Making Wordnet Mappings Robust*

Jordi Daudé y **Lluís Padró**
TALP Research Center
Universitat Politécnica de Catalunya
Barcelona.
{daude,padro}@lsi.upc.es

German Rigau
IXA Group
Euskalerriko Unibersitatea
Donosti.
{rigau}@si.ehu.es

Resumen: La construcción de los recursos necesarios para el procesamiento semántico a gran escala es una tarea que implica a grandes grupos de investigación durante largos periodos de desarrollo. Los resultados de estos proyectos son, normalmente, grandes y complejas estructuras semánticas, no compatibles con otros recursos desarrollados en proyectos y esfuerzos anteriores. Para mantener la compatibilidad entre wordnets de distintas lenguas y versiones es fundamental disponer de una herramienta automática de alta precisión. Este artículo presenta una validación precisa, tanto cuantitativa como cualitativa de la metodologia usada por (Daudé, Padró, and Rigau, 2001) para conectar dos versiones distintas de WordNet. Comprobamos la precisión de la técnica usándola para enlazar una versión de WN con ella misma, lo que permite no sólo la evaluación cuantitativa, sino también un estudio cualitativo de los casos de error y un afinado del algoritmo.
Palabras clave: Mapping de ontologías, WordNet, Etiquetado por relajación

Abstract: Building appropriate resources for broad–coverage semantic processing is a hard and expensive task, involving large research groups during long periods of developement. The outcomes of these projects are, usually, large and complex semantic structures, not compatible with resources developed in previous projects and efforts. To maintain compatibility between wordnets of different languages and versions, past and new, it is fundamental to dispose of a high accurate tool. In this paper we present an accurate, quantitative and qualitative validation of the methodology used by (Daudé, Padró, and Rigau, 2001) to map two WordNet versions. We check the accuracy of the technique by applying it to map a WN version onto itself, which enables not only quantitative evaluation but also a qualitative study of the error cases and algorithm tuning.
Keywords: Omtology mapping, WordNet, Relaxation labelling

1 Introduction

Using large scale lexico-semantic knowledge bases, as WordNet, has become a usual practice for most Natural Language Processing. The diffussion and success of WordNet have determined the emergence of several projects that aim either to build wordnets for languages other than English[1] (Hamp and Feldweg, 1997; Artale, Magnini, and Strapparava, 1997) or to develop multilingual wordnets. The most important project in this line was EuroWordNet (EWN) (Vossen, 1998), a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet (Fellbaum, 1998).

Now MEANING (Rigau et al., 2002) is the cutting edge project in this line. MEANING[2] has designed the Multilingual Central Repository (MCR) to act as a multilingual interface for integrating and distributing all the semantic knowledge acquired in the project. The MCR follows the model proposed by the EuroWordNet project.

Building appropriate resources for broad–coverage semantic processing is a hard and expensive task, involving large research groups during long periods of developement. For example, dozens of person-years are be-

* This research has been partially funded by the Spanish Research Department (TIC2000-0335-C03-02, TIC2000-1735-C02-02), by the European Comission (IST-2001-34460), and by the Catalan Research Department (CIRIT 1999SGR-150).

[1] see those wordnets currently under developement at http://www.globalwordnet.org/

[2] http://www.lsi.upc.es/ nlp/meaning/meaning.html

ing invested world–wide into the development of wordnets for various languages. The outcomes of these projects are, usually, large and complex semantic structures, not compatible with resources developed in previous projects and efforts. This fact has severely hampered Human Language Technology (HLT) development.

MEANING plans to integrate into the MCR several large-scale resources developed in previous projects and efforts. Initially, most of the knowledge acquired in MEANING will be derived from WN1.6 (selectional preferences automatically acquired from SemCor and BNC). The Italian WordNet and the MultiWordNet Domains are aligned to WN1.6 (Bentivogli, Pianta, and Girardi, 2002; Magnini and Cavaglià, 2000), but the Spanish, Catalan and Basque wordnets are aligned to WN1.5 (Atserias et al., 1997; Benítez et al., 1998). Further, the EuroWordNet Base Concepts where selected from WN1.5, and they serve to hock the EuroWordNet Top Ontology via the Base Concepts.

To solve this version gap and in order to minimize side effects with respect other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association, MEANING plans to provide a generic, powerful and robust mapping tool and a new set of improved mappings.

