A DESCRIPTIVE STUDY ABOUT WORDNET (MCR) AND LINGUISTICS SYNSETS

Besharat Fathi*
Jorge Vivaldi Palatresi**

Resumen: Este artículo presenta el trabajo realizado para extender WordNet MCR al dominio lingüístico, que analiza las situaciones problemáticas ocasionadas por la estructura de WordNet y las características inherentes del dominio. Se ha empleado el enfoque descriptivo para poder explicar cómo el hecho de mantener la estructura original de WordNet puede llegar a afectar las extensiones de un dominio específico. Nuestros resultados demuestran que, para poder extender grupos de sinónimos cognitivos de dominios específicos, es indispensable realizar una reorganización estructural.

Palabras clave: WordNet, Lingüística, Base de datos de conocimiento de un dominio específico, Conceptos de un dominio específico.

Resumo: Este artigo apresenta o trabalho realizado para aplicar a WordNet MCR ao domínio linguístico e discute as situações problemáticas geradas pela estrutura WordNet e pelas características inerentes ao domínio. Foi empregado o enfoque descritivo para explicar como a manutenção da estrutura original da WordNet pode afetar as extensões de um domínio específico. Nossos resultados mostram que, para poder ampliar os synsets de domínios específicos, é inevitável uma reorganização estrutural.

Palavras-chave: WordNet, Linguística, Base de dados de conhecimento de um domínio específico, Conceitos de um domínio específico.

Abstract: This paper presents the work carried out towards enlarging WordNet MCR in linguistics domain, which discusses about problematic situations caused by WordNet structure and inherent characteristics of the domain. The approach employed in this paper is descriptive to explain how maintaining the original structure of WordNet might affect domain-specific extensions. Our results show that for any enlargement in domain-specific synsets a structural rearrangement is inevitable.

Keywords: Wordnet, Linguistics, Domain-Specific Knowledge Data Base, Domain-Specific Concepts.

Cómo citar este artículo: FATHI,Besharat; VIVALDI PALATRESI, Jorge. A descriptive study about Wordnet (MCR) and linguistics synsets. *Debate Terminológico. No. 12, Dic. 2014;* pp. 43-54

1. Introduction

WordNet (WN) is a lexical database whose development started in the eighties at Princeton University under the direction of Prof. G. Miller to test some psycholinguistic theories (Fellbaum, 1998). Since the very beginning it has been used for a number of interesting applications in natural language processing ranging from information retrieval, word sense disambiguation or lexicographical applications (e.g. Agirre,2007;Rosso, 2004; Gonzalo *et al*, 1998; Baker,1998; Miller,1995; Voorhees, 1993). Also it has been developed for many languages other than English[1] (Vossen, 2002).

The increasing number of WNs and related projects, and the variety of perspectives and multilingual approaches, mostly associated with general language, demonstrate an interest in understanding the structure and characteristics of those WNs and in particular a tendency to compare and align them to facilitate the extension process. Among those studies, still there are a few which have focused on specialized domains. A starting point for domain-based studies in WN might be to study their linguistic properties by considering their lexical characteristics and probable complexities. As the following stage, it can be analyzed how WN have overcome problematic situations, or whether the WN construction could prepare a reasonable solution to organize and systematize specialized knowledge or not. A wider point of view can be mapping specialized synsets of one WN to the synsets of another in order to reduce ambiguities.

In this paper, we have analyzed Linguistics concepts in terms of super-ordinate and subordinate relations based on the English synsets to detect the challenges of extending current WN.Only noun-related part of WN is concerned for two main reasons. First, in domain-specific data bases nouns have the most important role; second, organizing all data about verbs and adjectives, and adverbs (if there is any) is much more complicated. Thus, this

*Insitut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra). E-mail: besharat.fathi@upf.edu
**Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra, Barcelona, España. E-mail: jorge.vivaldi@upf.edu
[1]http://globalWordNet.org/WordNets-in-the-world/

article focuses on Linguistics noun-related synsets to give an overview about WN structure; however it can be developed in the future through all lexical properties. The leading motivation of this work is extending WN in linguistics domain. For this aim, this article first gives a short description about WN and domain-specific knowledge; second, it explains the main characteristics of Linguistics domain. Third, it studies the current synsets structure in Linguistics domain and problematic situations that might affect the extension process. Finally, we discuss that for any multilingual extension not only we need to consider conceptual structure of target language but also a structural revision upon current English properties in WN. Thus, this paper has a descriptive and structural view to linguistics domain in WN, mainly for English synsets. Nevertheless, any correction or extension in the future necessitates a multidimensional approach. This paper is a part of activities we have done in enlarging and improving the synsets in the linguistics domain. This task is being done in the framework of Enlarging the Multilingual Central Repository (MCR) (see González-Agirre*et al*., 2012), a project currently under development by several Spanish public universities.

