Multidimensional Categorization in Terminological Definitions[1]

Pilar León Araúz & Antonio San Martín

Keywords: *multidimensionality*, *terminological definition*, *EcoLexicon*, *recontextualization*, *contextual variation*.

Abstract

EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on the environment that currently holds 3,351 concepts and a total of 17,475 terms in English, Spanish, German, Russian, French, and Modern Greek. Concepts are linked by means of hierarchical and non-hierarchical relations in dynamic networks and in definitions. The environmental domain is interdisciplinary and its concepts can be categorized from different perspectives, thus conceptual representation needs to be multidimensional. Although, unlike other knowledge resources, conceptual representations in EcoLexicon reflect multidimensional categorization, this has also produced an information overload, particularly at upper concept levels. This means that many concepts show overloaded networks partly caused by multiple inheritance, as many of them have several hyperonyms. However, all conceptual dimensions do not occur at the same time but rather are context-dependent. Since the context of a concept is the set of concepts relevant to its intended meaning, we solved the information overload problem by recontextualizing networks in terms of discipline-based domains. The recontextualization of concepts constrains their relations with other concepts, depending on the activation scenario. By no means, does this imply that these are different senses of a polysemic term, but concepts also vary by context regardless of sense variation. Given that terminological definitions are also an integral part of the representation of multidimensionality, we applied the same contextual constraints to definitional propositions. The result is what we call flexible terminological definitions. This paper describes the representation of context-dependent multidimensionality in EcoLexicon and, more specifically, how this phenomenon is managed in terminological definitions.

1. Introduction

EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on the environment that currently holds 3,351 concepts and a total of 17,475 terms in English, Spanish, German, Russian, French, and Modern Greek. Concepts are linked by means of hierarchical and non-hierarchical relations in dynamic networks. Moreover, each concept is defined in English as well as Spanish.[2] This paper describes the representation of context-dependent multidimensionality in EcoLexicon and, more specifically, how this phenomenon is managed in terminological definitions.

2. Recontextualized semantic networks

The environmental domain is interdisciplinary and its concepts can be viewed from different perspectives. They can thus be regarded as multidimensional since they can be configured in several dimensions within the same concept system, depending on which feature or set of features is being focused on (Bowker and Meyer 1993: 123). Multidimensionality can thus give rise to multiple inheritance from various hyperonyms. Although, unlike other knowledge resources, conceptual representations in EcoLexicon reflect multidimensional categorization, this has also produced an information overload, particularly at upper concept levels. Figure 1 shows a segment of the overloaded network of SAND, which has six hyperonyms: AGGREGATE, FILTRATION MEDIUM, SOIL COMPONENT, BEACH FILL, SETTLEABLE SOLIDS, and SEDIMENT.

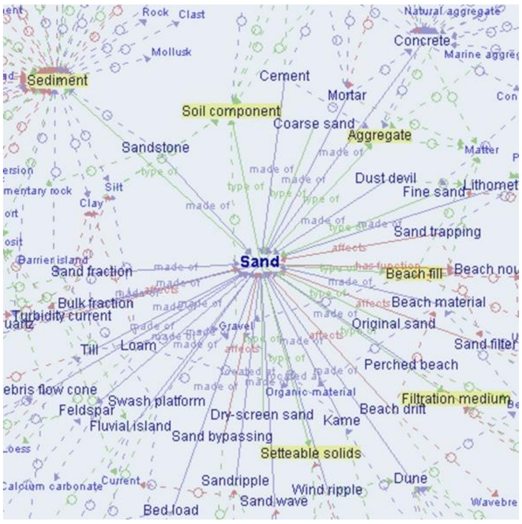

Figure 1. Information overload in the network of SAND.

Corpus evidence shows that all these hyperonyms generate other networks in which the concept has an important role. Nonetheless, they do not all occur at the same time (León Araúz et al. in press, León Araúz and Magaña 2010), but rather are context-dependent. For instance, SAND is mainly regarded as a type of SEDIMENT in geological scenarios, but other less common hyperonyms are more salient in other contexts, such as BEACH FILL in *COASTAL ENGINEERING* or SETTLEABLE SOLIDS and FILTRATION MEDIUM in *WATER TREATMENT AND SUPPLY*. Such contextual variation also has an important impact on non-hierarchical relations, such as *has_function*. For example, the proposition <SAND *has_function* BEACH NOURISHMENT> is only activated in a *COASTAL ENGINEERING* context, but not in any of the others.

