From linguistic patterns to ontology structures

Guadalupe Aguado de Cea, Inmaculada Álvarez de Mon and
Elena Montiel-Ponsoda

Ontology Engineering Group,
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain
{lupe, emontiel}@fi.upm.es
ialvarez@euitt.upm.es

Abstract: The aim of this paper is to contribute to the research on linguistic patterns focusing on the *subclassOf* relation for the semi-automatic construction of ontologies. Taking as a starting point those ontological structures corresponding to consensual modelling solutions, which are known as Ontology Design Patterns (ODPs), we identified the linguistic patterns that convey the relation captured in ODPs as Lexico-Syntactic Patterns (LSPs) and included them in an LSP-ODP pattern repository. LSPs will permit novice users the conversion of the domain field they want to model into an ontological structure. In the present contribution, the language of classification in Spanish is studied in order to collect the most common ways of verbally expressing the *subclassOf* relation. Then, the topology of the most common classification patterns is analysed to discover the type of ontological knowledge provided, i.e. which concept relation, and the two essential features in ontology knowledge: exhaustiveness and disjointness.

Keywords: lexico-syntactic patterns, classification language, ontology design patterns, ontologies.

1 Introduction

The importance of language for the extraction of knowledge and information has led to the use of texts and documents in the construction of several types of resources, such as dictionaries, terminologies, or ontologies, to mention but a few. The need for automating the process of knowledge extraction or semantic relation acquisition has constituted a field of research for more than fifteen years. Three main trends can be distinguished in the automatic identification of semantic relations: those relying on statistical measures about co-occurrence of terms (Maedche & Staab, 2002), those relying on regular expressions that usually convey a relation of interest, the so-called pattern-based approaches, and hybrid approaches that combine the two previous ones (Gillam *et al.*, 2005). Within the pattern-based approaches, which are the ones that interest us in the present research, most have focused on taxonomic and meronymic relationships (Marschman *et al.*, 2002; Cimiano *et al.*, 2005), and others have put the emphasis on the identification of non-taxonomic relationships of specific domains (Marshman & L'Homme, 2006; Sánchez & Moreno, 2008).

In Computational Linguistics, the idea of applying patterns to the discovery of semantic relations was introduced by Hearst (1992) in the early 1990s. In Hearst's classical work several key patterns for the extraction of hyponyms in English were shown. Hearst's patterns ("such as", "and other", "or other", "including", etc.) rely on a combination of words (prepositions, conjunctions, paralinguistic signs) with a certain grammatical function.

In Terminology, accelerating the extraction of concepts for the creation of terminologies by means of certain patterns was also seen as an attractive task. The main interest has been in *knowledge rich contexts* (Meyer, 2001) that contain definitional information mainly expressed by means of hyponymy, synonymy, meronymy, holonymy, function and causality relations. In this sense, these patterns, named knowledge patterns, have been defined as elements "that link two or more specialized knowledge units in a particular subject field" (Feliu, 2004: 27). Most of the research in this field has focused on verb-oriented knowledge patterns, i.e., patterns in which verbs are the ones that convey the semantics of the relation holding between two or more terms. Research work has been conducted for different languages such as French (Marschman *et al.*, 2002; Aussenac-Gilles & Jacques, 2006), Spanish (Alarcón & Sierra, 2003), Catalan (Feliu & Cabré, 2002) or German (Xu *et al.*, 2002).

The Ontology Engineering field has also benefited from the previously mentioned approaches, and has applied them to (semi)-automatically learn classes and/or instances to populate ontologies (Berland & Charniak, 1999; Cimiano *et al.*, 2004; Pasca, 2005; among others), or to directly learn ontological relations from texts (Kavalec & Svátek, 2005; Sánchez & Moreno, 2008).