That is, for MEANING it is fundamental to achieve a high performance and accurate tool to maintain compatibility between wordnets of different languages and versions, past and new. Nevertheless, automatic ontology mapping methods are difficult to evaluate. Hand checking of a small –statistically significant– sample of the performed connections, provides a quantitative idea of the accuracy of the technique, but does not allow to draw qualitative conclusions.

This paper presents an in depth study of the robustness and accurateness of the relaxation labelling algorithm for mapping already existing wordnets. The (Daudé, Padró, and Rigau, 2001) relaxation labelling based technique is used to map WN1.5 onto itself, which enables not only quantitative evaluation, but also the qualitative study of error cases. This study enables us to detect some anomaly cases which are probably inconsistences in the taxonomy. Examples and a typology for those cases is presented.

2 *Method Description*

Relaxation labelling (RL) is a generic name for a family of iterative algorithms which perform function optimization, based on local information, but with global effects. See (Torras, 1989) for a summary, or (Padró, 1998; Atserias, Padró, and Rigau, 2001) for previous applications to NLP tasks. One of its most remarkable features is that the focus problem is modelled in terms of compatibility/incompatibility constraints (which may be hand-written, statistical, machine-learned, ...) between variable–label pairs.

RL uses constraints to increase or decrease the weight for a variable label. In our case, constraints increase the weights for the connections between a source synset and a target synset. Increasing the weight for a connection implies decreasing the weights for all the other possible connections for the same source synset. To increase the weight for a connection, constraints take into account already connected nodes that have the same relationships in both taxonomies.

The problem is modelled with a variable for each node in the source taxonomy, which has as possible labels all candidate connections for that node (see Figure 1). Used constraints rely on checking the existence of a connected ancestor/descendant for both ends of a candidate connection. Complexity of constraints varies on the allowed distance from the candidate connection and in the simultaneously checked conditions. The RL algorithm will select the label assignment for all variables (i.e. the connection for each node) which better satisfies all constraints. More details on the algorithm and constraints can be found in (Daudé, Padró, and Rigau, 2000; Daudé, Padró, and Rigau, 2001).

Figure 1 shows an example of possible connections between two taxonomies. For source node S_1, connection C_4 will have its weight increased due to C_5, C_6 and C_1, while connections C_2 and C_3 will have their weights decreased. Eventually, label C_4 will be assigned to variable S_1.

3 *Validation via* automapping

In order to evaluate the performance of the algorithm, we mapped the nominal part of WN1.5 onto itself. The nominal WN1.5 is almost a tree –few nodes have more than one hyperonym– and consists of 60,557 nodes, 11 of which are roots, and 47,110 (77.79%) leafs.

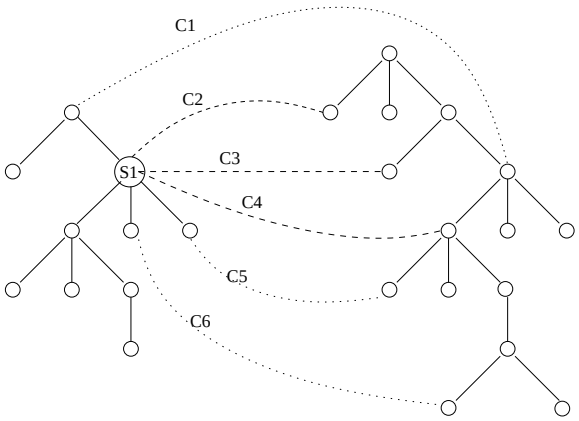

Figure 1: Example of candidate connections

The candidate conections for a source node are obtained retrieving all synsets in the target taxonomy for all words contained in the source synset. Since the target taxonomy contains a copy of the source synset, all synsets have at least one candidate connection. In WN1.5, 37,204 synsets are *single-link*, i.e. they have only one candidate connection. They don't need to be disambiguated, but are helpful to solve ambiguity for other nodes connected with them. The remaining 23,353 synsets (38.56%) are *multiple-link*, i.e. have more than one candidate connection. Each multiple-link sysnset has between 2 and 66 candidates, with an average of 4.26.

Using the algorithm with the same taxonomy as source and target not only is useful to evaluate its correctness and efficiency, but also to tune some of the used constraints, and to detect existing gaps and incorporate new constraints to cover them.

