Towards Merging Common and Technical Lexicon Wordnets

Raquel AMARO[1] *Sara MENDES*[1,2]

(1) Center of Linguistics of the University of Lisbon,
Av. Professor Gama Pinto 2, 1649-003 Lisbon, Portugal

(2) Universitat Pompeu Fabra
Roc Boronat, 138, Barcelona, Spain

{ramaro,sara.mendes}@clul.ul.pt

ABSTRACT

The growing amount of available information and the growing importance given to the access to technical information enhance the potential role of NLP applications in enabling users to deal with information for a variety of knowledge domains. In this process, lexical resources are crucial.

Using and comparing already existent wordnets for common and technical lexica, we set up a basis for integrating these resources without losing their specific information and properties. We demonstrate their compatibility and discuss strategies to overcome the issues arrising in their merging, namely aspects concerning conceptual variation, subnet and synset merging, and the incorporation of technical and non-technical information in definitions.

As we are using models of the lexicon that mirror the organization of the mental lexicon, the accomplishment of this goal can provide insights on the type of relations holding between common lexical items and terms. Also, the results of integrating such resources can contribute to the better intercommunication between experts and non-experts, and provide a useful resource for NLP, particularly for tools simultaneously serving specialist and non-specialist publics.

KEYWORDS : wordnet, technical lexicon, common lexicon, merging.

1 Introduction

Since its appearance, Princeton WordNet (Miller *et al.*, 1990; Fellbaum, 1998) has been the main database used in NLP research and applications. With a strong psychological motivation, relational models of the lexicon have played a leading role in machine lexical knowledge representation. WordNet potential as a resource for NLP has also been explored in tasks typically associated to domain-specific information, such as systems for information extraction and document indexing, retrieval and preservation, and applications for technical domains such as Law (Peters *et al.*, 2006), Medicine (Elhadad & Sutaria, 2007) or Urbanism (Lacasta *et al.*, 2008). Although manifesting a number of shortcomings (Bodenreider *et al.*, 2003; Bodenreider & Burgun, 2002; Burgun & Bodenreider, 2001; Magnini & Strapparava, 2001), which reflect the lack of domain expertise of lexicographers developing it and the fact that it was not originally built for domain-specific applications (Smith & Fellbaum, 2004), WordNet potential to model technical lexica is made apparent by research showing that concept-based resources (ontologies, thesauri and wordnets) have great usability in teaching

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 147–160,
COLING 2012, Mumbai, December 2012.

(Mudraya, 2006; Fuentes, 2001; Robinson, 1989; Hutchinson & Waters, 1981) or improving mutual understanding between specialist and non-specialist publics (Elhadad & Sutaria, 2007).

The globalization of most activities, alongside technology development, produced significant changes both in the relation between specialist and non-specialist publics and in different aspects of terminology. Recent studies on the use of computer-based tools for technical domains point to a mismatch between technical lexical information incorporated in such tools and non-expert discourse employed by lay users (Slaughter, 2002; Tse & Soergel, 2003; McCray & Tse, 2003). Moreover, while the use of terms by professionals is expected to be subject to control by standardization efforts, the highly contextually dependent usage of terms by lay persons is much more difficult to capture. All these factors make the combination of common and specialized language resources more and more crucial. The importance of encoding domain-specific information in the WordNet model has also been remarked in the last years. In this context, there has been a considerable amount of research dedicated to the integration of domain-specific information into generic synsets (Magnini *et al.*, 2002; Vossen, 2001; Magnini & Cavaglià, 2000) or to the determination of the relevance of common lexicon synsets with respect to specific domains (Buitelaar & Sacaleanu, 2001). In parallel, there have been several efforts to develop dedicated wordnets for technical domains, such as Medicine (Buitelaar & Sacaleanu, 2002; Smith & Fellbaum, 2004), Geography (Giunchiglia *et al.*, 2009), or the Maritime domain (Roventini & Marinelli, 2004).

Research on integrating specialist taxonomies and common lexicon taxonomies (Pedersen *et al.*, 2010) has also been developed, as well as on merging domain-specific lexical resources with WordNet (Bosch, n/d). Following from this research, in this paper we compare a common lexicon wordnet with wordnets for ten technical domains built for Portuguese, setting up the bases for integrating both resources without losing specific information and properties. We expect the merging of technical and common lexica to raise several challenges, particularly regarding mismatches in sense differentiation and the encoding of relevant conceptual relations in models that reflect the organization of the mental lexicon. Accomplishing our goal will set the grounds for providing a useful resource to the research community, particularly to researchers working with domain-specific NLP tools simultaneously serving specialist and non-specialist publics.