In the approach presented here, we want to contribute to the research on linguistic patterns in a new and promising way, namely, by identifying those linguistic structures that convey the relation expressed in the ontological structures considered *consensual modelling solutions* in Ontology Engineering. Those consensual modelling solutions have been identified as Ontology Design Patterns (henceforth ODPs) in Gangemi (2005). They are considered as highly beneficial for the development of ontologies because they allow the reuse of best practices and speed up the development process. Our objective is then to establish a correspondence between ODPs and the linguistic structures that realize them, which we have also named Lexico-Syntactic Patterns or LSPs. These two types of patterns, ODPs and LSPs, will be included in a *LSP-ODP pattern repository*. This repository will constitute the core of a system intended for a semi-automatic identification of ODPs, starting from formulations in Natural Language (NL) of the domain aspects novice users wish to model in ontologies. Therefore, the aim of this approach is to assist non ontology engineers in the reuse of ODPs for an easier, faster and more reliable development of ontologies.

In the present contribution, the focus is on the so-called *subclassOf* relation ODP (see Suárez-Figueroa *et al.*, 2007), classified as a Logical ODP. Logical ODPs include domain independent patterns, i.e., patterns that can be used across domains, in opposition to Content ODPs, which allow representing relations that may only happen in certain domains (e.g.: *Role-task* ODP, *Participation* ODP). For an online repository of ODPs see: www.ontologydesingpatterns.org.

In (Aguado de Cea *et al.*, 2008), a preliminary version of the *LSP-ODP pattern repository* was already introduced for English. Our aim now is to analyse the language of classification in Spanish and collect the most common ways in this language to express the *subclassOf* relation. Once those linguistic structures are discovered, the ones that show more appropriate and efficient for ontology building will be formalized and included in the LSP-ODP pattern repository. The main difference with the previous approaches is that they focused on definitional (Meyer, 2001), exemplifying language (Hearst 1992), that can also provide this *subclassOf* relation, whereas in our approach we pay attention to the classification language usually employed by the user in a direct interaction with a system that transforms NL formulations into ODPs. In this sense, LSPs will have two main characteristics: 1) they convey in an assertive way how things are organized in a certain domain of knowledge; 2) they express the modelling aspect in a concise, compact, natural way. These two characteristics will certainly restrict the set of patterns to be included in the repository.

The remaining of this paper is structured as follows. Section 2 provides a brief state of the art about the language of classification, and describes the special features of classifications when being modelled in ontologies. Section 3 is devoted to the methodology followed in discovering classification knowledge-rich contexts, and the results obtained. Then, the criteria for the selection of the linguistic patterns to be included in the LSP-ODP pattern repository are explained. In section 4, an extract of the repository is shown. Finally, the paper is concluded in section 5.

2 The language of classification

In spite of the interest in hypernym/hyponym relations for extracting knowledge from texts, the language of classification has been understudied. From a pure theoretical view, Levin's (1993) exhaustive study of the semantics and syntactic properties of English verbs does not pay any attention to the verbs used in classifying expressions. The verb "classify" appears as belonging to the "characterize" verbs (1993: 181), that is, those verbs that "characterize" or describe properties of entities. In systemic functional approaches to language, classification verbs are considered to be relational. Following these functional approaches, it is in the field of scientific and technical language in English where this topic has deserved some attention. Halliday (1989) claims that explicit classification is a property of the scientific language used in didactic texts. Wignell *et al*, (1993:137) corroborate that classifying is a further step of description, establishing the difference between everyday language and technical language. In fact, classification is typical of certain scientific domains such as biology, botany, entomology, histology, zoology and many others. And they account for the fact of subjectivity in classifications: "Naming a thing always implies a classification, and the same thing with the same name can be classified differently depending on who is doing the classification."

Trimble (1985), concerned with the teaching of writing for university non native students, has considered classification as one of the five rhetorical functions found in written texts for the transmission of knowledge in technical language This author (1985: 86) identifies three types of classification depending on the amount of

information provided: complete, partial and implicit. A *complete* classification gives three kinds of information: the name of the class, the hypernym, the members of the class, the hyponyms, and the basis for classification. A class that has all its members listed is considered to be a 'closed class'. When the classification is *partial*, only the name of the class and the members are given. *Implicit* classification corresponds to pieces of text that have a different rhetorical function, for instance, definition.