1.1. WordNets and specialized knowledge

In any language, WN is a construction of *synset*s. A synset must be seen as a set of words (also called *variants* in the WN's jargon) where each one is interchangeable in a given context. Synsets are linked through taxonomic and non taxonomic relations, although the latter is poorly developed in all languages. WN is often named as a lexical ontology and it is intended to be general rather than cover some domains. Among those relations, "taxonomic relations i.e. hyperonymy/ hyponymy usually dominate (J. Ramanand, 2007:6), particularly in domain-specific synsets. Unfortunately at the beginning it has been manually created without any corpora support. It must be noted that although it is claimed that it is a general purpose resource, some domains (like Medicine) are much more developed than others. Also, the semantic relations are not developed to the same level in all languages. In order to obtain some ontological coherence it has been connected with other resources (WN domains, SUMO, BabelNet, etc.)

The most significant characteristics of WNs are that they usually maintain Princeton WN structure and its coherence has not been checked for specialized domains (see for example the analysis of the law domain in Zanotti*et al*., 2012). Often creation of WN in other languages just means translate nominal hierarchy without giving the necessary attention to the knowledge that is behind its structure. However, recently there is an interest either to employ WN model in domain-specific knowledge databases (e.g. ArchiWordNet) or to carry out studies on domain-specific concepts in WN (e.g. Bodenreider and Burgun, 2002).

1.2. Taxonomy and linguistics terminology

Linguistics is a broad discipline which contains concepts ranging from general linguistics and grammar to historical linguistics and sociolinguistics, and etc. Besides, various sub-domains are formed to elaborate either types of analysis or theories or thoughts. This feature caused a wide range of synonymous and polysemous terms. The existence of synonymous terms may refer to the chronological evolutions from traditional schools to modern theories and the rapid expansion of the domain which affect the usage of terms. Besides, those polysemous terms emerge inevitably due to the growth of sub-domains and are very frequent as a consequence of extending the meaning of a previously existing word to adapt it for specific needs. Any thought or school may define concepts from their own perspectives. Sometimes a concept possesses both changes, either chronological or polysemous, which causes a complexity for showing hierarchical coherence. This characteristic becomes more significant when super-ordinate and subordinate relations are concerned. For instance the term "predicate" represents various related senses from different points of view:

term	**definition**	**resource**
predicate	(In modern theories of syntax and grammar) the predicate of a sentence corresponds mainly to the main verb and any auxiliaries that accompany the main verb, whereas the arguments of that predicate (e.g. the subject and object noun phrases) are outside the predicate.	1.Wikipedia 2.http://projecteuclid.org/download/pdf_1/euclid.ndjfl/1093891495
predicate	(In traditional grammar) [Predicate, verb phrase] one of the two main constituents of a sentence; the predicate contains the verb and its complements.	1.WordNet MCR 2.http://www.oxforddictionaries.com/definition/english/predicate
predicate	(In linguistics) predicates are words that describe certain relations and properties, usually verbs and adjectives	http://faculty.simpson.edu/lydia.sinapova/www/cmsc180/LN180_Johnsonbaugh-07/L05-

		Predicate%20Calculus-11.pdf
predicate, logical predicate	(In semantics/ predicate logic) one of the meaning constituents of a proposition that is the smallest unit to which we can assign a truth value. * (logic) what is predicated of the subject of a proposition; the second term in a proposition is predicated of the first term by means of the copula.	1. Key concepts in language and linguistics. By R. L. Trask 2.https://cs.uwaterloo.ca/~plragde/ cs245/07-predsem.pdf 3. * WordNet MCR

Table1. Polysemy instances in linguistics

Moreover, because of the wide range of topics and interdisciplinary issues in linguistics, and the nature of the WN itself, the lexical domain borders (see definition in section 2) are not clear-cut. There are many terms from other related domains widely employed in linguistics that after a while due to the terminological needs their related concepts and notions are created, particularly belonging to the linguistics domain. In some cases the term belongs to linguistics but not its entire hyponyms or taxonomic forms. For example, the term *computational linguistics* is defined as "the use of computers for linguistic research and applications" (given by WN) which can be considered not only a computer science matter but also linguistics; while its hyponym *machine translation* is categorised in computer science and not linguistics. Another example, as it is illustrated in diagram (1), is *markup language* which is an interdisciplinary term belonging to linguistics and computer science; while types of *markup language* seems to be computer subject matters that should be categorised in computer science.

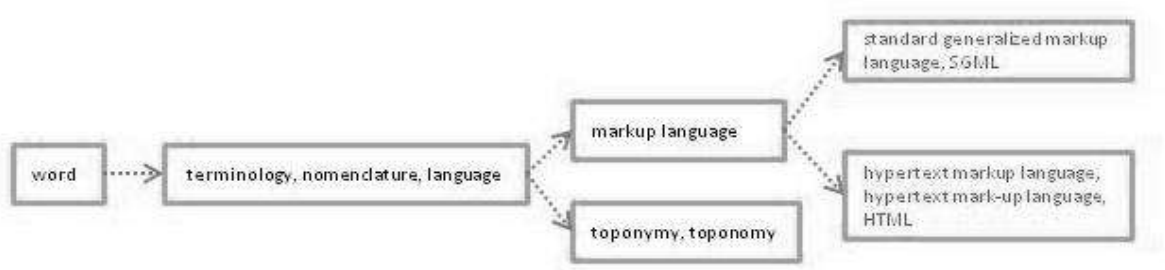

Diagram1. Examples of blurred lexical domain borders in linguistics

Although this situation is a general debate for all fields with an interdisciplinary nature, it is explicitly controversial in taxonomy process which is our case.