Since the context of a concept is the set of concepts relevant to its intended meaning (Michalski 1991: 6), we solved the information overload problem by recontextualizing networks in terms of discipline-based domains organized in a hierarchical structure (e.g. *HYDROLOGY*, *CHEMISTRY*, *METEOROLOGY*, *WASTE MANAGEMENT*, *ENERGY ENGINEERING*, etc[3]). Nevertheless, concepts can still be visualized in a context-free mode, which corresponds to their general environmental hierarchy.

A concept is recontextualized in a certain domain by dividing the linking propositions into two sets[4]: active and inactive propositions (Figure 2). Active propositions are those that are relevant to the concept in a particular contextual domain. Inactive propositions are those that are either irrelevant to that domain ('latent propositions') or that pertain to another facet of the concept ('incompatible propositions'). Incompatible propositions would never occur in a particular context. Thus, the recontextualized semantic network only includes active propositions, such as the functional facet of SAND in natural contexts, whereas latent propositions are overridden based on the prototypical behavior of the concept, but not because they would never coincide in a real-world scenario.

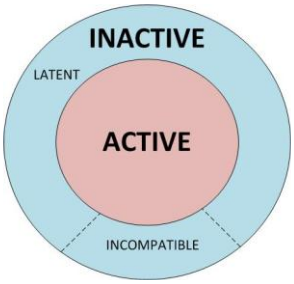

Figure 2. Types of recontextualized propositions.

In Figures 3 and 4 the concept SAND is recontextualized in the domains of *WATER TREATMENT AND SUPPLY* and *SOIL SCIENCES*. SAND is the same concept in all domains, but its relational behavior changes from one network to another. These works are more informative because of the straints applied. For example, some of the hyperonyms no longer appear and the proposition <SAND TRAPPING *affects* SAND> is only present in the *WATER TREATMENT AND SUPPLY* network, which becomes an incompatible proposition for *SOIL SCIENCES* and any other domain. Both networks also share certain propositions, such as <FINE

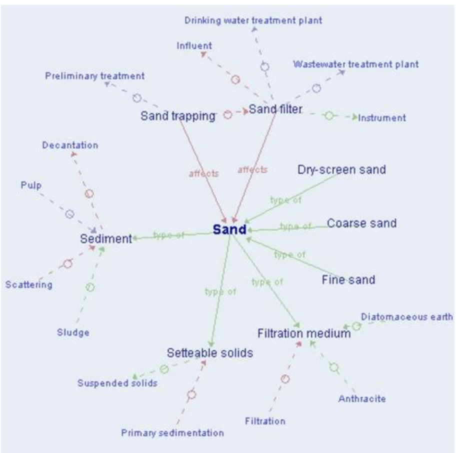

Figure 3. SAND in *WATER TREATMENT AND SUPPLY*.

SAND *type_of* SAND> or <COARSE SAND *type_of* SAND>, but the second hierarchical level in *SOIL SCIENCES* is broader, since it is composed of certain latent propositions (e.g. <SEDIMENT *made_of* SILT>) for the *WATER TREATMENT AND SUPPLY* context.

It should be highlighted that the recontextualization of concepts only constrains their relations with other concepts depending on the activation scenario. By no means, does this imply that these are different senses of a polysemic term. In EcoLexicon, contextual variation is also reflected at the microstructural level through flexible definitions, since semantic relations do not only vary by word sense. They vary by context as well, regardless of sense variation (Murphy 2003: 30).

3. Flexible definitions

Definitions in EcoLexicon have genus-differentiae structure. They also follow a template, based on their category membership, which reflects their conceptual structure (Faber et al. 2007). In order to account for multidimensionality, we propose the creation of 'flexible definitions' (San Martín in press, León Araúz and San Martín in press). This entails providing a set of recontextualized definitions for concepts with a high level of contextual variation. These definitions complement and enrich the general environmental definition (GED) of the concept. They are also standalone, and thus convey all the necessary information to define a concept in a certain domain, independently of the other definitions in the set. Flexible definitions follow the same premises used in the recontextualization of semantic networks. This means that only the most relevant propositions are represented.