From a different point of view, semantics places the emphasis on the relationship between lexical items. It is seen as a relation of inclusion. Lyons (1977: 291) defines hyponymy as "the relation which holds between a more specific, or subordinate, lexeme and a more general, or superordinate, lexeme". Hyponymy is also a basic relationship in Wordnet where the opposite of a hyponym is referred to as a hypernym. In this semantic line, Cruse (1986) draws a distinction between hyponymy and taxonomy, which stands for the expression 'X is a kind/ type of Y'.

Therefore, approaches to classification can be said to be covered, in a way, by the study of definitional contexts in terminology and lexicography (as already mentioned in Section 1) considering that a definition should describe the concept and its relations to other concepts in the concept system. However, according to (ISO FDIS 704), there are different types of definitions, besides the traditional aristotelic one. These definitional contexts can encompass a wider range of semantic relations such as synonymy, meronymy, causality, or purpose. The hyponym-hypernym relation can be considered to overlap or even subsume the classification relation, which could be regarded as more specific of certain domains (e.g. biology or zoology) or for certain purposes (e.g., the construction of a taxonomy or an ontology).

From an ontological perspective, the classification relation is of great interest because it helps to organize ontology concepts into taxonomies and thus define the concept hierarchy, which is the basic organizational form in ontologies. By means of this relation, those identified as subclasses in the ontology inherit the properties and relations of the parent or superclass. In this sense, we are more interested in a *partial* classification than in a *complete* one, as very rarely are they produced, following Trimble's classification. The "name of the class" in the partial classification would correspond to the superclass or hypernym in the *subclassOf* relation, and the "members" would be equivalent to the subclasses. The "basis for classification" given in the *complete* classification would provide information about the criteria that determine a certain classification, or the author or school supporting it. In principle, there would be no need for reporting about that information at the conceptual level, since it is the author of the ontology the one who decides to include one classification and not another.

What is also highly recommendable in ontology modeling is to further specify some of the specificities of the *subclassOf* relation necessary for assuring correctness in subsequent reasoning possibilities offered by the ontology. These particular features are *disjointness* and *exhaustiveness*. Disjointness accounts for the fact that subclasses that belong to the same superclass do not share any instances, as pointed out in (Gómez-Pérez *et al.*, 2003: 134). In fact, *Disjoint Classes* is a Logical ODP that allows representing a set of disjoint classes. Additionally to disjointness, there is a further characteristic of the *subclass of* relation that has to do with Trimble's

"closed class", namely, exhaustiveness. That is to say, identifying whether the set of spotted subclasses are all the possible ones to be covered by the superclass (Gómez-Pérez *et al.*, 2003), or at least all the classes the author wants to include in the ontology as subclasses of a superclass. In the same sense, there is a Logical ODP for representing *Exhaustive Classes*.

3 Methodology and results obtained

Apart from the studies on definitional patterns in Spanish already mentioned in Section 1, we have found other patterns that express hierarchical relationships by means of classification sentences. Since the purpose of our research is to discover LSPs to assist novice ontology builders in the reuse of ODPs when modelling a domain of knowledge, the identification of linguistic patterns directly from domain documents is just the first stage of the knowledge acquisition activity. For this aim, we prepared an "ad hoc" corpus which focused on the patterns targeted (Pearson 1998: 48). To build this "ad hoc" corpus, several textbooks of subjects mainly concerned with the classification of natural phenomena in a domain were selected, such as histology, biology and zoology. This "ad-hoc" corpus served to establish a tentative list of 'seed' words and to discover its main patterns of use. These patterns were used to search for sample sentences in the on-line Corpus of current Spanish of *Real Academia Española* (CREA), with more than 150 million words. We focused only on the Science and Technology subcorpus, as it is where more phenomena and entities are classified. This subcorpus includes 10 % of the documents of the CREA. In order to discover the topology of classification patterns, we used the distance operator to determine what particles, words or prepositions appeared accompanying the verb.