2. Methodology

The first phase of enlarging WN is recognising the current construction and detecting problematic situations as well as having an overview about the whole lexical property and the extension level of synsets. For carrying out this phase, we have extracted automatically all the variants associated with the synsets linked with *linguistics* or *grammar* domains (WN uses semi-automatically *IRST-Domains Hierarchy* marksto all its synsets[2]). It is worth to mention that, due to the semi-automatic process of assigning domain labels of synsets, there might be some synsets related to the linguistics domain which are not marked as linguistics or grammar terms, so it is not a highly accurate data. However, these assigned labels facilitate the process of extracting the majority of domain-specific synsets to have an overview to the domain structure. Then we have classified all these extracted terms[3] due to their hyperonym in order to derive a core base of basic concepts in linguistics for subsequent extension. We consider linguistics head synsets those linguistics synsets that are hyponyms of non-linguistics synsets and do not have any siblings related to a linguistics synset. We have considered them as domain border synsets, so called linguistics head-terms.

[2]See Magnini*et al*. (2000) for details

[3]Those terms from other domains, although might have some links to linguistic terms in upper taxonomic levels, are excluded. For instance the term *voice* in one of its meanings [*voice_6*] is hyponym of *communication* but it is a psychology term; so, this registered synset is excluded from our list.

These data are visualized in columns where the first column represents the head-term and the following columns represent hyponyms in different levels. Table (2) shows a sample of our data. In this table some hyponyms of the term "linguistics" and two other head-terms are shown in which any hyponym can be tracked to the first column:

e.g.: **orthoepy** is [HYPONYM of **phonology** (syn. phonemics) [HYPONYM of **descriptive linguistics** [HYPONYM of **linguistics]]]**

This linear form of representation gives us a possibility of collecting all synsets together in a same document and as a consequence we obtained a horizontal view to all synsets, illustrating which synsets are more developed or which synsets need more concerns. The facility of looking for a specific term and tracking its taxonomic route to its head is another positive point of this classification.

HEAD TERM	**HYPONYM1**	**HYPONYM2**	**HYPONYM3**
linguistics	neurolinguistics		
linguistics	pragmatics		
linguistics	semantics	deixis	
linguistics	semantics	formal semantics	
linguistics	semantics	cognitive semantics, conceptual semantics, semasiology	
linguistics	descriptive linguistics	phonology, phonemics	orthoepy
linguistics	descriptive linguistics	morphophonemics	
linguistics	prescriptive linguistics		
linguistics	etymology	lexicostatistics	
psycholinguistics			
speech perception			

Table2. WordNet representation sample based on hyponyms

The aim of this classification is to figure out what the current structure of linguistics synsets is and to what extent it can be useful for further extensions either in English or other languages.

In parallel, we have drawn synset diagrams whenever we needed more visualized data, particularly for complex synsets or for those that may require some structural modification (Diagram 2). Our modification proposals are in progress and they may change either by experts or informants in the future or just by elaborating our synsets. However, they can show how complicate synsets trees are in particular when it necessary to take profit of the multiple inheritance in order to represent adequately a given linguistic phenomena.. In this diagram dark blocks are our synsets proposals and white blocks are existing synsets in WN. The arrows here are used for connecting super-ordinate words to subordinates. Since in this diagram we proposed a restructuring in synsets arrangement all arrows are shown with continuous line; otherwise, if the relation between two synsets will remain without any change we used dotted lines.