In this paper we analyze the behaviour of the algorithm on an incremental basis, starting with the simplest constraint configuration, and progressively extending the used model to enhance its performance.

3.1 Immediate connection (II) constraints

The simplest constraint set checks for the existence of a connection between *immediate* (II) hypernyms or hyponyms at both ends of the candidate connection, such as (C_4, C_1) in Figure 1.

Table 1 presents the results obtained using II constraints. Precision and recall are given over single and multiple link synsets. Recall is computed as the percentage of source nodes that keep the correct connection among their proposed targets. Precision is computed as the number of proposed targets that are correct connections.

Over trivial single-link synsets, the performace is obviously perfect. Over the multiple-link subset, some correct links are discarded by the algorithm, yielding a recall below 100%. There are only ten error cases – grouped in four clusters– which can be found in Figure 2, where the arrows show the correspondence betweeen synsets selected by the algorithm.

In each cluster, the error in one of the synsets causes the error in the others. For instance, case A in Figure 2 is more detailed in Figure 3, where we can observe that the target synset `00145061` is only reinforced by constraint C1, while target `08150656` receives support from constraints C2 and C3, causing it to be wrongly selected.

		II	IIB
	#NODES	PREC.-REC.	PREC.-REC.
single-link	37,204	100%-100%	100%-100%
multiple-link	23,353	93.80%-99.96%	93.86%-100%
Total	60,557	97.51%-99.98%	97.54%-100%

Table 1: Precision-recall results obtained using II and IIB constraint sets

II constraints provide support for a link from the existence of either a linked hyperonym or hyponym, but not from the simultaneous existence of them both. IIB constraints extend the II set with an extra support for those links with a *simultaneously* linked hyperonym and hyponym. This is precisely the case in the above mentioned errors, since for instance in case A, both hyperonym and hyponym for the source `00145061` are linked with the respective hyperonym and hyponym for target `00145061`, while the hypernym for source `00145061` is not linked with the hyperonym for the other candidate target `08150656`.

The use of IIB constraint will provide additional evidence in favour of the correct link, that should overwhelm the evidence provided by two hyponym constraints supporting the wrong candidate. As can be seen in Table 1, the use of these constraints produces a recall of 100% and an increment in precision, solving all wrong links presented in Figure 2.

This confirms the need for B constraints to help the disambiguation in cases such as those presented in the example. Note that this is a general statement, valid for any hierarchy, since only class/subclass relationships

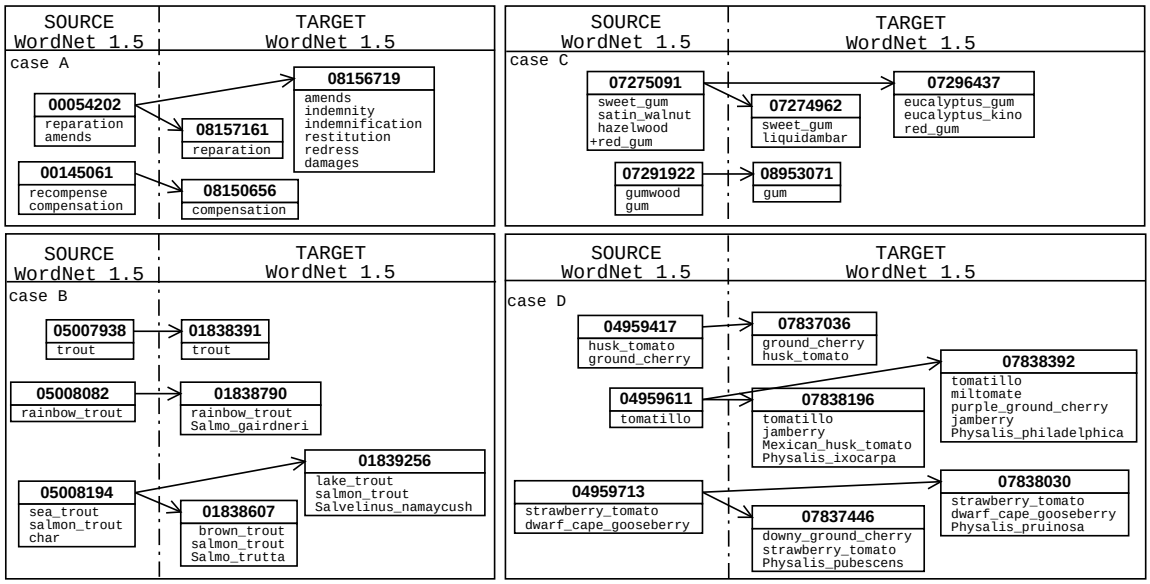

Figure 2: All wrong links selected by II constraint.

