Metonymy as a Cross-lingual Phenomenon

Wim Peters
NLP Group
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield S1 4DP
U.K.
w.peters@dcs.shef.ac.uk

Abstract

The question of whether metonymy carries across languages has always been interesting for language representation and processing. Until now attempts to answer this question have always been based on small-scale analyses. With the advent of EuroWordNet (Vossen 1998), a multilingual thesaurus covering eight languages and organized along the same lines as WordNet (http://www.cogsci.princeton.edu/~wn/) we have a unique opportunity to research this question on a large scale. In this paper we systematically explore sets of concepts comprising possible metonymic relations that have been identified in WordNet. The sets of concepts are evaluated, and a contrastive analysis of their lexicalization patterns in English, Dutch and Spanish is performed. Our investigation gives insight into the cross-linguistic nature of metonymic polysemy and defines a methodology for dynamic extensions of semantic resources.

1. Introduction

Viewed traditionally, metonymy is a non-literal figure of speech in which the name of one thing is substituted for that of another related to it. It has been described as a cognitive process in which one conceptual entity, the vehicle, provides mental access to another conceptual entity (Radden 1999). In its basic form, it establishes a semantic relation between two concepts that are associated with word forms. The semantic shift expressed by the relation may or may not be accompanied by a shift in form. The semantic relation that is captured by metonymy is one of semantic contiguity, in the sense that in many cases there are systematic relations between metonymically related concepts that can be regarded as slots in conceptual frames (cf. Fillmore 1977).

For example, in the sentence 'The colonies revolted against the crown.' *crown* is used as a symbol for the monarchy as well as denoting the traditional head ornament worn by the monarch. As the example above shows, polysemy is a common way in which metonymically related concepts manifest themselves in language.

It is to be expected that any systematic semantic relations between concepts expressed by these sense distinctions are lexicalized, i.e. they are explicitly listed in dictionaries and independent of a pragmatic situation. For example, *university* is on the one hand an institution and on the other a building. The semantic relation between the two senses is 'is housed in'.

Regular polysemy is a subset of metonymy that covers the systematicity of the semantic relations involved. It can be defined as a subset of metonymically related senses of the same word displaying a conventional as opposed to novel type of semantic contiguity relation. This relation holds for related senses of two or more words (Apresjan, 1973), i.e. is a lexicalized pattern, not a nonce formation (a pragmatically defined novel metonymy), and can therefore be

called regular. It is this subtype of metonymy that we concentrate on in this paper.

2. Regular polysemy across languages

The question whether regular polysemy is a cross-linguistic phenomenon has until now only been approached by small scale analyses.

For instance, Kamei and Wakao (Kamei, 1992) approached the question from the perspective of machine translation and conducted a comparative survey of the acceptability of metonymic expressions in English, Chinese and Japanese consisting of 25 test sentences. The results they report show that in some cases English and Japanese share metonymic patters to the exclusion of Chinese, but that in others English and Chinese team up.
(Seto1996) performed a study into the lexicalization of the container-content schema in various languages (Japanese, Korean, Mongolian, Javanese, Turkish, Italian, Germanic and English). This pattern is lexicalized in English by 'kettle':
1. A metal pot for stewing or boiling; usually with a lid
2. The quantity a kettle will hold
His observation was that the pattern is observable in all languages, and can be considered cross-linguistic. This small study seems to indicate that the regular polysemic pattern extends over language family boundaries to such an extent that it almost seems universal. This could suggest that the pattern is rooted in general human conceptualisation, and reflects an important non-arbitrary semantic relation between concepts or objects in the world. Indeed, if we describe the relation between *container* and *content* in terms of Aristotle's qualia structure (Pustejovsky 1995), we see that it is the function of a container to hold an object or substance (telic role) and that a container is normally brought into existence for this purpose.

More small-scale studies like the ones described above have been performed, mostly relying on introspection and small-scale dictionary analysis. A limited number of patterns that are valid in more than one language have been identified such as container/content and producer/product (Peters 2000). With the availability of WordNet and EuroWordNet it has become possible to investigate the cross-linguistic nature of metonymy on a large scale.

3. EuroWordNet

EuroWordNet (EWN) (Vossen 1997; Peters 1998) is a multilingual thesaurus incorporating wordnets from eight languages: English, Italian, Dutch, German, Spanish, French, Czech, Estonian. The wordnets have been built in various ways. Some of them have been created on the basis of language specific resources and matched onto the original Princeton WordNet (Fellbaum 1998) when the interlingual relations were created. They therefore reflect the language specific lexicalization patterns and semantic organization. Others have been built from the start on the basis of a match between WordNet and bilingual dictionaries. In this case the conceptual structure is less language specific but can be regarded as the conceptual overlap between the structure of the English WordNet and the ontological structure associated with that particular language.