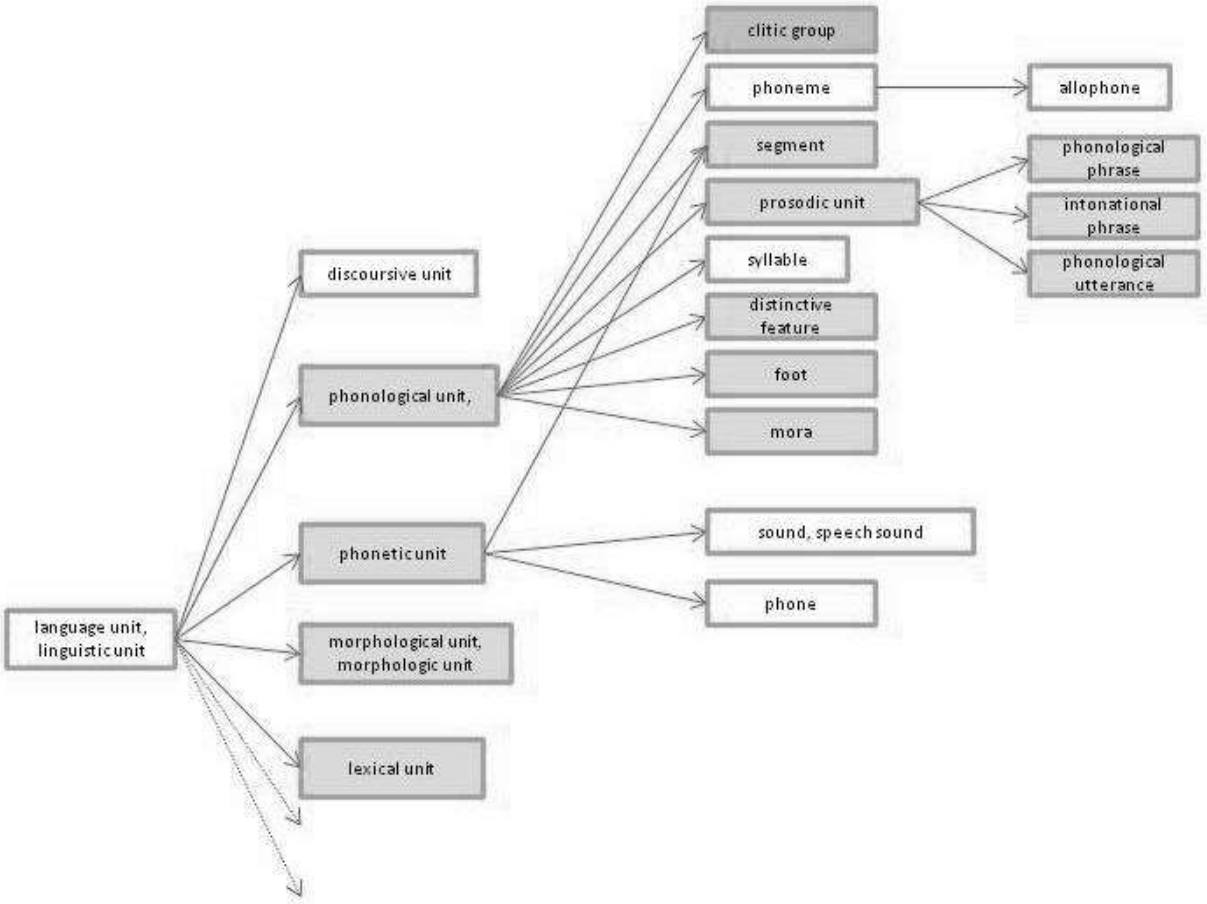

Diagram2. Simplified example of structural modification proposals for *language unit*synsets tree

3. Results and discussions

Our survey shows 1548 English records which are connected via hyponymy and hyperonymy relations and amongst these terms we have found 57 domain border synsets. Once we individuated the linguistics head-terms, we studied characteristics and common patterns of their hierarchical structure and we followed our analysis to the deep structure of our synsets. Besides, we have considered and studied corresponding glosses of each synset in terms of accuracy and to figure out the intended meaning of each.

Head-terms: We have recognized some ambiguities and problematic situations related to linguistics head-terms (table 3). The main characteristics of linguistics head-terms are:

a) According to the WN definitions, some of these head-terms are not specifically linguistics terms:

Ex.1. *damn, darn, hoot, red cent, shit, shucks, tinker's damn, tinker's dam* [definition] something of little value
It can be considered as a general term.

Ex.2. *rambler* [definition] a person whose speech or writing is not well organized
It can be considered as a general term.

Ex.3. *bodice ripper* [definition] a romantic novel containing scenes in which the heroine is sexually violated
It is more a literature term rather than linguistics.

Ex.4. *logomach, logomachist* [definition] someone given to disputes over words
It can be considered as a general term.

b) In terms of taxonomy relations some related terms are missed to be marked as linguistics terms, although they exist in WN:

Ex.1. *Armenian alphabet* and *Arabic alphabet* as hyponyms of *alphabet* are considered as linguistics head-terms, while the other types of alphabets like *Roman alphabet* or *Hebrew alphabet* or *phonetic alphabet* are not even labeled as linguistics terms.

Ex.2.*Postposition* as hyponym of *place, position* is considered as linguistics head-term, glossed as:

[definition] (Linguistics) the placing of one linguistic element after another (as placing a modifier after the word that it modifies in a sentence or placing an affix after the base to which it is attached)

While another related term *preposition* in one of its meanings is not even labeled as linguistics term, glossed as:

[definition] (Linguistics) the placing of one linguistic element before another (as placing a modifier before the word it modifies in a sentence or placing an affix before the base to which it is attached).

c) In terms of synsets structure we have found a lack of coordination amongst some head synsets. For instance, neurolinguistics and psycholinguistics are two related fields that each one studies one of the aspects of language. However, the former is classified as hyponym of *linguistics*, and the latter not.

