked to FunGramKB. The linking approach chosen is *alignment*. This means that EcoLexicon will eventually be a satellite ontology of the FunGramKB core ontology though each resource will continue to be independent (Hameed *et al.* 2004).

This alignment is based on the use and extension of FunGramKB's basic and terminal concepts in the deep semantic representation of concepts in EcoLexicon. A crucial aspect of this linking is the mapping of *overlapping concepts,* or concepts that are represented in both FunGramKB and EcoLexicon (Faber and San Martín 2011). They are part of basic knowledge, but in the environmental domain they also acquire specialized meaning. In EcoLexicon, general concepts thus act as a scaffold for specialized meaning (Faber and San Martín *in press*).

2. Ontological mismatches

Most of the difficulties in any alignment stem from heterogeneity, which can take several forms. Of the many classifications of ontological heterogeneities, the most important are syntactic, terminological, conceptual, and pragmatic (Bouquet *et al.* 2004; Euzenat and Shvaiko 2007: 40-42). In order to define the best strategy for such an alignment and guarantee efficient maintenance afterwards, the dissimilarities between the ontologies need to be carefully studied. The following sections describe some of the potential heterogeneities in the alignment of FunGramKB and EcoLexicon.

2.1. Syntactic heterogeneity

Syntactic heterogeneity occurs when two ontologies do not have the same representation format (Bouquet *et al.* 2004: 6). Regarding FunGramKB and EcoLexicon, FunGramKB uses logically-connected predications in COREL to formalize meaning (Periñán and Arcas 2005), whereas EcoLexicon represents meaning by conceptual relations and natural-language definitions. This mismatch can be resolved by adding a COREL deep semantic representation of EcoLexicon concepts. When the linking process is completed, there will be three interrelated levels of concept representation in EcoLexicon:

Level 1: deep semantic representation of concepts in terms of COREL MPs, which can be accessed as natural language translations.

Level 2: a surface-semantic representation, which users can interact with via ThinkMap software.

Level 3: meaning definitions, encoded as the natural-language translation of COREL MPs. This translation must be adapted since features in the MPs will have to be added, omitted, or modified for the sake of explanatory adequacy. MPs are designed to be interpreted by a machine, therefore certain information that would be deemed obvious for a human needs to be encoded. Similarly, knowledge that would be easily inferred by the machine may be particularly useful if made explicit to the user for him/her to gain a better understanding of a concept.

2.2. Terminological heterogeneity

Terminological heterogeneity is when the same concept in different ontologies is named differently (Bouquet *et al.* 2004: 7). This usually stems from the use of different natural languages or from terminological variation: different levels of specificity (WATER CYCLE vs. HYDROLOGIC CYCLE), synonymy (WEATHER FORECASTING vs. WEATHER PREDICTION), geographical variants (AUTUMN vs. FALL), etc. This kind of problem is an obstacle to automatic matching with a computer application.

EcoLexicon concepts are associated with their designations in English (and other languages). There is one main entry term while other terms referring to the same concept are labeled as some type of variant. If these terms are automatically compared with FunGramKB's lexicon, a list of potentially overlapping concepts can be obtained. Ideally, the machine would do this, based on conceptual information.

Human intervention would be required to check for the omission of overlapping concepts, to discard incorrect matchings, and to resolve multiple matchings. Some FunGramKB concepts may match more than one EcoLexicon concept because of the non-specialized nature of FunGramKB and cognitive clustering (quasi-synonyms are grouped under the same concept in FunGramKB [Periñán and Mairal 2011: 24]). This would be the case of +VAPOUR_00 (FunGramKB), and VAPOR and STEAM (EcoLexicon).