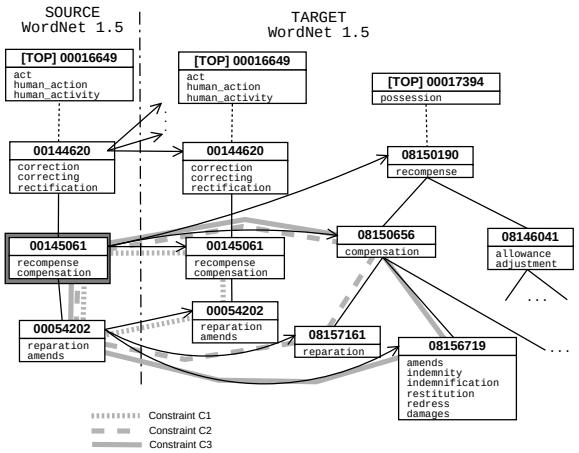

Figure 3: Details of wrong link in Fig. 2, case A.

are being used.

3.2 Using extra hyponym information

Although we have a 100% recall, precision is not perfect yet. This is due to remaining ambiguity in some nodes. Figure 4 presents an example of such a node (00026244) that occurs either with II or IIB constraints. Details on the involved relationships are also depicted: We can observe that source 00026059 is correctly linked since its hyponyms (00029218 and 00171746) provide the necessary evidence. Contrarily, source 00026244 is not disambiguated because both candidates have the same supporting evidence: constraint C1 for one candidate and C2 for the other.

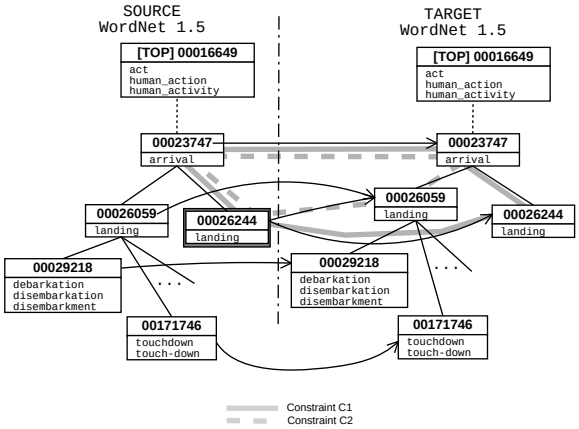

Figure 4: Relationship structure for ambiguous node example

This cases could be solved if knowledge about the number of daughters of each node was taken into account. We tested the following two ways of using this information (see Table 2 for results):

1. ZD constraint (Zero Daugthers): A simple boolean check consisting of a constraint that reinforces a connection between two leaf nodes (i.e. when both have zero daugthers).
2. ED constraint (Equal Daugthers): A generalization of the previous, consisting of a reinforcement of a connection between nodes with equal number of daughters.

	#NODES	IIB+ZD PREC.-REC.	IIB+ED PREC.-REC.
single-link	37,204	100%-100%	100%-100%
multiple-link	23,353	94.90%-100%	94.93%-100%
Total	60,557	97.97%-100%	97.98%-100%

Table 2: Precision-recall results when using constraints on the number of daughters.

When using constraints IIB+ZD, 1,136 nodes remain ambiguous, all but three of which are leaf nodes. One of these three is synset 02323757, presented in Figure 5.

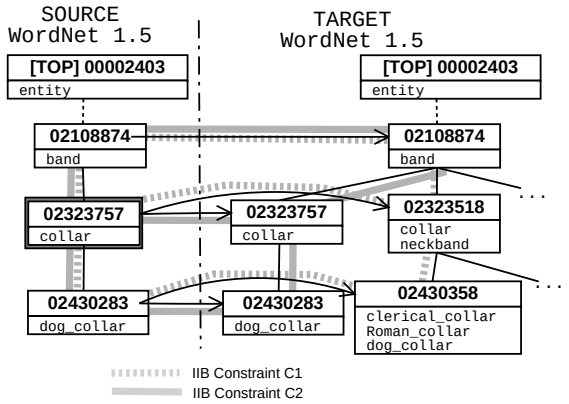

Figure 5: Example of non-leaf ambiguous node.