EuroWordNet gives us for the first time the opportunity to examine the question of the language independence of regular polysemy in a more systematic and automatic way.

4. Methodology

The following methodology has been followed:

First, the hierarchy of WordNet1.6 was analysed in order to obtain English candidates for regular polysemic patterns (section 4.1). Then a process we call lexical triangulation was applied to these data within EuroWordNet (section 4.2). The results were then manually evaluated.

4.1 Automatic candidate selection

A technique was developed (Peters 2000) for identifying sense combinations in WordNet where the senses involved potentially display a regular polysemic relation, i.e. where the senses involved are candidates for systematic relatedness.
In order to obtain these candidate patterns WordNet (WN) has been automatically analysed by exploiting its hierarchical structure. Wherever there are two or more words with senses in one part of the hierarchy, which also have senses in another part of the hierarchy, then we have a candidate pattern of regular polysemy. The patterns are candidates because there seems to be an observed regularity for two or more words. This follows the definition of (Apresjan 1973) mentioned in the introduction.
An example can be found in Figure 1 below.

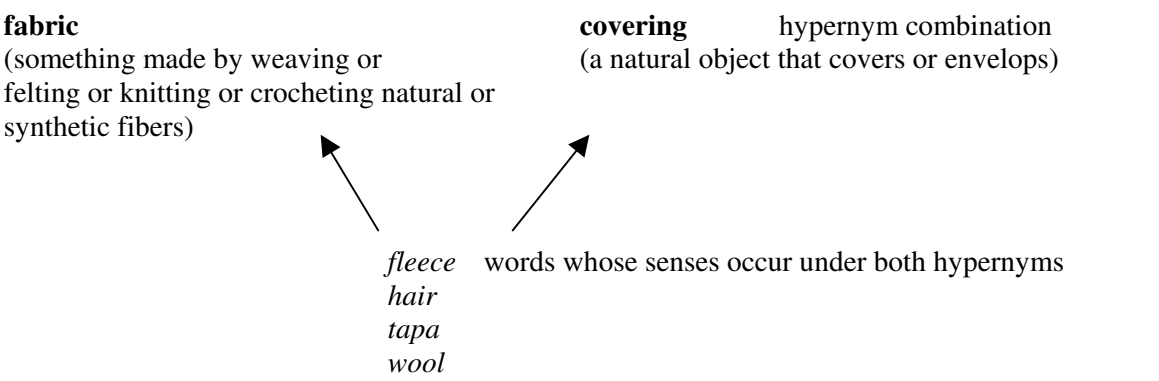

Figure 1: words in WordNet covered by the pattern fabric/covering

We have restricted our experiments to cases where the related meanings are of the same syntactic class (nouns). The procedure does not discover all regular polysemy relations, because the outcome is heavily dependent on the consistency of the encoding of these regularities in WordNet.

4.2 Lexical triangulation

In order to determine whether regular polysemy is indeed a cross-linguistic phenomenon, one needs to compare languages, preferably from different language families.

Data will depend heavily on vocabulary coverage in various languages, and until the advent of EuroWordNet no serious lexical data sets were available for analysis. The EuroWordNet database is the most comprehensive multilingual thesaurus to date. This resource not only provides us with an appropriate amount of lexical information in terms of vocabulary coverage, but also has the additional advantages that its taxonomic building blocks are identical for all languages involved and the language specific concepts are all linked to an interlingua which is based on the full set of the original Princeton WordNet (version 1.5), and is referred to as the interlingual index (ILI).

We started with a comparative analysis of Germanic and Romance languages. The main reason for this is that the size of the corresponding wordnets is large enough to yield significant results. For our analysis we used three languages: English, Dutch and Spanish, hence the term for this process: lexical triangulation.

Singling out areas where three language-specific lexicalization patterns converge

enabled us to identify metonymic patterns that supported the hypothesis that certain metonymic relationships have a higher degree of universality.

We extracted the sense combinations of Spanish and Dutch words that participate in any of the potential regular polysemic patterns from the initial large set described in section 4.1. In other words, we concentrate here on lexicalization patterns in three different languages: sense combinations that are lexicalized by one language-specific word in English, Spanish and Dutch.