Beside this problem, we can realize that if we do any change in terms of their classifications, a revision on glosses is necessitated for conceptual harmonization; as we can see below the differences between their glosses:

Psycholinguistics [definition] the branch of cognitive psychology that studies the psychological basis of linguistic competence and performance

Neurolinguistics [definition] the branch of linguistics that studies the relation between language and the structure and function of the nervous system

No.	Head-Terms	Hyperonym
1	communication	abstraction, abstractentity
2	punctuation	grouping
3	worddivision, hyphenation	division
4	neologism, neology, coinage[4]	invention
5	cryptograph	device
6	decoder	machine
7	tone	pitch
8	damn, darn, hoot, red cent, shit, shucks, tinker's damn, tinker's dam	worthlessness, ineptitude
9	lexis	cognition, knowledge, noesis
10	vocabulary, lexicon, mental lexicon	cognition, knowledge, noesis
11	speechperception	auditoryperception, soundperception
12	psycholinguistics	cognitivepsychology
13	linguistics[5]	science, scientific discipline
14	lexicology	linguistics, philology
15	soundlaw	law, natural law
16	languageunit, linguisticunit	part, portion, component part, component, constituent
17	line	text, textual matter
18	orthography, writing system	writing
19	Armenian, Armenianalphabet	alphabet
20	Arabic alphabet	alphabet
21	manual alphabet, fingeralphabet	alphabet
22	saying, expression, locution	speech, speech communication, spoken communication, spoken language, language, voice communication, oral communication

[4]In WN there are two synsets in linguistics domain stand for "*neologism, neology, coinage*", one which is recognized as a head-term is defined as "the act of inventing a word or phrase"; while the other is defined as "a newly invented word or phrase" which is hyponym of "word".
[5]In WN *linguistics* (syn. Philology) is not labeled as linguistics term, but as philology term classified in literature domain.

No.	Head-Terms	Hyperonym
23	non-standardspeech	speech, speech communication, spoken communication, spoken language, language, voice communication, oral communication
24	idiolect	speech, speech communication, spoken communication, spoken language, language, voice communication, oral communication
25	written symbol, printed symbol	symbol
26	monogram	symbol
27	sense, signified	meaning, significance, signification, import
28	sign[6]	gesture, motion
29	spiel, patter, line of gab	channel, communication channel, line
30	languagesystem	system, scheme
31	implosion	blockage, closure, occlusion
32	plosion, explosion	release, toneending
33	pronunciation	utterance, vocalization
34	speech	utterance, vocalization
35	rule, linguistic rule	concept, conception, construct
36	phylum	social group
37	French Academy	academy, honorarysociety
38	linguist, polyglot	person, individual, someone, somebody, mortal, soul, human
39	linguist, linguisticscientist	scientist
40	logomach, logomachist	disputant, controversialist, eristic
41	rambler	communicator
42	signer	communicator
43	linguisticprocess, language	highercognitiveprocess
44	linguisticprocess	human process
45	linguisticrelation	relation
46	imaginativecomparison	comparison
47	words per minute, wpm	rate
48	etymology[7]	history, account, chronicle, story
49	bodiceripper	romance
50	syntax, sentence structure, phrase structure	structure
51	slot	position, spatialrelation
52	grammaticalcategory, syntacticcategory	class, category, family
53	postposition	place, position
54	grammaticalmeaning	meaning, significance, signification, import
55	declension	class, category, family
56	conjugation[8]	set
57	conjugation	class, category, family

Table3. Linguistics domain border terms (head-terms) in WordNet

Linguistics synsets: While we have been classifying linguistics synsets, we figured out some controversial synsets that could be occurred by the domain characteristics or WN structure or simply inaccurate information. Some examples are given below:

a) Inaccurate definitions that contradict their current hyperonymy relations:

> Ex.1. *grammar*[definition] the branch of linguistics that deals with syntax and morphology (and sometimes also deals with semantics)
> [HYPONYM of **descriptive linguistics** [HYPONYM of **linguistics**]]

[6]In WN there are two synsets in linguistics domain stand for *sign*, one which is recognized as a head-term is defined as "a gesture that is part of a sign language"; while the other is defined as "a fundamental linguistic unit linking a signifier to that which is signified" which is hyponym of "*language unit, linguistic unit*".
[7]In WN there are two synsets in linguistics domain stand for *etymology*, one which is recognized as a head-term is defined as "a history of a word"; while the other is defined as "the study of the sources and development of words" which is hyponym of *linguistics*.
[8]In WN there is another synset standing for *conjugation* which is defined as "the inflection of verbs" and hyponym of "*inflection, inflexion*".

Ex.2. *descriptive linguistics* [definition] a description (at a given point in time) of a language with respect to its phonology and morphology and syntax and semantics without value judgments
[HYPONYM of **linguistics**]

Ex.3. *prescriptive grammar*[definition] a grammar that is produced by prescriptive linguistics
[HYPONYM of **grammar** [HYPONYM of **descriptive linguistics** [HYPONYM of **linguistics**]]]

None of these definitions are in accordance with their hyperonymsynsets; besides, these problematic synsets unfold a structural contradiction where *prescriptive grammar* is linked to *descriptive linguistics*. This situation also may refer to the lack of different notions of *grammar* in WN which is discussed later.

b) Heterogeneous taxonomy which might affect synsets extension:

There are some examples in WN in which we cannot figure out what classification pattern is employed or have been followed.