2 Comparing common and technical lexicon wordnets

The work depicted in this paper is framed by research on wordnets developed for technical domains and on the characteristics of terms and specialized language, as well as on the interface between common and technical lexicon. We use two existing resources, a common lexicon wordnet – WordNet.PT[1] – and ten domain-specific wordnets for different technical domains – LexTec[2], and compare them with regard to different aspects, namely the amount of variants per synset, the type of relations used and the density of the network of relations. Both resources have been independently encoded and revised manually within the general framework of EuroWordNet. WordNet.PT (WN.PT) currently has about 18,000 lexical entries, covering all the main part-of-speech (PoS). We consider a subset of the database (15,000 lexical units) which covers the most salient daily life communication topics (food, clothing, sports, education, geography, transportation, etc.). LexTec covers more than 8,000 lexical units from all the main PoS and was built following the same development strategies

[1] WordNet.PT (Marrafa 2001, 2002), available online at http://www.clul.ul.pt/clg/wordnetpt/index.html.
[2] LexTec (Marrafa *et al.* 2009), available online at http://www.instituto-camoes.pt/lextec/.

and relations used in WN.PT. LexTec is balanced between ten different domains: Banking, Commerce, Economy and Business Management, Energy, Environment, Insurance, International Trade Law, Telecommunications, and Tourism.

We expect this comparison to allow us to identify contrasts and similarities between common and technical lexicon, which not only can be contrasted to previous work but also can be used for designing sound strategies for integrating both resources without losing their specific information and properties. This is not a trivial task, particularly since the common lexicon tends to reflect and integrate popular lexicalizations in specific domains. The taxonomies reflecting popular lexicalizations have been argued to be significantly less elaborate at both the upper and lower levels than in the corresponding technical lexica (Medin & Aran, 1999). Also, popular terms tend to cover a larger range of referent types than technical terms, i.e. to be less precise, while others may cover only part of the extension of their technical counterparts. The information in Table 1 allows for identifying similarities and differences between technical and common lexica regarding phenomena such as PoS distribution and synonymy.

		N	V	Adj.	PN	Average
WN.PT	lexical entries (%)	74.3%	8.4%	8.5%	8.9%	
	synsets (%)	73.6%	8.5%	9.3%	8.7%	
	average variant/synset	1.28	1.26	1.16	1.30	**1.27**
LexTec	lexical entries (%)	77.1%	3.6%	3.3%	16.0%	
	synsets (%)	77.5%	5.2%	5.0%	12,4%	
	average variant/synset	1.71	1.18	1.14	2.21	**1.71**

TABLE 1 – PoS distribution and density in terms of synonymy relations of WN.PT and LexTec

In terms of PoS distribution, the larger percentage of nominal nodes in LexTec (77.5% of nouns and 12.4% of proper nouns), and consequent smaller percentage of the other PoS, is consistent with what is generally assumed, specifically that the description of a given domain is mainly constituted by nominal expressions (Cabré, 1998: 36). However, when it comes to the ratio between variants and synsets, technical lexica would be expected to have a lower ratio, since the "form and content of terms tends towards an unambiguous relationship" (Cabré, 1998: 116). Despite the precision characteristic of specialized discourse, the existence of synonymy in terminology has long been acknowledged (Daille *et al.*, 1996; Freixa, 2002; Cabré, 2008; Montiel-Ponsoda *et al.*, 2011, Aguado-de-Cea & Montiel-Ponsoda, 2012). Moreover, the integration of English terms in the terminology of other languages, sometimes co-existing with variants in these languages, is also to be considered. Table 1 confirms this and makes apparent that synonymy is a distinctive feature of the technical lexicon with regard to the common lexicon. To verify whether these characteristics apply generally and equally to different domains, we looked into the numbers regarding individual domains (Table 2).

Table 2 presents the PoS distribution and the density of synonymy relations for 6 technical wordnets. These regard specifically chosen domains: Banking; Environment; Energy; Telecommunications; Construction; and Tourism. The first four are more classical knowledge domains, rich in terminology. Construction was selected as it includes terms from Civil Engineering, Architecture, but also lexicalizations of traditional construction methods and materials. As to Tourism, its selection was motivated by the fact of it being a more recent and interdisciplinary area, including aspects of Social Sciences, Economics and Commerce, but also very familiar to lay publics, as they interact directly and regularly with tourism products.