Propositions are activated in a hierarchical structure similar to that of contextual domains. Thus, the propositions activated in a superordinate definition will also be activated in the definition corresponding to a subordinate contextual domain. Nevertheless, flexible definitions follow a bottom-up approach. The GED is at the top of the hierarchy and is a generalization based on the shared information of the subordinate definitions. This definition encompasses the whole environmental domain, and it represents the most prototypical features of the concept across all contextual domains. For this reason, it includes those propositions shared by all the recontextualized definitions (e.g., in the definition of SAND: <SAND *made_of* QUARTZ>). Also, certain propositions that are only partly shared are represented due to what we call 'general non-specification' and 'disjunctive generalization'.

General non-specification occurs when recontextualized definitions activate the same type of proposition, but one of the concept values differs in its intension. As a consequence, the value used in the general definition is a superordinate concept. For instance, when defining SEDIMENTATION in *WATER TREATMENT AND SUPPLY*, the proposition <SEDIMENTATION *affects* WASTEWATER> is activated, whereas in *PHYSICS*, it becomes <SEDIMENTATION *affects* SOLUTION>. As a result, in the GED, the proposition represented is <SEDIMENTATION *affects* FLUID>, since FLUID is a superordinate concept of both WASTEWATER and SOLUTION.

As for disjunctive generalization, it occurs when a conceptual relation is indispensable, but because of the concept's nature, the resulting definitional propositions are variable across contextual domains. In this case, all propositions will be activated in the GED as a disjunction. For instance, both <SEDIMENTATION *result_of* GRAVITY> and <SEDIMENTATION *result_of* CENTRIFUGATION> are necessary to define SEDIMENTATION in *WATER TREATMENT AND SUPPLY*. However, only <SEDIMENTATION *result_of* GRAVITY>

is relevant in *GEOLOGY*. As a consequence, the GED would activate <SEDIMENTATION *result_of* GRAVITY or CENTRIFUGATION>.

As previously shown, each discipline within the environmental domain gives rise to different conceptual propositions for the same concept and that also includes hyperonymic relations. In other words, disciplines categorize the same concepts differently, and this is reflected both in the networks and in the genus of flexible definitions (León Araúz and San Martín in press). Because of the coexistence of several hierarchies, concepts may have as many three hyperonyms depending on the contextual domain:

-The general environmental hyperonym corresponds to the superordinate concept in the general environmental hierarchy and the genus of the GED. It is applicable to all contextual domains, even though in some cases, this wider categorization may not be the most prototypical for certain contextual domains. For example, the general environmental hyperonym of SAND is MINERAL MATERIAL.

-Preferential contextual hyperonyms correspond to the superordinate concepts in each contextual domain hierarchy. The preferential hyperonym in a certain contextual domain is the genus of the recontextualized definition. For example, the preferential *SOIL SCIENCES* hyperonym of SAND is SOIL COMPONENT.

-Non-preferential hyperonyms correspond to all other superordinate concepts not included in the other two types. These hyperonymic relations come from those *type-of* propositions that have not been chosen to structure definitional hierarchies but are relevant to the recontextualization of networks. For example, a non-preferential *SOIL SCIENCES* hyperonym of SAND is SEDIMENT.

Hyperonymic relations always entail feature inheritance. Consequently, in EcoLexicon even recontextualized concepts inherit propositions from more than one hyperonym. For instance, SAND in *SOIL SCIENCES* inherits propositions from MINERAL MATERIAL, SOIL COMPONENT, and SEDIMENT. This gives rise to 'indirect definitional propositions' as opposed to 'direct definitional propositions', which are directly established by a concept. Following the genus-differentiae definition model, only one hyperonym can be the genus. However, some of the inherited propositions also need to be represented in the definition. Consequently, indirect definitional propositions may be of two types (Figure 4):

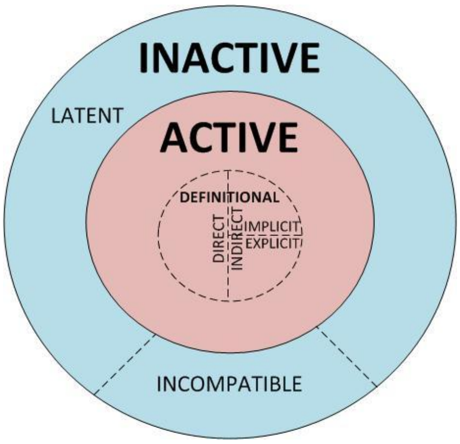

Figure 4. Recontextualized propositions in view of definitional purposes.