Ex.1. *apocope*[definition] abbreviation of a word by omitting the final sound or sounds
[HYPONYM of **abbreviation** [HYPONYM of **word form** [HYPONYM of **word** [HYPONYM of **language unit**]]]]

This synset is classified as hyponym of *abbreviation*, while some other types of *abbreviation* are specialized under different synsets and many of them (e.g. initialism, hybrid abbreviation, clipping) are missed.

What reason stands for categorizing *abbreviation* or *acronym* as a type of *word form* and *blend* not? This problem becomes more highlighted when we intend to extend these synsets. Considering the *word form* as it is glossed in WN, another question that comes to the mind is why *neologism* is not a type of *word form*, or why *acronym* or *abbreviation* cannot be considered as a *neologism*:

- *form, word form, signifier, descriptor*[definition] the phonological or orthographic sound or appearance of a word that can be used to describe or identify something

This heterogeneous structure do not provide any hint to solve this puzzle that due to which pattern we can add related synsets like *initialism* or *clipped word*.

Ex.2. *contraction*[definition] a word formed from two or more words by omitting or combining some sounds: `won't' is a contraction of 'will not'
[HYPONYM of **word** [HYPONYM of **language unit**]]

According to the given definition *contraction* can be better categorised as a type of *word form*. In other words, a connection between *word* and *contraction* is missed which is *word form*.

Ex.3. *loanblend, loan blend, hybrid*[definition] a word that is composed of parts from different languages

The same problem has been occurred for this synset. *Loan blend* and *loan word* are considered as hyponyms of *word*, while *blend* is classified as hyponym of *neologism*. The reasons are still unclear and might affect any attempt to extend current synsets.

HEAD TERM	HYPONYM1	HYPONYM2	HYPONYM3	HYPONYM4
languageunit, linguisticunit	word	form, word form, signifier, descriptor	acronym	
languageunit, linguisticunit	word	neologism, neology, coinage	blend, portmanteauword, portmanteau	
languageunit, linguisticunit	word	form, word form, signifier, descriptor	abbreviation	apocope
languageunit, linguisticunit	word	contraction		
languageunit, linguisticunit	word	loan word, loan		
languageunit, linguisticunit	word	loanblend, loan-blend, hybrid		

Table4. Some examples of heterogeneous taxonomy

c) Domain-dependent situations that entail specialists' revision:

Amongst those problematic situations we have found some synsets that need to be revised by experts in terms of their content accuracy.

Ex.1. *semantic role, participant role*[definition] (linguistics) the underlying relation that a constituent has with the main verb in a clause

In linguistic contexts there are occurrences in which *semantic role* and *semantic relation* are employed interchangeably. Moreover, there are some other terms considered as their synonyms, like *thematic role*, *case relation*, *theta role*, *deep case*, *semantic case*, (abbreviated form) *SR*, etc. (Dowty, 1989; Primus, 2008; Payne, 1997). This variety of synonyms along with differentiating between roles and relations in many WN synsets, have brought about a debatable issue. In this case, first, we need to collect all possible synonymous forms of this concept; second, we need to reach a consensus amongst experts or representative informants.

Another related issue might emerge during the extension process of its hyponyms, namely different types of *semantic role*, since there is a wide range of semantic role and thematic role proposals. Although, there are some common classifications, there is no concord and these concepts are defined from different perspectives due to the considerable number of schools and frameworks.

> The term "thematic relation" is frequently confused with theta role. Many linguists (particularly generative grammarians) use the terms interchangeably. This is because theta roles are typically named by the most prominent thematic relation that they are associated with. To make matters more confusing, depending upon which theoretical approach one assumes, the grammatical relations of subject and object, etc., are often closely tied to the semantic relations. For example, in the typological tradition agents/actors are tied closely to the notion of subject (S). [Source: Thematic relation, Wikipedia]

There are some more instances in WN that are in the same boat as *semantic role* and *thematic relation*. The most significant characteristic of these synsets is that their complexity is a consequence of the linguistics domain nature. For these instances it is crucial to clarify our approach, whether generalizing is appropriate or they need to be more elaborate.

> Ex.2. *grammar*[definition] the branch of linguistics that deals with syntax and morphology (and sometimes also deals with semantics)

It is interesting that this basic concept in linguistics has not been concerned adequately in WN. In linguistics, *grammar* has different notions that one of them is the study of syntactic and morphological rules. However in WN only one of them is considered. Besides, there is a variety of definitions from traditional view to modern thoughts that complicate any taxonomy effort for this concept. Diverse types of each notion and their functions can cause a complex node in WN extension. Considering this fact that *grammar* is a core concept in linguistics domain, it needs a high precision and some modification in the current structure of WN is inevitable.

d) Inaccurate taxonomy:

In WN There are some hyponym synsets that do not comply a *type-of* relationship. They can be better classified as meronyms (*part-of* relationship) or can be considered as instances.