The result of this initial step was a set of candidate knowledge-rich contexts in which different classification patterns were present. However, the obtained contexts can not be directly used in the development of ontologies without previous filtering, since the "reliability or certainty" of the information needs to be previously assessed, as claimed in (Marschman, 2008). "Certainty" has to do with the presence of some lexical indicators of quantification (e.g.: *some X are classified into…*), hedging (e.g.: X *is basically classified into…*), the use of modal verbs (e.g.: *X groups of Y may be distinguished*), and the negation. (e.g.: *one cannot distinguish X from Y*). In the present approach, we trust the filtering task to the novice user who, despite not being an expert in ontology engineering, is nevertheless expected to have a good command of the domain in question. In this sense, we assume that when interacting with the system for creating ontologies, the user will either discard those contexts (or parts of it) of no interest or reformulate them. However, we have foreseen to provide the user with some language recommendations when interacting with the system before (s)he starts using it for the first time. The reason for this sort of "guided natural language approach" is that allowing users to introduce statements in full natural language would require powerful and consistent processing resources in order to deal with language ambiguities, what is not fully possible nowadays. On the other hand, we consider that the proposed approach is sufficiently "natural" (or close to natural

language) so that users can concentrate on the content they need to model in the ontology, instead of being worried about how to express it.

Therefore, we can conclude that the translation of extracted knowledge into an ontology structure is not completely automatic. It rather requires the user intervention when formulating the modelling aspect. Despite that, it is still a great help to the user, since manually selecting the most suitable Ontology Design Pattern (ODP) has proven to be a complex task even for users with some modelling background, as some experiments have revealed (Aguado de Cea et al., 2008; Blomqvist et al., 2009). The main reason for the proposal here presented is the lack of any kind of support or tool that guides them in the selection. Thus, the linguistic patterns we propose will help bridge the gap between naïve users and a subset of Logical and Content ODPs.

3.1. Results

The main verbs forming part of classifying patterns extracted from the corpus are the following: *clasificar* (classify), *figurar* (figure), *distinguir* (distinguish), *dividir* (divide), and *comprender* (comprise). One may wonder why "to be" (*ser*) is not included, since it has been taken into account in most pattern studies dealing with definitions and taxonomical relations. The main reason is that this verb *per se*, without taking part in a more complex construction or collocation, produced too much noise. In order to restrict to those contexts dealing with classification, the present verbal form in plural *son* (are) had to be combined with *tipos de* (types of) or *clases de* (classes of). The singular use of both lexical forms (*es un tipo/clase de*) also produced a lot of noise.

In the following, we will analyse the main characteristics of those constructions, both from the linguistic perspective and from the ontological one:

Clasifica (classifies)

(1) *Un AGENTE* ***clasifica H en*** *N tipos/grupos/clases* ***según/ de acuerdo con*** *(criterio): X, Y y Z* (An AGENT …classifies H into N types/groups/classes according to (basis): X, Y, and Z)

Se clasifica (is classified)

(2) *X se clasifica* ***como*** *H* (X is classified as H)
(*3) X* ***se clasifica dentro de*** *la familia o del grupo H,* ***con la denominación****...* (X is classified within the family or group H under the name of …)

Se clasifican (are classified)

(4) ***Según/de acuerdo con*** *(criterio), las/los H* ***se clasifican en*** *X, Y y Z* (According to … H are classified into X, Y andY)
(5*) Los/las H* ***se clasifican en*** *X* ***o*** *Y* (H are classified into X or Y)
(6) Los/las H ***se clasifican generalmente/básicamente/comúnmente en diversos tipos****: X, Y y Z* (H are generally/basically/commonly classified into several/various types: X, Y, and Z)
(7) Los/las H se clasifican como X / X y/o Y (H are classified as X / X and/or Y)

a. The verb *clasifica* shows different syntactical constructions. It can be followed by different prepositions, such as: *como*, *de acuerdo con, dentro, en, para,*

según, each resulting in a different meaning for the pattern. This shows the difficulties of straightforwardly reusing that information for ontology development.