-Implicit definitional propositions are those propositions inherited from the genus itself. Therefore, they do not need to be represented, since that would be redundant.

-Explicit definitional propositions are those propositions inherited from hyperonyms other than the genus, and thus must be explicitly represented.

4. The case of SAND

All of the previously mentioned factors are represented in Table 1. Only the GED along with the *GEOLOGY*, *SOIL SCIENCES*, and *CIVIL*

ENGINEERING definitions of SAND have been reproduced as an example of flexible definitions.

Table 1. Flexible definitional template of SAND.

SAND			
General environmental definition	Unconsolidated **mineral material** consisting mainly of fragments of quartz ranging in size of 0.05-2 mm.	*GEOLOGY* definition	**Sediment** consisting mainly of fragments of quartz ranging in size of 0.05-2 mm that is part of the soil and can be found in great quantities in beaches, river beds, the seabed, and deserts.
Type_of	MINERAL MATERIAL	*Type_of*	SEDIMENT
Has_attribute	UNCONSOLIDATED	*Made_of*	0.05-2 MM FRAGMENTS + QUARTZ
Made_of	0.05-2 MM FRAGMENTS + QUARTZ	*Part_of*	SOIL (inherited from non-preferential genus SOIL COMPONENT)
		Located_at	BEACH + RIVER BED + SEABED + DESERT
SOIL SCIENCES definition	Unconsolidated **soil component** consisting mainly of fragments of quartz ranging in size of 0.05-2 mm that are the result of weathering and erosion.	*CIVIL ENGINEERING* definition	**Natural aggregate** consisting mainly of fragments of quartz ranging in size of 0.05-2 mm that is a component of diverse construction material such as concrete and mortar.
Type_of	INORGANIC SOIL COMPONENT	*Type_of*	NATURAL AGGREGATE
Has_attribute	UNCONSOLIDATED	*Made_of*	0.05-2 MM FRAGMENTS + QUARTZ
Made_of	0.05-2 MM FRAGMENTS + QUARTZ	*Component_of*	CONCRETE + MORTAR
Result_of	WEATHERING + EROSION (inherited from non-preferential genus SEDIMENT)		

Following the general non-specification rule, the general environmental genus acts as the hyperonym of all of the others. Accordingly, flexible definitions show how different genus candidates (SEDIMENT, SOIL COMPONENT, NATURAL AGGREGATE) highlight the changing nature of the concept in each recontextualized scenario. In this respect, SEDIMENT emphasizes the *result* dimension in the *GEOLOGY* domain; SOIL COMPONENT highlights the *partitive* dimension of the SAND in *SOIL SCIENCES*; and NATURAL AGGREGATE highlights the *functional* dimension of the concept in *CIVIL ENGINEERING* scenarios.

The GED only activates those propositions shared by the recontextualized definitions. For instance, <SAND *made_of* 0.05-2 MM FRAGMENTS + QUARTZ> and <SAND *has_attribute* UNCONSOLIDATED> are contained in all the recontextualized definitions. However, the latter is not explicitly represented in the *GEOLOGY* and *CIVIL ENGINEERING* definitions. In those two cases, it is an indirect implicit proposition derived from the genus SEDIMENT and NATURAL AGGREGATE respectively, since both of them are, by definition, UNCONSOLIDATED MATERIALS.

The *GEOLOGY* and *SOIL SCIENCES* definitions both inherit a proposition from a non-preferential genus. Given that both contextual domains share a close relationship, the genus of each is the non-preferential genus of the other. On the one hand, this implies that <SAND *part_of* SOIL> acts as an indirect implicit proposition in the *SOIL SCIENCES* definition, since it is inherited from the preferential genus SOIL COMPONENT. Conversely, in the *GEOLOGY* definition, the same proposition is inherited from the now non-preferential genus SOIL COMPONENT, and so it becomes an indirect explicit proposition. On the other hand, <SAND *result_of* WEATHERING + EROSION> works in the opposite direction. It is an implicit proposition in the *GEOLOGY* definition (it is contained in the genus SEDIMENT) and an explicit proposition in the *SOIL SCIENCES* definition (inherited from the non-preferential genus).