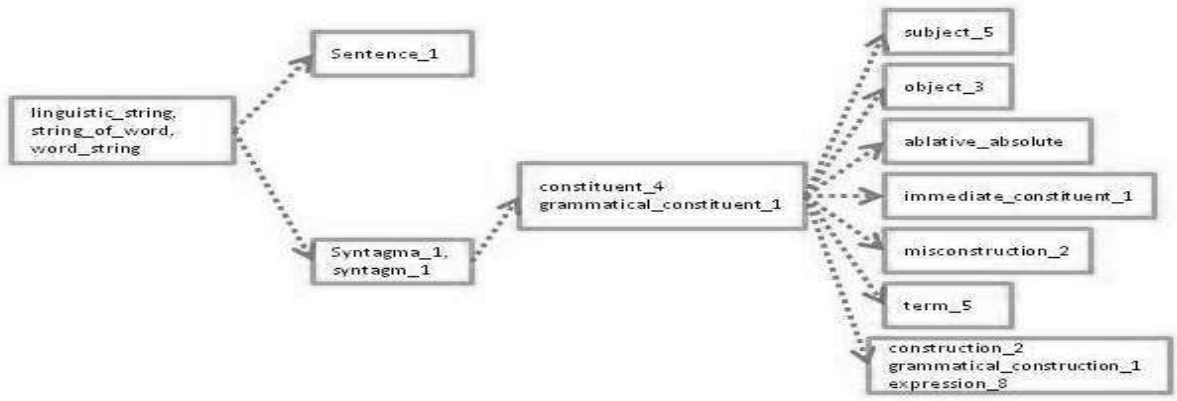

Diagram3. Some examples of inaccurate taxonomy in WordNet MCR (hyponymy vs. meronymy)

Diagram (3) shows some hyponyms of *linguistic string* as they are classified in WN. If we consider their given definitions we figure out some structural errors.

> *Syntagma*[definition] a syntactic string of words that forms a part of some larger syntactic unit

Constituent, grammatical constituent[definition] (grammar) a word or phrase or clause forming part of a larger grammatical construction
Construction, grammatical construction, expression[definition] a group of words that form a constituent of a sentence and are considered as a single unit

In all these given glosses it is mentioned that the concepts refer to a part of a larger unit which are classified as their hyperonyms. Furthermore, there is indeterminacy for *grammatical constituent* and *grammatical construction* that one cannot get if *grammatical constituent* forms part of a *grammatical construction* or vice versa.

Diagram (4) illustrates another mistake in WN classifications where *old man* is considered as hyponym of *dysphemism* and *Murphy's Law* as hyponym of *gnome*. In these examples not only some general words are labeled as linguistics terms, but also they can be better expressed as instances and not hyponyms.

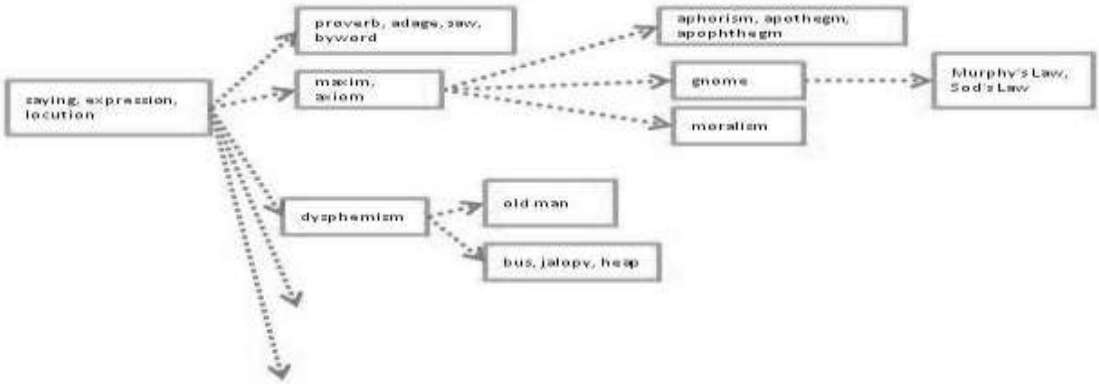

Diagram 4. Some examples of inaccurate taxonomy in WordNet MCR (hyponymy vs. instances)

Our aim was to build descriptive data on linguistics domain in WN for hierarchical extensions. Due to the fact that the WN domain label assignmentwas originally created by asemi-automatic process, obtaining fully accurate data of all linguistics lexical properties is not possible. We thus relied on maximal elicitation by *linguistics* and *grammar*labels in order to acquirestructural knowledgeabout current situation of the domain in WN. We further planned to rearrange our data in tables of hyponyms and hyperonymsin combination with diagrams to observe our data in a more visualized format.Although the main objective of the project was to extend synsets in terms of Spanish and Catalan variants, we have realized that there are some inconsistencies in English synsets that may affect any attempt to enlargecurrent synsets. Besides, in many cases, given glosses in WN not only are not useful for disambiguation, but also might bring about contradictions. This is highly crucial in multilingual extensions where original glosses are used for direct translations.

In WN we have found some inconsistencies and errors in linguistics domain that can be classified as:

1. Hierarchical contradictions amongst domain-specific synsets (caused by simple inheritance or lack of distinction between different notions)
2. Lack of coordination in domain border synsets (caused by mis-categorization)
3. Lack of precision in given glosses(semantic ambiguities)
4. Contradictions between definitions and hierarchical relations (caused by lack of precision or missed links)
5. Heterogeneous taxonomy
6. Inaccurate taxonomy (no distinction between types and instances or parts)
7. Lack of reliable domain label assignment

In spite of significant numbers of problematic situations, working on data elaboration along with increasing the precision of glosses and special concern to hierarchical expansion may reduce complexity of nodes. However, currently, such an inconsistent structure will cause serious problems in specialized information retrievals and domain-specific practices.

4. Conclusion

Taking into account subordinate and super-ordinate relations in WN MCR, we have found some problematic situations in which the extension process of linguistics synsets can be affected by complexities either in English or other languages. The results of our study show that these problematic situations are not exclusively related to the WN structure, but also some inherent characteristics of linguistics domain.

We believe that with current WN hierarchy, any multilingual extension in linguistics domain will be a hard task. Besides, existing problems might perform more complex situations in any automatic or semi-automatic process in domain-specific practices. Although our study is carried out in a specialized domain, the results are partially similar to some general studies about WN structure (Atserias*et al.*, 2005; Martin, 2003;Oltramari*et al.*, 2002). Our study shows that any problem in WN structure influences domain-specific synsets and much further the inherent characteristics of domains can create more crucial issues. We do agree with Oltramari*et al.* (2002:23) where he expressed that in WN "a serious taxonomy rearrangement is needed". Our study also proves that for multilingual WNs that are vastly based on Princeton WN structure, it is better to revise their policies for providing possibilities of structural modifications.

Acknowledgment

Weacknowledge Dr. MercèLorente and Dr. EsteveClua for their noteworthy collaboration,leading us to obtaina better knowledgeof the whole domain.

Thisworkwaspartiallysupportedbythe SKATER project (SpanishMinisterio de Economía y Competitividad, TIN2012-38584-C06-05).

References

Agirre, Eneko, and Philip Glenny Edmonds, eds. Word sense disambiguation: Algorithms and applications. Vol. 33.Springer, 2007.

Atserias, Jordi, et al. "A proposal for a Shallow Ontologization of WordNet."Proceedings of the 21th Annual Meeting of the Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN. Vol. 5. 2005.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe."The berkeleyframenet project."Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1.Association for Computational Linguistics, 1998.

Bentivogli, Luisa, Andrea Bocco, and EmanuelePianta. "ArchiWordnet: integrating Wordnet with domain-specific knowledge." Proceedings of the 2nd International Global Wordnet Conference. 2004.

Bodenreider, Olivier, and Anita Burgun."Characterizing the definitions of anatomical concepts in WordNet and specialized sources."Proceedings of the first global WordNet conference.Vol. 223. 2002.

Dowty, David R. "On the semantic content of the notion of ‘thematic role’."Properties, types and meaning.Springer Netherlands, 1989.69-129.

Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.

Gonzalez-Agirre, Aitor, EgoitzLaparra, and German Rigau. "Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base."Proceedings of the 6th Global WordNet Conference (GWC 2012). 2012.

Gonzalo, Julio, et al. "Indexing with WordNetsynsets can improve text retrieval." arXiv preprint cmp-lg/9808002 (1998).

Magnini, Bernardo, and Gabriela Cavaglia."Integrating Subject Field Codes into WordNet."LREC. 2000.

Martin, Philippe. "Correction and extension of WordNet 1.7."Conceptual Structures for Knowledge Creation and Communication.Springer Berlin Heidelberg, 2003.160-173.

Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

Oltramari, Alessandro, et al. "Restructuring WordNet's top-level: The OntoClean approach." Workshop Proceedings of OntoLex.Vol. 2. 2002.

Payne, Thomas Edward. Describing morphosyntax: A guide for field linguists.Cambridge University Press, 1997.

Primus, Beatrice. "Case, grammatical relations, and semantic roles." The Handbook of Case (2009): 261-275.

Ramanand, J., et al. "Mapping and Structural Analysis of Multi-lingual Wordnets." IEEE Data Eng. Bull. 30.1 (2007): 30-43.

Rosso, Paolo, et al. "Text categorization and information retrieval using wordnet senses." Proceedings of the Second International Conference of the Global WordNet Association". 2004.

Voorhees, Ellen M. "Using WordNet to disambiguate word senses for text retrieval." Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval".ACM, 1993.

Vossen, Piek. "Wordnet, eurowordnet and global wordnet." Revue française de linguistiqueappliquée 7".1 (2002): 27-38.

Zanotti, Cristian, Jorge Vivaldi, and MercéLorente."Upgrading WordNet: a Terminological Point of View."GWC 2012 6th International Global Wordnet Conference.