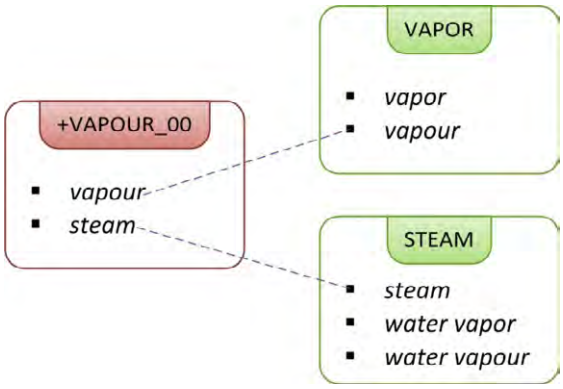

Figure 4. Simplified representation of the detection of EcoLexicon matches for +VAPOUR_00.

In this case, two EcoLexicon concepts could be mapped to +VAPOUR_00 and their specificities subsequently recorded in MPs. Alternatively, +STEAM_00 or $STEAM_00 could be included in FunGramKB and mapped to the EcoLexicon concepts.

2.3. Pragmatic and conceptual heterogeneity

Since pragmatic mismatches are caused by divergences in the interpretation of concepts or the intended usage of ontologies (Bouquet *et al.* 2004: 9). They thus have conceptual implications and are a source of conceptual heterogeneity.

Conceptual heterogeneity stems from divergences in conceptual modeling. It occurs at a metaphysical level when the concepts represented are not the same or when they are categorized differently. It occurs at an epistemic level when the semantic content of the concepts differs (Bouquet *et al.* 2004:7).

Conceptual heterogeneity at the metaphysical level is managed during the hierarchical rearrangement phase of the alignment, when the hierarchies of FunGramKB and EcoLexicon are rendered parallel in the transition zone (the area where concepts overlap). Heterogeneity at the epistemic level is handled during the conceptual modeling and mapping phase[1] to avoid inconsistencies. According to Bouquet *et al.* (2004), there are three types of conceptual modeling difference that can lead to conceptual heterogeneity: difference in coverage, in granularity, and in perspective.

2.3.1. Difference in coverage

Difference in coverage occurs when two ontologies (e.g. FunGramKB and Ecolexicon) focus on different portions of knowledge (Euzenat and Shvaiko 2007: 41). FunGramKB stores general concepts pertaining to everyday situations (Mairal and Periñán 2009: 219). In contrast, EcoLexicon is a domain-specific knowledge base. Some of the coverage areas in both resources overlap since many environmental concepts are part of everyday life.

For example, the FunGramKB core ontology stores concepts such as +LADDER_00 or +SPOON_00 which fall out of the scope of EcoLexicon. Conversely, EcoLexicon has concepts like SUPERFICIAL_RUNOFF or HEADLAND_BREAKWATER that are unsuitable for FunGramKB. Also, some concepts (overlapping concepts) appear in both ontologies, e.g. WATER, VEGETABLE or OXYGEN.

2.3.2. Difference in granularity

Difference in granularity occurs when two ontologies describe knowledge at different levels of detail (Euzenat and Shvaiko 2007: 41). FunGramKB is limited to commonsense knowledge, and does not include expert knowledge. In this respect, FunGramKB and EcoLexicon complement each other since EcoLexicon represents in-depth specialized information, whereas FunGramKB offers a uniform representation of upper-level concepts (Speranza and Magnini 2010: 230).

For example, +RAIN_00 in FunGramKB is defined as "the falling of water from the sky", whereas RAIN in EcoLexicon includes specialized information such as the type of cloud from which rain falls or the processes that cause it.

Granularity also differs because EcoLexicon, unlike FunGramKB, was not originally designed for NLP. In FunGramKB, some commonsense knowledge that is usually not covered in lexicographic resources need to be encoded for successful reasoning by the machine (Mairal and Periñán 2009: 218).

2.3.3. Difference in perspective

Difference in perspective occurs when two ontologies describe knowledge from different viewpoints (Euzenat and Shvaiko 2007: 41). FunGramKB depicts reality from the perspective of a person prototypically interacting with his/her surroundings. In contrast, EcoLexicon represents knowledge from the perspective of an environmental expert.