Apart from what can be implicitly or explicitly inherited from the three kinds of hyperonym, there are also two examples of direct propositions related to non-hierarchical relations. They are only present in the *GEOLOGY* and *CIVIL ENGINEERING* domains: <SAND *located_at* BEACH + RIVER BED + SEABED + DESERT> and <SAND *component_of* CONCRETE + MORTAR>.

5. Conclusions

Multidimensional categorization is a well-known phenomenon that causes multiple inheritance. It is often regarded as a way to enrich traditional static taxonomies. However, multiple inheritance should be modulated according to context, since not all propositions should be inherited from all hyperonyms. Contextual variation is a dynamic construct that triggers or restricts knowledge, and which includes all possible hyperonyms and the propositions to be inferred. Particularly when it comes to constructing definitions in a systematic way, only one genus can be preferential. Nevertheless, that does not mean that it should be preferential in all contextual domains.

It has been widely acknowledged that even within specialized domains, concepts can be defined according to many different facets. In this sense, we believe that terminological definitions are an integral part of the representation of multidimensionality and should not be left aside. Initially, recontextualizing definitions may seem excessively time-consuming. However, thanks to a controlled language formulated for this purpose, most of the propositions in flexible definitions can be inferred from information previously stored in EcoLexicon.

Notes

[1] This research was funded by the Spanish Ministry of Economy and Competitiveness (project FFI 2011-22397) and the Spanish Ministry of Education, Culture, and Sports (FPU Program AP2009-4519).

[2] For more information on how EcoLexicon was built and the different modules of information it has, see León Araúz et al. (2011).

[3] A comprehensive list of all contextual domains in EcoLexicon can be found in León Araúz and San Martín (in press).
[4] The following classification is an adaptation and extension of the one in Seppälä (2009).

References

Bowker, L. and I. Meyer 1993. 'Beyond Textbook. Concept systems: handling multi-dimensionality in a new generation of term banks.' In K.D. Schmitz (ed.), *TKE '93 Terminology and Knowledge Engineering*. Frankfurt am Main: Indeks Verlag, 123–137.

Faber, P., P. León Araúz, J. A. Prieto Velasco and A. Reimerink 2007. 'Linking Images and Words: the Description of Specialized Concepts.' *International Journal of Lexicography* 20.1: 39–65.

León Araúz, P., A. Reimerink and A. García Aragón. In press. 'Dynamism and Context in Specialized Knowledge.' *Terminology*.

León Araúz, P., A. Reimerink, and P. Faber 2011. 'Environmental knowledge in EcoLexicon.' In K. Jassem, P. Fuglewicz and M. Piasecki (eds.), *Proceedings of the Computational Linguistics Applications Conference*. Jachranka, Poland, 9–16.

León Araúz, P. and P. J. Magaña 2010. 'EcoLexicon: Contextualizing an Environmental Ontology.' In *Proceedings of the Terminology and Knowledge Engineering (TKE) Conference 2010*. Dublin, 341–355.

León Araúz, P. and A. San Martín. In press. 'Distinguishing Polysemy from Contextual Variation in Terminological Definitions.' In *10th International Conference of AELFE*. Valencia: AELFE.

Michalski, R. S. 1991. *Concepts as Flexible and Context-dependent Sets: the Two-tiered View*. Fairfax: George Mason University. Available at: http://www.mli.gmu.edu/papers/91-95/91-49.pdf

Murphy, M. L. 2003. *Semantic Relations and the Lexicon*. Cambridge: Cambridge University Press.

San Martín, A. In press. 'Hacia la flexibilización de la definición terminológica.' In *Proceedings of the XXIX AESLA International Conference*. Salamanca: AESLA.

Seppälä, S. 2009. 'A Proposal for a Framework to Evaluate Feature Relevance for Terminographic Definitions.' In *Proceedings of the 1st Workshop on Definition Extraction*. Borovets: ACL, 47–53.