# Incremental Knowledge Acquisition from WordNet and EuroWordNet

**Wim Peters**
University of Sheffield
United Kingdom
w.peters@dcs.shef.ac.uk

**Abstract**

This paper describes the process of the creation and extraction of implicit knowledge from WordNet (Fellbaum, 1998) and EuroWordNet (Vossen, 1998). This knowledge is an extension of the explicit knowledge structures already provided by the wordnets in the form of synsets and semantic relations, and is contained both within (Euro)WordNet's hierarchical structure and the glosses that are associated with each WordNet synset.

## 1.Introduction

WordNet (Fellbaum, 1998) and EuroWordNet (Vossen, 1998). Have been and are being used as resources for a number of NLP tasks such as semantic tagging (Volk et al, 2002) and information retrieval (Gonzalo et al., 1998). Although the thesauri contain a wealth of information, only part of their lexical knowledge is available in explicit form, organised in synsets and a number of semantic relations between these synsets such as synonymy, hypernymy and thematic relations. Much of the information is hidden away in both (Euro)WordNet's structure, and the glosses that are associated with the synsets. This paper describes an attempt to discover and extract at least part of that implicitly available lexical knowledge. The point of departure is the detection of patterns of figurative language use, more particularly metonymic cases of regular polysemy (Apresjan, 1973). This type of figurative language use exploits semantic regularities in language that, when captured, offer valuable additional semantic information for the concepts involved. Examples of regular polysemy have until now been the product of linguistic introspection and manual lookup in dictionaries and texts. The availability of electronic semantic resources such as WordNet makes it possible to extract and investigate regularities between sense distinctions in a data-driven way. These regularities form the core data set for the derivation of extended knowledge fragments that complement (Euro)WordNet.

## 2 Automatic Selection of Regular Polysemy

The work consisted of three phases. First, an automatic selection process identifies candidates for instantiations of regular polysemy (For a detailed description see Peters and Peters, 2001; Peters and Wilks, 2002) in WordNet on the basis of systematic sense distributions of nouns. These systematic distributions can be characterized by a pair of hypernyms taken from the WordNet hierarchies that subsume the sense combinations of the words involved. For instance, in two of its senses 'law' falls under the pattern *profession* (an occupation requiring special education) and *discipline* (a branch of knowledge). This pattern is also displayed by four other words in WordNet, namely 'architecture', 'literature', 'politics' and 'theology'. Figure 1 below illustrates this case.

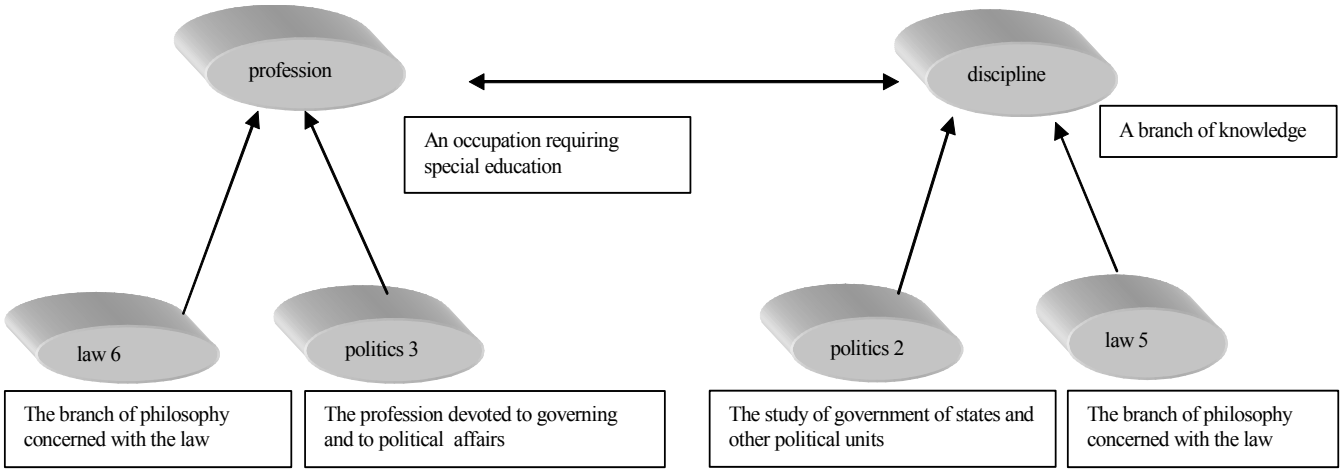


Figure 1: The Regular Polysemic Pattern *Profession - Discipline*

## 3 Extraction of Semantic Relations

In the second stage, the relations that exist between the word senses that participate in patterns are acquired in an automatic fashion. This additional information is obtained by analyzing the glosses that are associated with the synsets of the word senses involved and their hypernyms. After part of speech tagging and lemmatization all synonyms and hypernym synset members of the participating words are grouped together into two bags of words. These are then mapped onto the glosses that bare associated with all synsets in the hypernymic paths. If a verb occurs between pairs of words from each bag this is taken as the semantic relation that holds between the word senses. Figure 2 below illustrates the process.