It can be observed that the ambiguity between targets 02323757 and 02323518 is caused by IIB constraints C1 and C2 in Figure 5, and since 02323757 is not a leaf, ZD constraint does not apply. If constraint ED is used instead, the ambiguity is correctly solved, since the synset for dog_collar is correctly linked, causing its hyperonym to be also correcly disambiguated.

When using IIB+ED constraints, the amount of remaining ambiguous nodes is 1,129, all of them leafs. Leaf nodes are the weakest point of the algorithm, since they have no descendants to provide information. Thus, when a node has as candidate targets two leaf sibling synsets, disambiguation is not possible using only hyper/hyponymy relationships. Example of such cases are the three leaf nodes in Figure 6, which keep as possible all their target synsets, since there is no difference between them that may help to prefer one of the targets.

3.3 Using other relationships

Although the main structure of WordNet relies in the taxonomical hyper/hyponymy relationships, it contains many other relationships. The nominal part includes also antonymy, meronym, holonymy and attribute. The former three are *noun-to-noun*, i.e. internal to the nominal part, and the latter relates *noun-to-adj*.

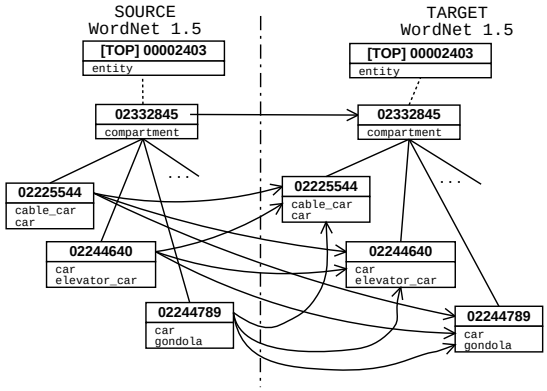

Figure 6: Example of ambiguity in leaf nodes

Since each ambiguous synset has different meronyms, using an II constraint on this relationship enables the algorithm to solve those ambiguity cases. Results when using all *noun-to-noun* relationships (plus ED constraints) are presented in the *Structural* column in Table 3.

With this model, there are 765 nodes that still remain ambiguous, since they do not have any other relationship we can use to provide extra information to help the disambiguation process. Thus, the use of non-structural information (i.e. not related to node relationships but to node similarity measures) will be necessary. Some of those cases appear in Figure 7.

	#NODES	Structural PREC.-REC.	Structural+WG PREC.-REC.
single-link	37,204	100%-100%	100%-100%
multiple-link	23,353	96.54%-100%	99.991%-100%
Total	60,557	98.64%-100%	99.997%-100%

Table 3: Precision-recall results obtained with each constraint model

3.4 Using non-structural information

To disambiguate cases in which a decision is not possible using only relationship-based constraints, we may extend our model with non-structural information which supports the connection between similar nodes. This obviously requires a way of computing node similarity that does not depend on the relationships among them. In the case of WN we

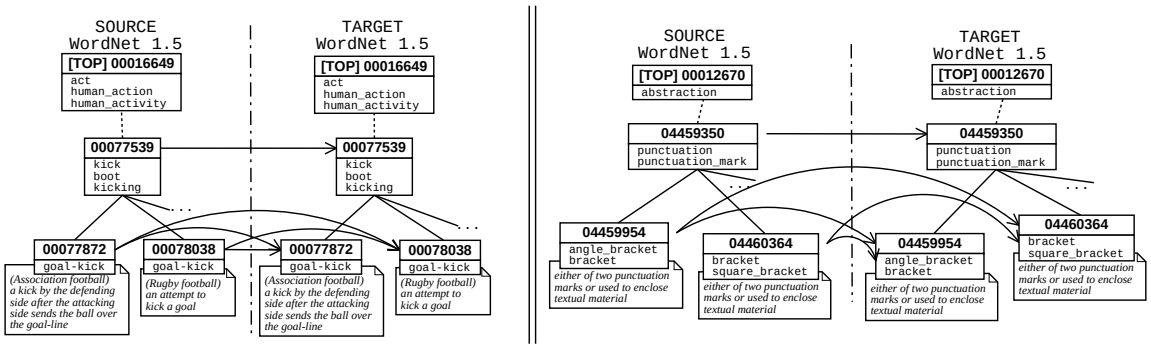

Figure 7: Example of nodes than can not be disambiguated with only relationship structure information

may use information internal to the node:

1. W constraint (coincident Words). The larger the number of coincidences in the words of two synsets, the more similar they are considered.
2. G constraint (coincident Gloss). The larger the number of coincidences in the words of both synsets glosses, the more similar they are considered. Non-content words (articles, prepositions, etc.) are excluded.