		N	V	Adj.	PN	Average
Environment	lexical entries (%)	66.5%	3.0%	7.0%	23.6%	
	synsets (%)	67.5%	4.9%	11.0%	16.6%	
	average variant/synset	1.75	1.07	1.13	2.53	1.78
Energy	lexical entries (%)	78.8%	2.5%	3.8%	14.8%	
	synsets (%)	80.2%	4.3%	6.2%	9.3%	
	average variant/synset	1.77	1.01	1.10	2.87	1.80
Telecom	lexical entries (%)	77.9%	3.8%	0.9%	17.4%	
	synsets (%)	82.1%	3.3%	1.5%	13.1%	
	average variant/synset	1.98	2.38	2.76	2.76	2.08
Banking	lexical entries (%)	87.5%	2.0%	1.1%	9.4%	
	synsets (%)	87.0%	3.8%	2.3%	7.0%	
	average variant/synset	2.19	1.12	1.10	2.94	2.17
Construction	lexical entries (%)	83.2%	5.1%	5.2%	6.5%	
	synsets (%)	83.5%	6.8%	6.5%	3.2%	
	average variant/synset	1.49	1.12	1.20	3.00	1.49
Tourism	lexical entries (%)	48.1%	4.5%	4.1%	43.4%	
	synsets (%)	50.9%	5.5%	5.2%	38.3%	
	average variant/synset	1.34	1.15	1.11	1.61	1.42

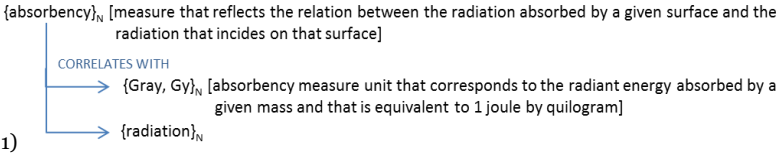
TABLE 2 – PoS distribution and synonymy relation density per technical domain

PoS distribution in these individual domains reflects the general tendency of technical lexica: nominal nodes are predominant, although the proportion between nouns and proper nouns can be considerably different, ranging from 87% of common nouns and 7% of proper nouns (Banking) to 51% of common nouns and 38% of proper nouns (Tourism). The ratios of variants per synset also show significant differences, ranging from an average of 2.17 (Banking) to an average of 1.42 (Tourism). Construction and Tourism are the two domains with the lower ratio, hence closer to WN.PT in this regard. These numbers seem to indicate a higher proximity of these technical domains to common lexicon, which is not surprising since non-specialist speakers interact regularly and directly with contents from these domains.

2.1 Lexical-conceptual relations and network density

WordNet.PT and LexTec are lexical-conceptual databases built within the same theoretical framework, using the same set of relations (exceptions being MANNER and CAUSE relations, not used in LexTec). In what concerns the relations used, LexTec presents a higher percentage of HYPERONYMY (24.8% VS. 19.4%), INSTANTIATION (8% VS. 4.2%) and CORRELATION (18.7% vs. 9.3%) relations. In contrast, WN.PT has a higher incidence of MERONYMY and HOLONYMY relations (10.4 vs. 6.1% and 10.1% vs. 5.6%, respectively), and IS A CHARACTERISTIC OF/HAS AS A CHARACTERISTIC relation (12.2% vs. 3.9%). Some of these differences are directly related to the PoS distribution in both resources: INSTANTIATION is the relation linking proper nouns to the nominal nodes they instantiate, thus the higher incidence rate of this relation in LexTec. With regard to the IS A CHARACTERISTIC OF/HAS AS A CHARACTERISTIC relation, it establishes a link between nominal nodes and their salient and definitional characteristics, denoted by adjectives (see Mendes (2009)). The higher incidence of this relation in WN.PT is not independent from the higher proportion of adjective nodes in this resource. The higher percentage of CORRELATION relations in LexTec is also expected since "concepts are related to other concepts in the specific field they together constitute" (Cabré, 1998:116). Also, since

nominal expressions are more common in technical language, it is more likely for this type of relations to be more relevant in technical wordnets given that there are not many technical verbs mediating the nodes in domain-specific wordnets, as shown in (1).



(1)

Also, it is predictable that HYPERONYMY relations have a strong weight on the overall number of relations in technical lexica, since the specification of concepts, expressed in wordnets through HYPERONYMY/HYPONYMY relations, is known to be quite productive in terminology (Daille *et al.*, 1996; Freixa, 2002; Burgun & Bodenreider, 2001; Roventini & Marinelli, 2004; Cabré, 2008; Montiel-Ponsoda *et al.*, 2011; among others). Moreover, it is generally assumed that when a term, for some reason, becomes part of the common lexicon, it usually loses some of its technical meaning, denoting a broader, less specialized concept (Aguado-de-Cea & Montiel-Ponsoda, 2012; Meyer & Mackintosh, 2000). Being so, the less specification of the concepts denoted is bound to be correlated to shallower HYPERONYMY trees.