b. *Clasificar* appears in the active voice (*clasifica*) when the author of the classification is given. Information about the author would normally be omitted in the ontology.
c. The preposition *en* introduces the subclasses with or without a cardinal number followed by *tipos/grupos/clases*, and then the subclasses. Between the verb and the preposition *en* appears the superclass (e.g.: *Hoyle clasifica* ***las ideas*** *en 2 grupos:…* - Hoyle classifies ideas into 2 types:…). It can be deduced from this construction that the list of subclasses is disjoint and exhaustive.
d. The relation between syntax and semantics is clearly seen in some of the cases in which the preposition that follows the verb shows the criteria of classification (e.g.: *de acuerdo con la altura/el peso/etc.* – according to height/weight/etc.). In ontological engineering these criteria would correspond to ontology properties.
e. *Se clasifican en* has 60 occurrences; in 17 of them, it is followed by a cardinal number, and in 18 it is followed by colon and then the subclasses. In all cases the subclasses appear to the right. 8 out of 60 show those paralinguistic features typically used when presenting classification in texts: a), i) 1); 12 out of 60 are followed directly by the subclasses. In the remaining there are some indeterminate quantifiers (*varios*, *diversos*), or adverbs (*generalmente, básicamente, comúnmente)*. In those cases, exhaustiveness cannot be assured. These forms of inaccuracy have to be avoided when interacting with the system.
f. *Se clasifica dentro* gives 6 hits for subclass of, but the classification pattern topology differs from the previous ones, as the noun phrase that precedes the verb is the subclass, and the one that follows, the superclass (E.g.: *Esta grave enfermedad neurodegenerativa se clasifica dentro del grupo de las enfermedades hereditarias recesivas* – This serious neurodegenerative illness is classified within the group of the recessive hereditary illnesses).
g. When *clasificar* is followed by the preposition *para* no classification cases appear in the Science and Technology corpus. The 22 hits retrieved from the General Corpus refer to the sports domain. Thus, this pattern is dismissed as non-classificatory.
h. Certain combinations of verbs and prepositions revealed rich semantic contexts, for instance *se clasifican en,* whereas others were rarely used, for instance, *se clasifican dentro.* Spanish verb conjugation also proved to be very enlightening: the plural produced more accurate hits than the singular, for instance, *se clasifican* vs. *se clasifica*.

Figuran (figure, list, mention)

(8) ***Entre*** *los/las H* ***figuran*** *los/las* ***siguientes grupos/tipos/clases*** *: X, Y y/o Z* (Among H it is possible to list the following groups/types/classes: X, Y and/or Y)
(9) ***Entre*** *los/las H* ***figuran*** *X, Y y/o Z (*Among H it is possible to list X, Y, and/or Z)

i. *Entre los H figuran*: 16 hits in 16 documents. In all cases the concept relation is subclass of and the semantics of the pattern shows that there is no exhaustiveness in the classification. Some of the examples retrieved also show instances of a class. This construction is generally used when the main instances of a class are listed, but it would not be relevant for the ontology. Therefore, this pattern was discarded for the repository.