The first step in this process was the reduction of the search space for regular polysemic patterns in EuroWordNet. First we determined the conceptual overlap for nouns between the English, Dutch and Spanish wordnets. Table 1 below shows the number of nouns in the three wordnets involved.

language	Number of noun synsets	Number of corresponding ILI concepts
English	66025	66025
Dutch	28352	26779
Spanish	24073	24087

Table 1: conceptual coverage of English, Dutch and Spanish wordnets

The conceptual overlap between these wordnets is computed simply by determining the intersection of ILI noun concepts covered by each of the wordnets. The total overlap is 17007 ILI concepts. There are 920 English polysemous nouns with two senses or more within synsets linked to this set of ILI concepts. Their senses have identical language specific lexicalizations in Spanish and Dutch. For example, the English word *church* has one sense that is a building and another that is an institution. The same sense distinctions apply to the Spanish *iglesia* and the Dutch *kerk*. The senses in the different wordnets are linked through the ILI concepts by means of equivalence synonymy or near-synonymy relations (Vossen 1997).

The second step was to map these noun senses onto the results from the wordnet analysis described in section 4.1, and then to evaluate the cross-linguistic validity of the regular
polysemic patterns that have been projected from the English monolingual wordnet onto the Dutch and Spanish wordnets.

5. Evaluation

The cross-linguistic filter yields a subset of the monolingual analysis data described in section 4.1. It covers 404 distinct English nouns out of a total of 8062 (5%).

This original filter considered nouns satisfying the criteria of Apresjan (cf. section 1), i.e. they are one of at least 2 words with sense distinctions that exhibit a particular relationship.

The percentage covered by the cross-linguistic data compared to the original analysis gradually varies from a 100% for the very small potential classes of regular polysemy (2-3 words) to 1-2% for middle sized (30-50 words) and large classes (100+ words).

In order to create a set for manual evaluation, the set of 404 English nouns was reduced by strengthening the Apresjan criterion and requiring that a word be considered only if it was one of at least a three word set illustrating the regular polysemy (RP). We will refer to this as a three-word RP class. The rationale behind this was that two word candidate RP classes introduce noise because of the increased

probability of a fortuitous coincidence of senses belonging to a set of just two words. This step reduced the number of participating words to 394. At this point, 177 words were randomly chosen from this set for manual evaluation. The evaluation consisted of examining the hypernym pairs that reflect a candidate regular polysemic relation.[1] The criteria used in this step are semantic homogeneity (the semantic relation that defines the candidate RP class should apply to the majority of the participating words) and specificity of the pattern (the lower the position of the hypernymic pair in the hierarchy, the more specific the semantic relation).

109 of these words displayed valid regular polysemic patterns (62%), 68 did not (38%). This means that by means of this automatic filtering method we have a 62% success rate for identifying valid regular polysemic patterns. Below are a few examples of cross-linguistic RP classes that have satisfied the criteria of the evaluation.

Hypernymic Pair: Control (the activity of managing or exerting control over something) – **Trait** (a distinguishing feature of one's personal nature)
English RP class (7 total): *abstinence, sobriety, inhibition, restraint, self-control, self-denial, self-discipline*
Dutch RP class (2 total): *zelfcontrole, onthouding*
Spanish RP class (3 total): *autodiscipline. abstinencia, abnegación, inhibición*
Coverage of the intersection between all three languages: 36% of set derived from WordNet

Hypernymic Pair: Fabric (something made by weaving or felting or knitting or crocheting natural or synthetic fibers) - **Covering** (a natural object that covers or envelops)
English Rp class (4 total): *wool, hair, fleece, tapa*
Dutch RP class (1 total): *wol*
Spanish RP class (1 total): *lana*
Coverage of the intersection between all three languages: 25% of set derived from WordNet

Hypernymic Pair: Plant (a living organism lacking the power of locomotion) - **Edible fruit** (edible reproductive body of a seed plant especially one having sweet flesh)
English RP class (159 total): *apple, boxberry, blackcurrant, banana, fig . . .*
Dutch RP class (9 total): *banaan, vijg, persimoen, meloen…*
Spanish RP class (20 total): *banana, plátano, melón, caqui, higo…*
Coverage of the intersection between all three languages: 2.5% of set derived from WordNet

Hypernymic Pair: Person (a human being) - **Quality** (an essential and distinguishing attribute of something or someone)
English RP class (11 total): *attraction, authority, beauty, . . .*
Dutch RP class (1 total): *schoonheid*
Spanish RP class (4 total): *belleza,atracción, autoridad, imagen*
Word intersection between all three languages: 9% of set derived from WordNet

Hypernymic Pair: Substance (that which has mass and occupies space) - **Drug** (something that is used as a medicine or narcotic)
English RP class (25 total): *alcohol, bromide, dragee, histamine, iodine, liquor…*
Dutch RP class (2 total): *broom, cocktail*
Spanish RP class (10 total): *bromuro, histamina, muscatel, yodo…*
Word intersection between all three languages: 4% of set derived from WordNet

Hypernymic Pair: Occupation (the principal activity in your life) – **Discipline** (a branch of knowledge)

[1] A complication arises because many combinations of hypernym pairs can be considered for the same set of words. (In fact the possibilities are the Cartesian product of the ancestors of each of the hypernyms in the pair). If all hypernymic combinations were taken into account this amounts to an average of 17 classes per word.

English RP class (6 total): *architecture, literature, politics, law, theology, interior design*
Dutch RP class (1 total): *architectuur*
Spanish RP class (2 total): *arquitectura, teología*

6. Universality of regular polysemy

It is possible to view these results as an indication of the cross-linguistic validity of the regular polysemic patterns and their level of universality relative to the language families represented by the wordnets. The hypothesis is that if a metonymic pattern occurs in several languages, there is stronger evidence for a higher level of universality of the regular polysemic pattern.

Of course there is interference with the coverage of the wordnets in EuroWordNet. Since the Dutch and Spanish wordnets are only half the size of the English wordnet only limited coverage can be expected. Still, the coverage seems to be consistently low in most cases, often not more than 2-5%. On the basis of wordnet size only one would expect a higher coverage.
There are other explanations for the lack of identical lexicalizations in other target language wordnets:

1. The metonymic pattern is language specific, and is not realised as a polysemous word in the target language. For example, the Dutch *kantoor* is synonymous to the English *office* in the sense 'where professional or clerical duties are performed', but its sense distinctions can not mirror the regular polysemic relation in English with 'a job in an organization or hierarchy'.

2. The pattern is unattested in the target language in terms of usage but forms a potential sense extension in that language. For instance, the Spanish *iglesia* and the Dutch *kerk* both mean 'building for worship' and 'a service conducted in a church'. The Spanish wordnet has an additional systematically related sense for iglesia ('institution to express belief in a divine power') that is not shared by its Dutch counterpart but is a valid new sense.

3. The missing sense can in fact only be lexicalized by another word or compound or derivation related to the word with the potentially missing sense. For example, the Dutch *vereniging* has the sense (an association of people with similar interests). The English equivalent is *club*, for which there is another sense in Wordnet (a building occupied by a club). This is not a felicitous sense extension for the Dutch *vereniging*, because the favoured lexicalization is the compound *verenigingshuis* (club house).

4. The metonymic pattern is in fact attested in the language, but one or more senses participating in the patterns has not yet been captured in the wordnet. One of the reasons could be the sense granularity of the resource on the basis of which the wordnet has been built. For example, *embassy* has one sense in WordNet (a building where ambassadors live or work). The Dutch translational equivalent *ambassade* has an additional sense denoting the people representing their country. This sense can be projected to the English WordNet as a regular polysemy pattern that is also valid in English. In fact, LDOCE (Procter,1978) only lists the sense which is missing in WordNet.

Word intersection between all three languages: 16% of set derived from WordNet

7. Coverage and extendibility

There are many RP classes whose English word members do not all have a Dutch or Spanish counterpart. We wanted to evaluate the universality of the regular polysemic relations by testing native speaker intuitions about these regular polysemic gaps. This

was done by projecting the senses of the participating English words in an RP class onto Dutch and Spanish, and to assess whether the missing senses were adequate additional senses in these two languages.

The experiment we conducted was very small. We intend to perform more experiments of this kind in the future. The pattern we examined is the hypernymic combination
occupation (the principal activity in your life) – **discipline** (a branch of knowledge). This RP class has five members. Two Dutch and two Spanish native speakers were asked to judge the felicitousness of the senses that are missing in the Dutch and Spanish wordnets. Below is a short discussion of each member.

interior design
1. the trade of planning the layout and furnishings of an architectural interior
2. the branch of architecture dealing with the selection and organization of furnishings for an architectural interior

The corresponding Dutch word *binnenhuisarchitectuur* has only one sense which is linked to both WordNet senses by means of a near-synonymy relation. This means that the Dutch wordnet is underspecified for the distinction of these metonymically related senses and can be extended with the specific sense distinctions (see explanation 4 above). This coincided with the verdict of the Dutch jury.
The Spanish WordNet has a separate translation for each sense: *interiorismo* (corresponding to *interior design* 1) and *deseño de interiores* (corresponding to *interior design* 2). The latter translational equivalent was considered to also have a possible trade reading.

law
1. the learned profession that is mastered by graduate study in a law school and that is responsible for the judicial system
4. the branch of philosophy concerned with the law

The Dutch '*rechtswetenschap*' has only one sense, which is linked to both WordNet senses by means of a near-synonymy relation. This again means that the Dutch wordnet is underspecified for the distinction of these metonymically related senses and can be extended with the specific sense distinctions (see explanation 4 above). This coincided with the verdict of the Dutch jury.
The Spanish equivalent of *law* 4 is *jurisprudencia*, whereas *law* 1 does not have a correspondence in the Spanish wordnet. The profession reading was not considered a felicitous additional sense for this word. Both subjects remarked that another word captures both meaning: *leyes*, which is not present in the Spanish wordnet.

literature:
1. the profession or art of a writer
2. the humanistic study of a body of literature

The Dutch *letterkunde* is only linked up to sense literature no. 2. Sense no. 1 was not considered to be a straightforward new sense for this word by the judges.

The Spanish *literatura* lacks a profession reader in the Spanish wordnet. This sense was considered as valid by one subject, but rejected by the other subject.

politics
1. the profession devoted to governing and to political affairs
2. the study of government of states and other political units

The Dutch word *politicologie* also has only one sense that is linked to both WordNet senses by means of a near-synonymy relation. This again means that the Dutch wordnet is underspecified for the distinction of these metonymically related senses and can be extended with the specific sense distinctions. The Dutch subjects, however, were not happy with the profession reading.

The Spanish *política* lacks a profession reading in the Spanish wordnet. The Spanish subjects considered this a valid sense for this word.

theology
1. the learned profession acquired by specialized courses in religion (usually taught at a college or seminary
2. the rational and systematic study of religion and its influences and of the nature of religious truth

The Dutch *theologie* has no profession reading. This reading was considered valid by the Dutch subjects.

The Spanish *teología* has both senses in the Spanish wordnet, and this coincides with the subjects' intuition.

The results are summarized in table 2 below. Overall, the projection of the word senses onto the Dutch wordnet yields a sense extension for one word out of a possible two. For the Spanish wordnet the same process creates valid new senses for two out of four words.

	Senses available in NL	**Sense possible in NL**	**Senses available in ES**	**Sense possible in ES**
interior design	lumped into one sense	**+**	**-**	**+**
law	lumped into one sense	**+**	**-**	**-**
literature	**-**	**-**	**-**	**+/-**
politics	lumped into one sense	**-**	**-**	**+**
theology	**-**	**+**	**+**	**+**

Table 2: experimental results summary

8. Conclusion

We have described a methodology for systematically and automatically investigating the question of whether or not certain kinds of metonymic relationships hold across languages. We used EuroWordNet as the test bed for this investigation, although the technique is not limited to EuroWordNet and will apply to any multilingual resource that captures hypernymic relationships and has some notion of correspondences among languages. Our manual evaluation shows regular polysemy patterns to be valid across the three languages examined, and indicates that regular polysemy has a certain level of universality. Also, the results of the experiment indicate a potential for enhancing the semantic compatibility and consistency of wordnets. Wordnets are automatically extendable on the basis of regular polysemic information available from other wordnets. In our small experiment 50% of the Dutch and Spanish words that do not display a WordNet-derived regular polysemic pattern were successfully semantically enriched with this pattern. Future work will involve further experimentation.

Acknowledgement
This work was supported by the UK Engineering and Physical Sciences Research Council under grant M73521 (MALT).

References

Apresjan, J. (1973), *Regular Polysemy*
In: Linguistics 142: 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.
Amsterdam: Benjamins, 55-81.

Kamei, S. and Wakao, T. (1992), *Metonymy: Reassessment, Survey of Acceptability, and its Treatment in a Machine Translation System*
In: Proceedings of ACL

Peters, W., Vossen, P., Diez-Orzas, P., Adriaens, G. (1998), *Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index,*
In: Ide, N., Greenstein, D.,Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 221-251.

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

Pustejovsky, J. (1995), *The Generative Lexicon*,
MIT Press, Cambridge MA, U.S.A.

Radden, G. and Kövecses (1999), *Towards a Theory of metonymy*
In: Panther, K.U. and Radden, G. (eds.) Metonymy in language and Thought.
John Benjamins, Amsterdam

Seto, Ken-ichi (1996), *On the Cognitive Triangle: the Relation of Metaphor, Metonymy and Synecdoche*
In: A. Burkhardt & N. Norrick, eds., Tropic Truth (working title). Berlin-NewYork: De Gruyter, 1996.

Vossen, P., 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.

Vossen, P., Peters, W. and Díez-Orzas, P. (1997), *The Multilingual design of the EuroWordNet Database*
In: Mahesh, K. (ed.) Ontologies and multilingual NLP, Proceedings of IJCAI-97 workshop,
Nagoya, Japan, August 23-29.