Finally, there is also a significant difference in terms of the density[3] of these networks: WN.PT presents a density of 4.5; while that of LexTec amounts only to 3.2. However, we feel that no strong claims can be made in this respect based on this data since WN.PT is a single wordnet, which potentiates the number of nodes available for linking, while for technical language we are working with a set of separate wordnets, each corresponding to a given domain and whose individual size is far from being close to that of WN.PT.

3 Merging technical and common lexicon wordnets

The merging of technical and common lexica raises several issues. Contrasts concerning sense differentiation and the establishment of the relevant semantic and conceptual relations with other lexical-conceptual units are bound to arise since these derive directly from the meaning of each unit. And yet, merging common and technical lexica is unquestionably linguistically motivated since specialists always maintain the ability to use common lexicon for communicating with non-specialist speakers, or even with other specialists when terminology for new concepts does not exist (Cabré, 1998), thus never entirely replacing common lexicon with specialized language. This way, the study of the issues involved in the merging of technical and common lexica in models mirroring the organization of the mental lexicon, besides contributing to address a growing need in the scientific community and provide it with a useful and differentiated language resource, can also provide some insights on the type of relations existing between these differentiated subsets of the lexicon. In this section, we present a typology of cases we are confronted with when merging two resources with the characteristics described earlier, illustrating each situation with examples from the databases, and focusing on the issues to be accounted for.

[3] Network density is calculated by summing all the relations encoded in the database and dividing them by the number of synsets represented.

3.1 Conceptual variations

Sense discrimination covering domain-specific concepts and common lexicon can result in polysemy and semantic overlapping (Sagri *et al.* 2004, Pederson *et al.* 2010, Chen *et al.* 2011). Differences in the ontological nature of the concepts are expected and can range between what we will call compatible, semi-compatible and incompatible conceptual variations.

Compatible conceptual variations correspond to the cases where the concept denoted by technical synsets is more precise and specialized, but otherwise similar to and compatible with the concept denoted by a corresponding common lexicon synset (see (2))[4]:

(2) **WN.PT**: {gasóleo}$_N$ [liquid fuel, oil derivative, used in diesel engines] (diesel)
HYPONYM OF {combustível}$_N$ (fuel)
{derivado}$_N$ (derivative)

LexTec: {gasóleo, <u>diesel$_{english}$</u>}$_N$ [liquid fuel, composed mainly of hydrocarbons and obtained by oil distillation, brown colored, with an intense smell, denser and less inflammable than gasoline, used in compression combustion engines] (diesel)
HYPONYM OF {combustível}$_N$ (fuel)
{derivado do petróleo}$_N$ (oil derivative)

Semi-compatible conceptual variations include cases like that of (3), where concept specialization entails intermediary hyperonyms – expressing technical specification not existing in the common lexicon –, the concept denoted by both technical and common lexicon synsets being nonetheless the same.

(3) **WN.PT**: {ladrilho, mosaico}$_N$ [flat building material, square or rectangular, typically made of ceramic, used to cover walls and floor] (tile)
HYPONYM OF {material de construção}$_N$ (building material)

LexTec: {ladrilho, mosaico}$_N$ [covering that consists of one piece, typically a rectangular ceramic plate, that is applied on the floor or on the wall](tile)
HYPONYM OF {revestimento}$_N$ (covering)
HYPONYM OF {material de construção}$_N$ (building material)

Incompatible conceptual variations, in (4), refer to cases where the concepts denoted by technical and common lexica, though closely related, are not the same, as made apparent by the hyponymy chain.

(4) **WN.PT**: {sótão}$_N$ [floor of a building, with a low ceiling, immediately under the roof] (attic, garret, loft)
HYPONYM OF {piso, andar}$_N$ (floor, level, story)
HYPONYM OF {parcela}$_N$ (parcel)

LexTec: {sótão}$_N$ [annex situated immediately under the roof of a building, typically considered for storage] (attic)
HYPONYM OF {dependência}$_N$ (annex)
HYPONYM OF {construção}$_N$ (construction)
HYPONYM OF {estrutura}$_N$ (structure)

These three types of possible situations call for different merging strategies. Cases like (2) can be almost straightforwardly merged, involving only the use of labels already available in the WordNet model (see Section 3.3). In the case of semi-compatible conceptual variations,

[4] The information in the examples provided is given in the following format: {synset}$_{POS}$ [gloss] (English translation). Underlined expressions correspond to variants associated to usage information, given in subscript characters, such as registry or origin (in the case of borrowed expressions, for instance).

besides the merging of synsets, it is also necessary to assure an adequate subnet merger to integrate both technical and common lexica relations and nodes without information loss. Finally, in the case of incompatible conceptual variations, it is not possible to perform a direct merging, since the concepts denoted are distinct. Being so, these cases should be treated as any other case of homonymy in wordnets, where each concept denoted corresponds to a separate node in the network, as suggested by Pedersen *et al.* (2010:3184).