The description of basic scientific concepts for the general public is often at odds with their description for experts. Definitions of the same concept can vary greatly, depending on the knowledge level of the user group. For example, Lipschultz and Litman (2010) found that many entities defined as forces in

WordNet were really not forces, according to Physics. Consequently, an ontology reconciliation process will have to be carried out during the alignment of EcoLexicon and the FunGramKB core ontology.

3. Conclusions

During the linking process of two ontologies, heterogeneities are a key factor in the choice of methods and applications. This paper presents an approach to potential ontological mismatches between EcoLexicon and the FunGramKB core ontology. It is intended to be a guide for the forthcoming alignment, which is merely one of the steps towards the total integration of EcoLexicon and FunGramKB. This integration will include the development of a cognicon, an onomasticon, and various lexicons associated with an aligned deep-semantic EcoLexicon.

Notes

1. For an extended account of the alignment phases of EcoLexicon and FunGramKB, see Faber and San Martín (2011).

Acknowledgements

This research was funded by the Spanish Ministry of Economy and Competitiveness (project FFI 2011-22397).

References

Bouquet, P., Euzenat J., Franconi E., Serafini L., Stamou G. and Tessaris S. 2004. "D2.2.1 Specification of a common framework for characterizing alignment". (Available at: http://eprints.biblio.unitn.it/archive/00000653/01/090.pdf).

Euzenat, J. and Shvaiko, P. 2007. *Ontology Matching*. Berlin: Springer.

Faber, P. 2011. "The dynamics of specialized knowledge representation: Simulational reconstruction or the perception-action interface", *Terminology* 17 (1): 9-29.

Faber, P., León P., Prieto J. A. and Reimerink, A. 2007. "Linking Images and Words: the description of specialized concepts", *International Journal of Lexicography* 20 (1): 39-65.

Faber, P., Montero S., Castro M. R., Senso J., Prieto J. A., León P., Márquez C. and Vega, M. 2006. "Process-oriented terminology management in the domain of Coastal Engineering", *Terminology* 12 (2): 189-213.

Faber, P. and San Martín, A. 2011. "Deep Semantic Representation in a Domain-Specific Ontology: Linking EcoLexicon to FunGramKB". Paper presented at *44th Annual Meeting of the SLE*. Logroño.

Faber, P. and San Martín, A. (In pres). "Linking Specialized Knowledge and General Knowledge in EcoLexicon". *TOTh 2011. International Conference on Terminology & Ontology*. Annecy: Institut Porphyre.

Hameed, A., Preece A. and Sleeman D. 2004. "Ontology reconciliation", *Handbook on Ontologies*. Eds. S. y R. Studer. Berlin: Springer. 231-250.

León, P. and Magaña, P. 2010. "EcoLexicon: contextualizing an environmental ontology", *Proceedings of TKE Conference 2010*. Dublin.

León, P. and Reimerink, A. 2011. "EcoLexicon and FunGramKB : Applying COREL to Domain-Specific Knowledge", *Proceedings of the 24th International FLAIRS Conference*. 138-143.

Lipschultz, M. and Litman, D. 2010. "Correcting scientific knowledge in a general-purpose ontology", *Lecture Notes in Computer Science*, 6095: 374-376.

Mairal, R. and Periñán C. 2009. "The anatomy of the lexicon within the framework of an NLP knowledge base", *RESLA* 22: 217-244.
Periñán, C. and Arcas, F. 2005. "Microconceptual-knowledge spreading in FunGramKB", *9th IASTED International Conference on Artificial Intelligence and Soft Computing*. Anaheim-Calgary-Zurich: ACTA Press. 239- 244.
Periñán, C. and Arcas, F. 2010. "The Architecture of FunGramKB", *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta: ELRA. 2667-2674.
Periñán, C. and Mairal, R. 2011. "The COHERENT Methodology in FunGramKB", *Onomázein* 24: 13-33.
Speranza M. and Magnini, B. 2010. "Merging global and specialized linguistic ontologies". *Ontology and the Lexicon*. Eds. C. R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari and L. Prévot. Cambridge: Cambridge University Press. 224-238.