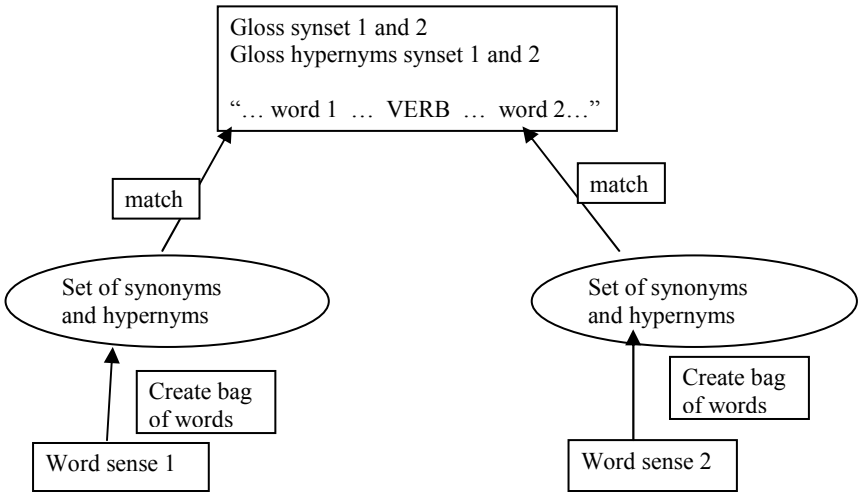


Figure 2: Extracting relations between the word senses

This method extracts, for instance, the relation 'master' between discipline and profession. By adding thematic roles to the concepts involved one can state that discipline is either the subject or instrument associated with 'master' and profession is the object. The explicit relations between word senses that can be gleaned from information implicit in glosses enriches the existing knowledge structures of WordNet, thereby expanding its coverage as a knowledge base. Also, it forms the start of the explicit encoding of metonymic potential of words that do not yet participate in the patterns. Overall, relations have been extracted for around 5000 candidate regular polysemic patterns.

## 4. Extended Knowledge Fragments

In the third phase, increasingly larger knowledge structures are built up on the basis of these sense pairs. The frame model (Minsky, 1975; Fillmore, 1977), elaborated in psychology and linguistics in the last two decades, provides us with a well established formalism for the illustration and representation of this systematic knowledge. Frames are conceptual wholes that can be found under various denominations in the literature, such as scenes, scenarios, domains, and Idealized Conceptual Models (Lakoff, 1987). The frame structure is defined as the relation that exists between elements of a frame or between the frame as a whole and its elements. The relation triples extracted in the second stage (e.g. **person-speak-language**) form the basic building blocks of the frames. Extension of these rudimentary frames takes place in two ways. First, the concept with which hypernyms from the regular polysemy patterns co-occur can be regarded as additional slots in a topical frame that characterizes a hypernym. For instance, the pattern 'music-dance' covers words such as *tango* and *bolero*. Music in its turn co-occurs with a number of other concepts within the hypernym pairs that characterize the regular polysemic patterns. These concepts and the relations that have been extracted between these hypernyms form a further extension of the MUSIC frame. For music, the following relations with other hypernyms have been extracted: **person-make/accomplish-music** and **music-accompany-activity**. Figure 3 below illustrates this. A further extension takes the semantic context of EuroWordNet into account. From the superset of all concepts and relations that are linked to MUSIC in all eight language specific wordnets that are contained within EuroWordNet the MUSIC frame is extended with this knowledge. The resulting structure is illustrated by figure 4 below.

## 5. Conclusion

In conclusion, the followed method demonstrates that it is possible to enrich the explicit structure of (Euro)WordNet with information that is implicitly available in the resource itself. These frame-like extensions can be incrementally extended by means other new and existing techniques, and form an increasingly rich basis for knowledge based NLP applications. For

instance, this knowledge base should contain enough information to allow the application of inferencing methods for the processing of coreference, anaphora and bridging expressions. For a text segment such as of linking relations between the nouns (*composer*, *sonata*, *music*). This is achieved by means of the hypernymic relation between *sonata* and *music*,

"The composer finished his sonata. Music had always been his first love.", the extended knowledge fragment for *music*, illustrated in figure 3, enables the detection

another hypernymic relation between *composer* and *person*, amd the 'make' relation between *person* and *music*.

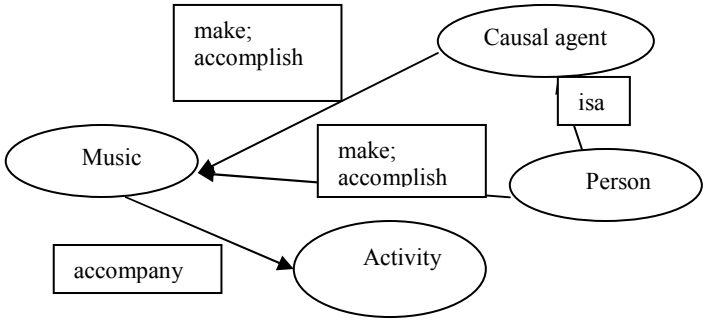


Figure 3: Extension of MUSIC Frame with Hypernymic Collocates

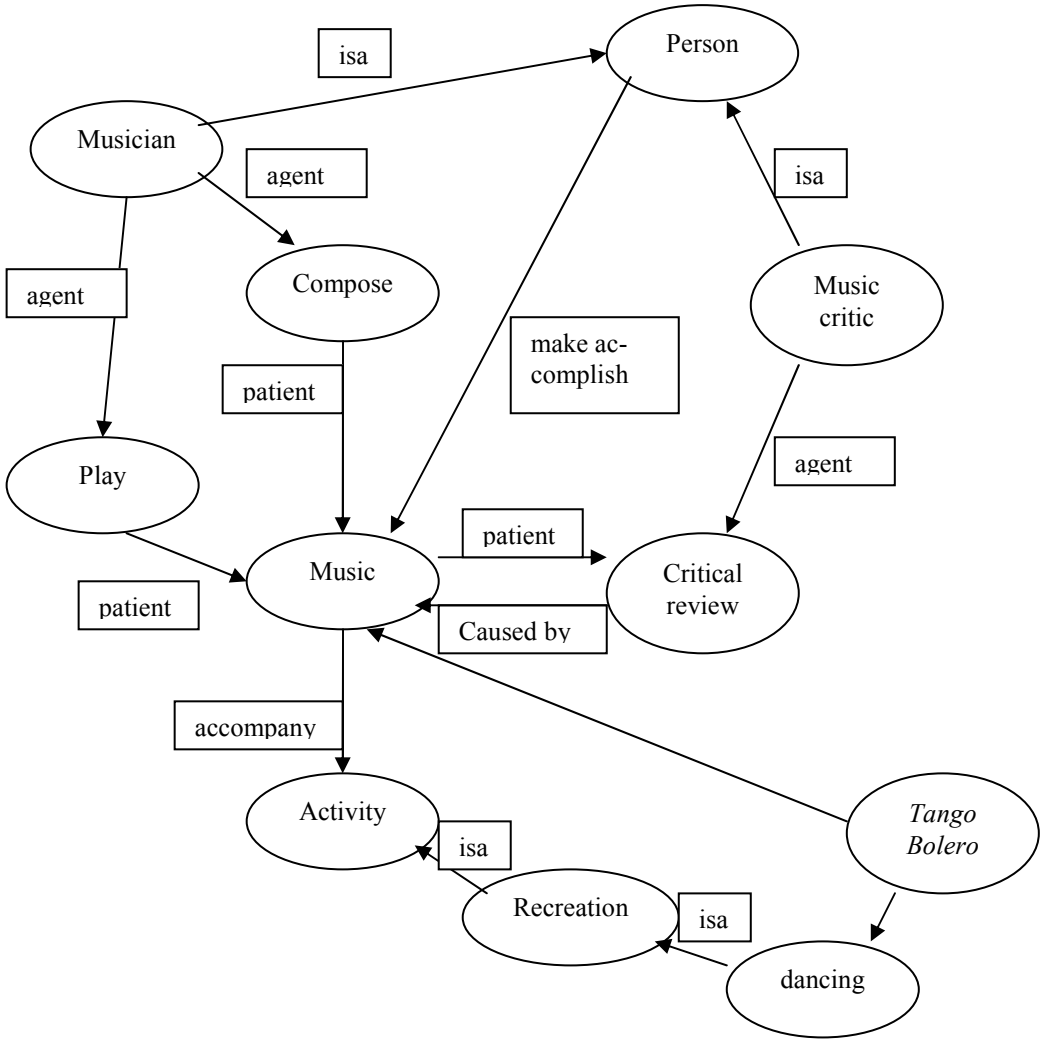


Figure 4: Extended Frame for 'music'

## References


Apresjan, J. (1973), *Regular Polysemy*
In: Linguistics 142, pp. 5-32

Fellbaum, Christiane (ed.) (1998), *WordNet: An Electronic Lexical Databas*e.
Cambridge, Mass.: MIT Press.

Fillmore, C (1977), *Scenes and frames semantics.*
In: Zampolli, A (ed.) Linguistic structures processing. Benjamins, Amsterdam, The Netherlands, pp. 55-81.
Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Gonzalo, J., Verdejo, F, Chugur, I. and Cigarrán, J. (1998), *Indexing with WordNet synsets can improve text retrieval*
In *ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*.

Minsky, M. (1975), *A Framework for Representing Knowledge*.
In: Winston, P.H. (Ed.), The Psychology pof Computer Vision, New York: McGraw-Hill, pp. 211-277.

Peters, W. and Peters, I. (2000), *Lexicalised Systematic Polysemy in WordNet*
In *Proc. Secondt Intnl Conf on Language Resources and Evaluatio*n
Athens, Greece

Peters, W. and Wilks, Y. (2001), *Distribution-oriented Extension of WordNet's Ontological Framework,*
Proceedings RANLP2001, Tzigov Chark, Bulgaria

Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., Raileanu, D. and Sacaleanu, B. (2002), *Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval*.
In: International Journal of Medical Informatics, Volume 67:1-3

Vossen, P.(1998), Introduction to EuroWordNet.
In: Nancy Ide, N., Greenstein, D. and Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.