Using W constraint (word coincidence count) correctly disambiguates the example presented on the left of Figure 7. Similarly, the G constraint (gloss coincidence count) correctly disambiguates the right hand side example. Thus, to disambiguate as many cases as possible, we will use both constraints, though since many WN1.5 synsets do not have a gloss, the coverage of the G constraint will be low.

Rightmost column in Table 3 shows the results obtained with all structural and non-structural constraints. There are only two remaining ambiguous synsets, one of which is presented as sample in Figure 8. It can be seen that there is not enough information in the taxonomy (even for humans) to disambiguate those cases, nevertheless, one may wonder if they are actually different senses or merely an error in the taxonomy.

Thus, our validation method via the mapping of a hierachy onto itself turns out to be also useful to detect possibly duplicated concepts –or at least, anomalous cases– in the semantic network.

4 Analysys of detected anomalies

Depending on the constraints used, the amount of unresolved nodes varies. As said

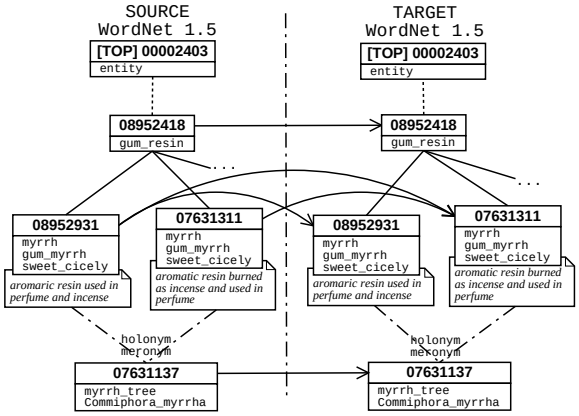

Figure 8: Example of node that can not be disambiguated with all the used constraints.

above, using all Structural+WG constraints only two cases which have identical structure, synset words and gloss words remain unresolved. The similarity criterion can be tuned by using a different set of constraints, for instance, if W constraint is not used, unresolved cases are those for which the only difference is some word in the synset.

Hand analysys of such cases in WN1.5, WN1.6 and WN1.7.1 for different constraint combinations discovers the anomaly patterns listed below, though it is difficult to assess which should be the appropriate correction without knowledge of the reasons that caused their inclusion in WN:

- Undistinguishable synsets, probably duplicates. This is the case of *[myrrh,gum_myrrh,sweet_cicely]* above, or *[Plantae,kingdom_Plantae,plant_kingdom]* presented in Figure 9a.
- Disinguishable synsets that should probably be joined in one. This happens in the case of the pairs (*[tolu]*–

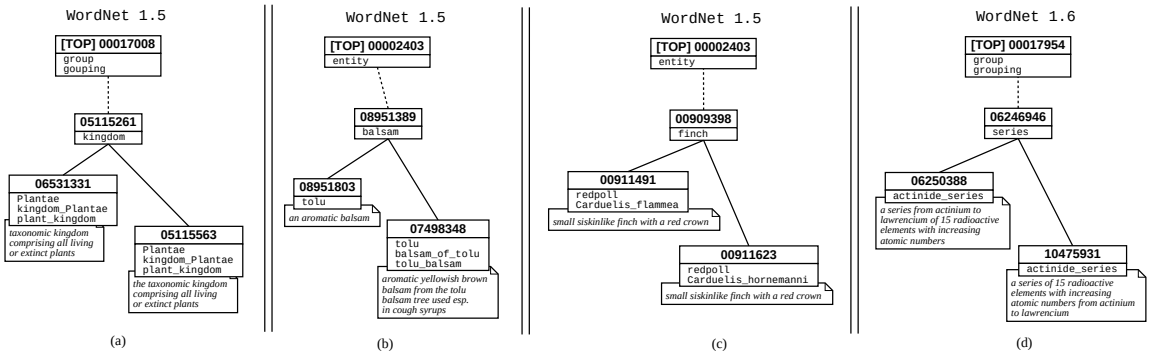

Figure 9: Examples of detected anomaly synsets.