Se distinguen (are distinguished)

(10) *N* ***grupos/tipos/clases*** *de H* ***se distinguen*** *atendiendo a (criteria):* X, Y y Z (N groups/classes of H are distinguished in terms of (criteria): X, Y and/or Z)
(11) ***X e Y se distinguen por*** *F* (criterios/elementos) (X and Y are distinguished by F (criteria/element)
(12) *Podemos* ***distinguir*** *N/los* ***siguientes grupos/tipos/clases de*** *H: X, Y y/o* Z (We can distinguish N groups/types/classes of H: X, Y, and/or Z)

j. Usually, the distinguishing elements or criteria, as well as the number of groups, types or classes are given. In any case, exhaustiveness is assumed.
k. Disjointness is also present in any construction with *se distinguen*, because the semantics of the verb indicates that totally separate groups are listed.
l. The construction in (11) is not a classification pattern *per se*. The pattern just highlights the disjoint relation between two classes. The preposition *por* introduces the differentiating criteria that could be translated into a property in the ontology.
m. The presence of modals, like in (12), should also be avoided when interacting with the system for the development of ontologies, since it would indicate that the classification is "possible rather than certain" (Marshman, 2008).

Se dividen (are divided)

(13) *Los/las H* ***se dividen en****: X, Y y/o Z* (H are divided in: X, Y and/or Z)
(14) *Los/las H* ***se dividen en*** *N* ***grupos/tipos/clases***: *X, Y y/o Z* (H are divided in N groups/types/classes: X, Y and/or Z)

n. In 33 of the 64 cases the verb was followed by a numeral, whereas in 5 there was a colon, and in the rest, a list of subclasses. The list of subclasses introduced by this construction can be said to be disjoint and exhaustive. Interestingly enough, the singular form *se divide en* always (108 hits) presented a meronymic relation, and was consequently discarded.

Incluyen (include)

(15) *Los/las H* ***incluyen*** *X, Y y/o Z* (H include X, Y and/or Z)

o. The number of hits obtained with this verbal form, as well as with *comprenden*, was quite negligible. Moreover, it was prone to providing examples of meronimic relations rather than taxonomic ones. Therefore, this would be considered an ambiguous pattern that needs disambiguation when introduced by

the user in the system. For more details on this we refer to Aguado de Cea *et al.*, (2008).

4 LSP-ODP pattern repository

In total, two new LSPs were added to the *LSP-ODP pattern repository*. The first one (Table 1) can be directly identified with the Logical ODP for *subclassOf* relation (LP-SC-01), but no information is given about disjointness or exhaustiveness. The second one (Table 2) corresponds to the Logical ODPs for *subclassOf* relation, *Disjoint Classes*, and *Exhaustive Classes* (LP-SC-Di-EC). This is the most complete pattern and it will not require any further refinement to be directly translated into the corresponding ontological structure.

The patterns derived from this study represent the basic form of the nuclear pattern, as it can be called, i.e., the elements that compulsorily have to be present in the sentence. For example, in Table 1, in which an LSP corresponding to *subclass of* is represented, we can see that there is a noun phrase followed at some stage by *se clasifica como* as the main verb of the sentence, and also followed by an additional noun phrase: *NP <subclass> se clasifica como NP<superclass>*.

However, the use of certain modifiers is anticipated, although we expect that the recommendations given to the user by the system can deter him or her from using them. Moreover, the presence of less explicit or ambiguous patterns in the *LSP-ODP pattern repository* is also foreseen, as the experiments reported in (Aguado de Cea *et al.*, 2008) for the English language have proved.

Table 1. LSPs corresponding to the s*ubclassOf relation* ODP in Spanish

LSP Identifier : LSP-SC-ES	
NeOn ODPs Identifier : LP-SC-01	
Formalization	
1	NP[1]<subclass> se clasifica como NP<superclass>
2	NP<subclass> se clasifica dentro de [CN] NP<superclass>
Examples	
1	*La pimienta común (Piper nigrum) se clasifica como perteneciente al género Piper.*
2	*Esta grave enfermedad neurodegenerativa se clasifica dentro del grupo de las enfermedades hereditarias recesivas.*

Table 2. LSPs corresponding to the *subclassOf relation*, Disjoint Classes, Exhaustive Classes ODPs in Spanish

[1] NP stands for NounPhrase; CN stands for Class Name, and includes the generic names: group, type, class; CD stands for Cardinal Number; PARA stands for Paralinguistic Sign; Elements in brackets […] are meant to be optional, which means that they can be present either at that stage of the sentence or not, and by default of appearance, the pattern remains unmodified; Parentheses (…) group two or more elements; Asterisk * indicates repetition.

LSP Identifier : LSP -SC-Di-EC- ES	
NeOn ODPs Identifier : LP-SC-01+LP-Di-01+LP-EC-01	
Formalization	
1	Los/las NP<superclass> se clasifican en|como | se dividen en [CD] [los/las siguientes] [CN] [PARA] [(NP<subclass>,)* and] NP<subclass>
2	Se distinguen CD CN de NP<superclass> : [(NP<subclass>,)* y] NP<subclass> CD CN de NP<superclass> se distinguen : [(NP<subclass>,)* y] NP<subclass>
Examples	
1	*Los hongos se clasifican en cuatro grandes grupos: Ficomicetos, Ascomicetos, Basidiomicetos y Deuteromicetos.* *Las grasas se dividen en saturadas e insaturadas.*
2	*Se distinguen dos tipos de tilacoides: los tilacoides de las granas y los tilacoides del estroma.*

The tables contain information about a) the LSP Identifier (in which ES stands for Spanish); b) the NeOn ODPs Identifier, i.e., the identifier given to those patterns in (Suárez-Figueroa *et al.*, 2007); c) the set of LSPs formalized according to an extension of the Bakus-Naur Form; and d) the examples in NL.

5 Conclusions

To conclude, our study confirms that some classification patterns reliably convey information that can be directly transformed into ontological structures. Consequently, we are building a repository of Lexico-Syntactic Patterns (LSPs) that correspond to Ontology Design Patterns (ODPs), and which will be the core of a system for a semi-automatic construction of ontologies starting from NL formulations.

We have shown that obtaining knowledge rich contexts by means of seed words is a valuable stage in the acquisition of knowledge for the development of ontologies. However, those contexts cannot be directly reused in ontology development, but have to undergo a process of filtering or refinement on the side of the user. For that reason, only those classification patterns that certainly convey the knowledge targeted are included in the repository. Special attention has been paid to those patterns that, additionally to the hypernym-hyponym relation, provide information about disjointness and exhaustiveness, essential characteristics of the *subclassOf* relation in ontology modelling.

To sum up, this empirical research, originated from the need to assist novice users in the reuse of ODPs for the construction of ontologies has led us to search for those linguistic forms that straightforwardly express classification. Thus, we aimed more at a qualitative rather than at a quantitative study of classification patterns. Therefore, recall was not so relevant as precision at this stage.

In the future, our plan is to validate the LSP-ODP repository with real users formulating in NL what they want to introduce in the system for ontology modelling.

Acknowledgements

This research has been supported by the European project NeOn (FP6-027595), and the National project GeoBuddies (TSI2007-65677C02). We would also like to thank the reviewers, mainly one of them, for detailed and helpful comments.

References

AGUADO DE CEA G. GÓMEZ-PÉREZ A. MONTIEL-PONSODA E. & SUÁREZ-FIGUEROA M.C. (2008). Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns. In A. GANGEMI & J. EUZENAT Eds. *Proceedings of the 16th EKAW*.32-47. Springer Verlag.

ALARCÓN R. & SIERRA G. (2003). The Role of Verbal Predications for definitional context extraction. In *Proceedings of TIA 2003*. 11-20.

AUSSENAC-GILLES N. & JACQUES M.P. (2006). Designing and evaluating patterns for ontology enrichment from texts. In S. STAAD & V. SVATEK Eds. In *Proceedings of EKAW 2006*. 158-165. Springer Verlag.

BERLAND M. & CHARNIAK E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th ACL*. 57-64. Montreal, Canada.

BLOMQVIST E. GANGEMI A. & PRESUTTI V. (2009). Experiments on Pattern-based Ontology Design. *In Proceedings of 5th K-CAP Conference*. 41-48. Redondo Beach, California, USA.

CIMIANO P. PIVK A. SCHMIDT-THIEME L. STAAB S. (2005). Learning taxonomic relations from heterogeneous evicence. In P. BUITELAAR, P. CIMIANO & B. Navigli Eds. *Ontology Learning from Text: Methods, evaluation and applications*. 59-73. Amsterdam: IOS.

CIMIANO P. HANDSCHUH S. & STAAB S. (2004). Towards the Self-Annotating Web. In *Proceedings of the WWW2004*. 12-22. New York, USA.

FELIU J. & CABRÉ M.T. (2002). Conceptual relations in specialized texts: New typology and extraction system proposal. In *Proceedings of TKE'02*. 45-49. Nancy, France.

FELIU J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomática*. Ph.D. Thesis. Institut Interuniversitari de Lingüística Aplicada. UPF.

GANGEMI A. (2005). Ontology Design Patterns for Semantic Web Content. In Y. GIL, E. MOTTA, R. BENJAMINS & M. MUSEN Eds. Proceedings of the *4th ISWC*, Springer Verlag.

GILLAM L. TARIQ M. & AHMAD K. (2005). Terminology and the construction of ontology. *Terminology* 11 (1): 55-81.

GÓMEZ-PÉREZ A., FERNÁNDEZ M. & CORCHO O. (2003). *Ontology Engineering*. Springer.

HALLIDAY, M.A.K. (1985). *An Introduction to Functional Grammar*. London: Baltimore, Md, USA: Edward Arnold.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the *14th Conference on Computational Linguistics*. 539-545. USA: New Jersey.

ISO FDIS 704 (2009) *Terminology work. Principles and methods*

KAVALEK M. & SVÁTEK V. (2005) A Study on Automated Relations Labeling in Ontology Learning. In P. BUITELAAR, P. CIMIANO, B. MAGNINI *Ontology Learning from Text: Methods, Evaluation and Applications*. 44-58. IOS Press.

LEVIN B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press.

MAEDCHE A. & STAAB S. (2000). Mining non taxonomic conceptual relations from text. In *Proceedings of EKAW 2000*. 189-202. Berlin: Sringer.

MARSHMAN E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts. In *Terminology* 14(1). 124-151.

MARSHMAN E. & L'HOMME M.C. (2006). Disambiguating lexical markers of cause and effect using actantial structures and actant classes. In *Proceedings of LSP*. p.261-285. Bergamo, Italy.

MARSHMAN E. MORGAN T. & MEYER I. (2002). French patterns for expressing concept relations. In *Terminology* 8(1).1-30. John Benjamins.

MEYER I. (2001). Extracting Knowledge-rich contexts for terminography. In D. BOURIGAULT, C. JACQUENIM & M.C. L'HOMME Eds. *Recent Advances in Computational Terminology*. 128-148. Amsterdam/Philadelphia: John Benjamins.

PASCA M. (2005) Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded. In *Computational Linguistic and Intelligent Text Processing*. 280-292. Springer.

PEARSON J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

REAL ACADEMIA ESPAÑOLA Banco de datos (CREA) [online]. Corpus de Referencia del Español Actual. http://www.rae.es.

SANCHEZ D. & MORENO A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. In *DKE* 64. 600-623.

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

SUÁREZ-FIGUEROA M.C. BROCKMAN S. GANGEMI A. GÓMEZ-PÉREZ A. LEHMANN J. LEWEN H. PRESUTTI V. & SABOU M. (2007). *NeOn D5.1.1 NeOn Modelling Components*. NeOn Project.

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

WIGNELL P. MARTIN J.R. EGGINS S. (1993). The discourse of Geography: Ordering and Explaining the Experiential World. In M.A.K. HALLIDAY & J.R. MARTIN Eds. *Science Literacy and Discursive Power*. 136-165. London: The Falmer Press.

XU F. KURZ D. PISKORSKI J. & SCHMEIER S. (2002). A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *Proceedings of the 3rd LREC*, Las Palmas, Spain.