However, as illustrated by (4), the relation between common and technical concepts is a very salient relation, which moreover can provide useful information both for NLP applications and human users. Considering the relations available in the WordNet model, the closest candidate to link these synsets would be the NEAR SYNONYMY relation[5], but this relation fails to cover this particular situation. Near synonyms are lexical units that do not pass the tests that motivate their belonging to the same synset: near synonyms are necessarily co-hyponyms, and have a stronger connection with each other than with their other co-hyponyms, which is not the case here. In this case, there are two different denotations (concepts), related to two different ways of conceiving and eventually lexicalizing a referent that can be, more often than not, the same. For instance, to use the example in (4), any utterance in which *sótão* (attic) occurs will refer to the upper part of a building, independently of whether the speaker is using the technical or common lexicon concept. This way, what seems to be at stake here is a shared reference, i.e. some type of co-reference relation, which requires a further and deeper study of this phenomenon and its properties.

3.2 Subnet variation and merging

One of the difficulties expected in the process of merging technical and common lexicon wordnets concerns the differences in the networks of relations established between compatible and semi-compatible synsets, which derive from conceptual variation. The example below illustrates this situation considering the synset {combustível} (fuel) and its relations in WN.PT (in black) and in LexTec (in orange)[6].

The graphical representation presented in Figure 1 illustrates the adaptations necessary, namely the overlapping, duplication and marking of the compatible synsets in both databases, as described in the literature (Roventini & Marinelli, 2004; Roventini *et al.*, 2000; Magnini & Speranza, 2001), to assure the visualization of each net individually. Roventini & Marinelli (2004) present a strategy to connect the databases through plug-in relations, considering that all upward relations (hyperonymy) from a given plugged-in node are taken from the common lexicon wordnet, while all other relations are taken from the technical one (Roventini & Marinelli, 2004: 196). This strategy does not prevent information loss, though.

To assure that all the relations in WN.PT and LexTec are considered, all relations are added, including those involving semi-compatible synsets (like {combustível}$_N$ (fuel) and {gás natural}$_N$ (natural gas)) and horizontal relations (such as ROLE relations) originally only present in one of the subnets. This strategy goes along the lines of the work of Bosch (n/d), although this author defends a partial merging that protects technical acceptions over general ones. In the strategy put forth in this work we do not argue for a proeminence of one resource

[5] We refer here to NEAR SYNOMYNY relation as defined in Vossen (2002:19). Near synonyms with different PoS are linked in EuroWordNet by the xpos_NEAR_SYNONYMY relation.

[6] The complete network of relations for these synsets in WN.PT and LexTec are available in http://www.clul.ul.pt/clg/wordnetpt/index.html and in http://www.instituto-camoes.pt/lextec/, respectively.

over the other, but rather outline a method for combining both resources maintaining their characteristics and properties and avoiding information loss. When overlapping, the relations and respective target nodes are analyzed regarding conceptual variation, in a new iteration of the process described above. In what concerns subtypes of relations (such as subspecified ROLE *vs.* ROLE PATIENT, for instance) finer-grained relations replace general ones. To maintain the possibility of separating the subnets merged, technical nodes have to be labeled, as well as individual lexicalizations in each synset, distinguishing lexical items pertaining to technical language, as described in the next section.

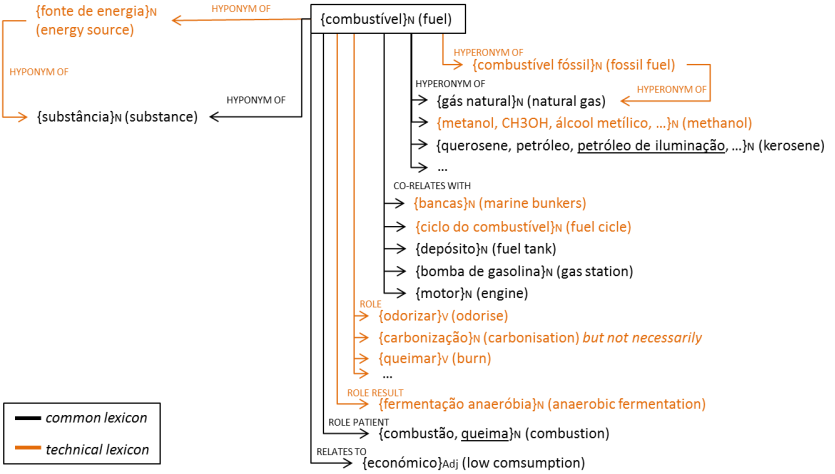


FIGURE 1 - merged network of relations for {combustível} (fuel)

3.3 Synset merging

The merging of compatible and semi-compatible synsets, besides requiring the insertion of intermediary hyperonyms when necessary, can also involve the treatment and encoding of lexical units in each set of synonyms. In the type of merging targeted in our work, lexical units can pertain both to common and technical lexica, and this information has to be overtly stated. EuroWordNet, the framework within which the resources considered in this paper have been developed, already allows for the tagging of technical lexical units through usage labels (Vossen 2002:106). This way, in merged synsets – which are part of both common lexicon and technical subnets – all lexical units have to be individually marked, as exemplified in (5), where *C* stands for common lexicon and *E* stands for the technical domain of Energy.

(5) a. {combustível$_{C,E}$}$_N$ (fuel)
b. {querosene$_{C,E}$, petróleo$_C$, petróleo de iluminação$_E$, petróleo iluminante$_E$}$_N$ (kerosene)
c. {combustão$_{C,E}$, queima$_E$}$_V$ (combustion)

The marking of the different lexical units requires only the definition of usage labels to include all the technical domains considered, as well as the common lexicon. With regard to making decisions involving the use of specific lexical units in common and technical

contexts, this calls for corpora analyses and experts' advice, as discussed in Burgun & Bodenreider (2001), Magnini *et al.* (2002) or Smith & Fellbaum (2004), among others.

3.4 Incorporating common and technical information in definitions

Wordnets are characterized by having synsets as their basic unit and by the fact that the meaning of each unit is determined by its relations in the network. This way, in wordnets, definitions or glosses constitute additional information used to aid human users, to provide examples of use or complementary information considered useful, especially when nodes are not available for linking. Even though not part of the WordNet model, definitions can provide helpful information in many situations, both to human users and to NLP tools. Considering this, in this section we focus on strategies to incorporate common and technical information in definitions avoiding potential incongruities and leaving open the possibility of using either subnet (common or technical) individually, in a process that can be developed automatically (Chen *et al.*, 2011).

Our basic methodology consists in considering the lexical-conceptual relations encoded in wordnets to build definitions. Beginning with the common lexicon subnet, the definition starts by stating the hyperonym and then all the horizontal relations which correspond to definitional properties of the concept. Non-definitional relations are disregarded, namely hyponymy relations and all relations marked as reversed. CO-RELATES WITH relations are typically accessory (i.e. not essential to the definition of the meaning of the lexical unit), although sometimes they provide relevant information, as illustrated in (6). The same procedure is applied with regard to the technical subnet. This methodology results in some level of repetition, as shown in the example below, which can be avoided by controlling the information in common in the first and second part of the definition and omitting it from the second part. The parts regarding the common and technical lexicon are separated by semi-colons and, following the previous color scheme, technical information is presented in orange. For purposes of explanation, redundant information is presented in brackets:

(6) a. **WN.PT definitional relations for {tile}$_N$:** IS HYPONYM OF {building material}$_N$, HAS AS A CHARACTERISTIC {flat}$_{Adj}$ and {glazed}$_{Adj}$, CO-RELATES WITH {wall}$_N$ and {floor}$_N$, IS INVOLVED IN {tile}$_V$

b. **LexTec definitional relations for {tile}$_N$:** IS HYPONYM OF {covering}$_N$, CO-RELATES WITH {wall}$_N$ and {fixative mortar}$_N$, IS INVOLVED IN {pave}$_V$, {paving}$_N$, {lay}$_V$, {laying}$_N$, {tile}$_V$ and {untile}$_V$

c. **definition:** flat and glazed building material used to cover walls and floor; constitutes a covering that is paved, layed or tiled (to walls and floor) with fixative mortar

This two-part definition can function for both subnets individually: in the case where redundant information is maintained, it is just a matter of presenting the first or the second part of the definition for an individual visualization of the common or the technical subnet, respectively; where redundant information is avoided, the first part of the definition is presented for common lexicon subnet visualization and the whole definition is presented for technical lexicon subnet visualization. In our perspective, it is preferable to maintain the redundant information, since on the one hand the individual visualization of technical subnets becomes more coherent, and on the other the visualization of both parts of the definition simultaneously can help to obviate the conceptual variations between common and technical lexica.

4 Final remarks and future work

Following from previous research on relational models of the lexicon and on the interface between common and specialized languages, this paper presents a comparison of existing wordnets for common and technical lexica for Portuguese, focusing on their contrasts and similarities, to set the basis for a merging that preserves the specific information and properties of these resources. We discuss strategies to overcome the issues to be accounted for in the merging of these particular lexica, namely in what concerns conceptual variation, subnet and synset merging and the incorporation of technical and non-technical information in the definitions associated to each node.

As pinpointed throughout the paper, several issues deserve nonetheless further attention and constitute topics for future work. In particular, concerning semi-compatible synsets, the number of intermediary hyperonyms allowed while preserving a compatible conceptual variation between common and technical synsets, directly related to the study of the depth of hyperonymy trees in both lexica, needs to be addressed and motivated. Also, research on possible co-reference relations between incompatible yet related synsets requires further work, possibly applying strategies of corpora analysis, and expert and non-expert users surveys, as suggested by Smith & Fellbaum (2004). The validation of the usage of specific lexical units in common and technical lexica is a related issue, which can be addressed using this kind of approaches. Finally, and based on the strategies defined and presented in this paper, future work naturally comprises the implementation of methods for collecting and merging synsets from both resources automatically or semi-automatically, based on approaches like the ones put forth, for instance, by Vossen (2001), Buitelaar & Sacaleanu (2002) or Tse & Soergel (2003), this way assuring a cost-efficient feasibility of the merging.

Acknowledgments

The work presented here has been supported by Fundação para a Ciência e a Tecnologia post-doctoral fellowships SFRH/BPD/75904/2011 and SFRH/BPD/79900/2011. The authors also whish to thank the remarks of the anonymous reviewers that contrubuted to the final version of this paper.

References

Aguado-de-Cea, G. & Montiel-Ponsoda, E. (2012). Term variants in ontologies. In *Proceedings of the 30th International Conference of AESLA*, April, Spain (pp. 19-31).

Bodenreider, O. & Burgun, A. (2002). Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In *Proceedings of the First Global WordNet Conference*, Mysore, India (pp. 223-230).

Bodenreider, O., Burgun, A. & Mitchell, J.A. (2003) Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. In *Studies in Health Technology and Informatics*, 95, 379-384.

Pedersen, B., Nimb, S. & Braasch, A. (2010), Merging specialist taxonomies and folk taxonomies in wordnets - a case study of plants, animals and foods in the Danish wordnet. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta (pp. 223-230).

Buitelaar, P. & Sacaleanu, B. (2001). Ranking and selecting synsets by domain relevance. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources,* Pittsburgh, Pennsylvania.

Buitelaar, P. & Sacaleanu, B. (2002). Extending synsets with medical terms. In *Proceedings of First Global WordNet Conference,* Mysore, India.

Burgun, A. & Bodenreider, O. (2001). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources,* Pittsburgh, Pensylvania (pp. 77-82).

Cabré, M. T. (2008). El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología. In *IBÉRICA* 16, 9-36.

Cabré, T. (1998). *Terminology. Theory, methods and applications*, Amsterdam: John Benjamins Publishing.

Chen, R.-C., Bau, C.-T. & Yeh, C.J. (2011) Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques. In *Applied Soft Computing*, 11(2), March 2011.

Daille, B., Habert, B., Jacquemin, C. & Royauté, J. (1996). Empirical observation of term variations and principles for theirdescription. In *Terminology* 3 (2), 197-257.

Elhadad, N. & Sutaria, K. (2007). Mining a Lexicon of Technical Terms and Lay Equivalents. In *ACL BioNLP Workshop Proceedings*, Prague, Czech Republic (pp. 49-56).

Fellbaum, C. (1998) (Ed.). *WordNet: an electronical lexical database*. Cambridge: The MIT Press.

Freixa, J. (2002). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient*. Doctoral dissertation, Universitat Pompeu Fabra, Barcelona.

Fuentes, A. C. (2001). Lexical Behaviour in Academic and Technical Corpora: Implications for ESP Development. In *Language Learning & Tecnhology*, vol.5, nº 3, 106-129.

Giunchiglia, F., Maltese, V., Farazi, F. & Dutta, B. (2009). *GeoWordNet: a resource for geo-spatial applications*. Technical Report #DISI-09-071: http://eprints.biblio.unitn.it/1777/1/071.pdf.

Hutchinson, T. & Waters, A. (1981). Performance and competence in ESP. In *Applied Linguistics* 2/1.

Lacasta, J., Nogueras-Isso, J., Zarazaga-Soria, P. & Muro-Medrano, R. (2008). Generating an urban domain ontology through the merging of cross-domain lexical ontologies. In *Conceptual Models for Urban Practitioners*, Bologna: Società Editrice Esculapio, pp. 69-84.

Magnini, B. & Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000),* Athens, Greece.

Magnini, B. & Speranza, M. (2001). Integrating Generic and Specialized Wordnets. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP-2001, Tzigov Chark, Bulgaria (pp. 149-153).

Magnini, B. & Strapparava, C. (2001). Using WordNet to improve user modelling in a web document recommender system. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources,* Pittsburgh, Pennsylvania.

Magnini, B., Strapparava, C., Pezzulo, G. & Gliozzo, A. (2002). Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet. In *Proceedings of First Global WordNet Conference,* Mysore, India (pp. 146-154).

Marrafa, P. (2002): "The Portuguese WordNet: General Architecture and Semantic Internal Relations", *DELTA*.

Marrafa, P. (2001): *WordNet do Português - Uma base de dados de conhecimento linguístico*, Instituto Camões.

Marrafa, P., R. Amaro, R. P. Chaves, S. Lourosa & S. Mendes (2009), *LexTec – Léxico Técnico do Português: Ambiente, Banca, Comércio, Construção, Energia, Seguros, Turismo, Telecomunicações, Direito Comercial Internacional e Economia e Gestão de Empresas*, Instituto Camões.

McCray, A.T. & Tse, T. (2003). Understanding search failures in consumer health information systems. In *Proceedings of the American Medical Informatics Symposium* (pp. 430-434).

Medin, D.L. & Adran, S. (1999) (Eds.). *Folkbiology*. Cambridge: The MIT Press.

Mendes, S. (2009) *Syntax and Semantics of Adjectives in Portuguese: analysis and modelling*, PhD thesis, Universidade de Lisboa.

Meyer, I. & Mackintosh, K. (2000). L'étirement du sens terminologique: aperçu du phénomène de la déterminologisation. In H. Béjoint & P. Thoiron (Eds.). *Le Sens en terminologie*, Lyon, France: Presses de l'Université Lumière Lyon 2.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K.J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3*(4), 235-44.

Montiel-Ponsoda, E., Aguado-de-Cea, G. & McCrae, J. (2011). Representing term variation in lemon. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence (TIA 2011*), Paris, France (pp. 47–50).

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. In *English for Specific Purposes* 25, Elsevier, 235-256.

Peters, W., Sagri, M., Tiscornia D. & Castagnoli, S. (2006). The LOIS Project. In *Proceedings of Linguistic Resources Evaluation Conference (*LREC'06*),* Genova, Italy (pp. 23-27).

Robinson, P.J. (1989). A rich view of lexical competence. In *ELT Journal*, vol. 43/3, Oxford University Press, 274-282.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B., Bertagna, F. (2000). ItalWordNet: a large semantic database for Italian. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC'00), Athens, Greece.

Roventini, A. & Marinelli, R. (2004). Extending the Italian WordNet with the Specialized Language of the Maritime Domain. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (Eds.). *Proceedings of the Global WordNet Conference 2004 (GWC 2004)*, (pp. 193-198). Brno: Masaryk University.

Sagri, M.T., Tiscornia, D. & Bertagna, F. (2004). Jur-WordNet. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (Eds.). *Proceedings of the Global WordNet Conference 2004 (GWC 2004)*, Brno: Masaryk University (pp. 305-310).

Slaughter, L. (2002). *Semantic relationships in health consumer questions and physicians' answers: a basis for representing medical knowledge and for concept exploration interfaces*. Doctoral dissertation, University of Maryland at College Park.

Smith, B. & Fellbaum, C. (2004). Medical WordNet: a New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Tse, T. & Soergel, D. (2003). Procedures for mapping vocabularies from non-professional discourse. A case study: 'consumer medical vocabulary'. In *Proceedings of the Annual Meeting of the American Society for Information*.

van den Bosch, A. (no date), *Merging domain thesauri with a generic Wordnet: A case study with the Dutch Army Museum*, http://www.rnaproject.org/whitepapers.aspx

Vossen, P. (2001). Extending, trimming and fusing wordnet for technical documents. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources,* Pittsburgh, Pennsylvania.

Vossen, P. (2002) (Ed.). *EuroWordNet General Document*, EuroWordNet Project LE2-4003 & LE4-8328 report, University of Amsterdam, http://vossen.info/docs/2002/EWNGeneral.pdf.