[tolu,balsam_of_tolu,tolu_balsam]) (see Figure 9b), or (*[myrrh,gum_myrrh]*–*[myrrh,gum_myrrh,sweet_cicely]*) in WN1.6 and 1.7.1 (see below).

- Distinguishable synsets (by differences in word list) that should probably be restructured. Most involve different subkinds of a plant or animal, or more generally, specializations of the same concept that are daughters of a too general concept. For instance, synsets for *[redpoll,Carduelis_flammea]* and *[redpoll,Carduelis_hornemanni]* in Figure 9c are children of *[finch]*, without any intermediate *[redpoll]* concept. The same occurs with *[angle_bracket,bracket]* and *[square_bracket,bracket]* in Figure 7, being both under *[punctuation,punctuation_mark]*, while probably an intermediate *[bracket]* concept would be necessary.

Regarding the evolution of those cases through increasing WN versions, we find that most of them are maintained. Nevertheless, changes exists, and may be classified as:

- Undistinguishable synsets that are slightly distinguished in a newer version. This is the case of *[myrrh,gum_myrrh,sweet_cicely]*, which is undistingushable in WN1.5, while in later versions only one of both synsets retains the *sweet_cicely* variant
- Single synsets that are duplicated in newer versions. This is the case of *[actinide_series]* (Figure 9d) which is a single synset in WN1.5 and appears duplicated in 1.6 and 1.7.1 versions.
- Synsets not included in older versions that appear duplicated in newer ones, as for instance *gutta-percha_tree* which is not in WN1.5, but duplicated in WN1.6 and WN1.7.1.

5 *Conclusions and future work*

We have validated a WN mapping technique based on relaxation labelling through a detailed analysis of the results of mapping several wordnet versions onto themself. Using this approach, we have quantitative and qualitatively evaluated the relaxation labelling technique and we justified the different types of constraints and the different kinds of knowledge used. However, the main conclusion of this work is that the proposed method is robust enough even to detect in several wordnet versions, unclear synset distinctions, duplicates and possibly errors and inconsistences. We expect to detect a larger amount of inconsistencies in other part-of-speech categories.

Based on the study presented on this paper, we will design a new set of constraints based on the number of direct hyponyms. Using these new constraints, we expect to improve the current WN1.5 to WN1.6 mapping accuracy (precision-recall of 98.8%-98.9% for the noun hierarchy).

We also plan to evaluate the algorithm robustness when mapping different hierarquies. We will start mapping two identical taxonomies, progressively introducing differences between them. The introduced differences will consist of random node deletions from one of the taxonomies (either target or source). Note that in this case, the deletion of one node in one taxonomy will be seen as an insertion of its corresponding synset in the other.

References

Artale, A., B. Magnini, and C. Strapparava. 1997. Lexical Discrimination with the Italian Version of WordNet. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources*, Madrid. Spain.

Atserias, J., S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria.

Atserias, J., L. Padró, and G. Rigau. 2001. Integrating Multiple Knowledge Sources for Robust Semantic Parsing. In *proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'01)*, Tzigov Chark, Bulgaria.

Benítez, L., S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and Tools for Building the Catalan WordNet. In *Proceedings of ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Bentivogli, L., E. Pianta, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

Daudé, J., L. Padró, and G. Rigau. 2000. Mapping WordNets Using Structural Information. In *38th Annual Meeting of the Association for Computational Linguistics (ACL'2000).*, Hong Kong.

Daudé, J., L. Padró, and G. Rigau. 2001. A Complete WN1.5 to WN1.6 Mapping. In *NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations" (NAACL'2001).*, Pittsburg, PA, USA.

Fellbaum, C., editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Hamp, B. and H. Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*, Madrid. Spain.

Magnini, B. and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece.

Padró, L. 1998. *A Hybrid Environment for Syntax–Semantic Tagging*. Phd. Thesis, Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, February. http://www.lsi.upc.es/~padro.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*, Taipei, Taiwan.

Torras, C. 1989. Relaxation and Neural Learning: Points of Convergence and Divergence. *Journal of Parallel and Distributed Computing*, 6:217–244.

Vossen, P., editor. 1998. *Euro WordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .