Proper Names in A Semantic Database

Rita Marinelli – Adriana Roventini

Istituto di Linguistica Computazionale del CNR Pisa Italy
e-mail: rita.marinelli@ilc.cnr.it – adriana.roventini@ilc.cnr.it

Abstract

Among the resources developed in SI-TAL (Integrated Systems for the Automatic Treatment of Language), ItalWordNet (IWN) was built as reference semantic database, enlarging the Italian WordNet developed in the framework of the European project EuroWordNet (EWN). The Italian lexical database was increased, by introducing and codifying, besides the new grammatical categories of the adjectives and adverbs, a subset of proper names. In the IWN context, the subset of proper names represents a quantitatively limited portion, about 3600 synsets, but it may become a qualitatively important extension. The ever growing amount of non-structured information, stored in natural language, requires the availability of computational instruments able to manage this kind of information where proper names show a remarkable incidence in any types of texts. The work here presented falls in this context, taking into account the proper names, and is focussed on: i) encoding in the IWN database; ii) more typical uses in either proper or metaphorical and metonymic ways such as textual corpora evidence; iii) possibility of a well reasoned and structured enlarging of this data on the basis of the recent experience carried out in IWN.

1. Building the set of proper names in the IWN database

IWN was first developed within the EWN[1] project (Vossen, 1999) and then extended in the framework of an Italian national project for automatic treatment of the language SI-TAL[2]. IWN (Roventini et al., 2002) is a lexical-semantic database containing semantic information for about 50,000 synsets of nouns, verbs, adjectives, adverbs, and a subset of proper names. The information is encoded in the form of lexical-semantic relations between pairs of synsets (synonym sets). The most important relations encoded, using machine-readable dictionaries as sources, are synonymy and hyponymy; however a rich linguistic model was designed containing many other lexical-semantic relations which are encoded for various subsets of Italian nouns, verbs and adjectives. All the synsets are also linked to WordNet 1.5, the Princeton Wordnet database (Miller et al., 1990).

In the framework of the SI-TAL project the lexical coverage of IWN has been extended by adding, besides two grammatical categories not encoded in EWN (i.e. adjectives and adverbs), a set of proper names which are taken into consideration in this paper. This decision was also due to the high degree of incidence of proper names observed in the corpus selected within SI-TAL for semantic annotation.

1.1 Coding proper names

In IWN proper names are connected to the class they belong to by means of the Belongs_ *to_ class* relation. This relation and its reverse *Has_instance are* only used to link instances with synsets. Indeed in the IWN database, unlike the well known Princeton semantic WordNet (Miller et al., 1990), hyp(er)onymy or 'is-a' relation was not used for this part of the lexicon. For proper names the "inherence propositions" between an individual and a class are allowed, not the "relation propositions" which are allowed only between classes (Blanché, 1968). What is denoted by the name belongs to a class, not to the name: using a name is not a matter of representing it as having certain properties but, as Russell (1919) said, "merely to indicate what we are speaking about…". Moreover, whereas Common Nouns may have some relation with the referent, so that they are almost all the same, the Proper Name is a target that does not depend on the context: "a Name refers to an individual. And once the meaning of the name has been established, a context cannot normally change very much of it." (Pamp, 1985).

However, in IWN there are other relations used to link proper names with common nouns and adjectives: *Derivation* and *Pertains_to*. *Derivation* is a morphological relation, which links the proper name with its derivatives and viceversa. As in EWN, it is used to encode derivation links when no other semantic relation is available. It connects variants belonging to different PoSs and applies both to the first and to the second order entities as shown in the examples below:

Grande (wide)	*Derivation*	grandemente (widely)
Marx	*Derivation*	marxismo (Marxism)
Romanità	*Derivation*	Roma (Rome)

[1] EWN was a project in the EC Language Engineering (LE4003) programme. Complete information on EWN can be found at its web site: http://www.hum/uva.nl/~ewn.

[2] The SI-TAL project : 'Integrated Systems for the Automatic Treatment of Language' was a National Project, coordinated by A. Zampolli, devoted to the creation of large linguistic resources and software tools for Italian written and spoken language processing. Besides IWN, the following were developed within the project: a treebank with a three level syntactic and semantic annotation, a system for integrating NL processors in applications for managing grammatical resources, a dialogue annotated corpus for applications of advanced vocal interfaces, software and tools for advanced vocal interfaces.

The P*ertains_to* relatio*n* and its reverse H*as_ pertained*, has been used both in WN and in EWN. It allows the link of a noun with a relational adjective.

In IWN this relation applies either between synsets or between synsets and istances: it connects 2° order entities with 1° entities, or 2° order entities and istances:

dantesco (dantean)	P*ertains_ to*	Dante
musicale (musical)	P*ertains_ to*	musica (music)

Also proper names were linked with WordNet 1.5 by means of equivalence relations. The *Eq_synonymy* is used to link proper names with an equivalent istance in WN; in IWN the *Eq_belongs_ to_ class*, that was not present in EWN, is used to map proper names to the generic belonging class when they have no equivalent in WordNet.

Summing up, the following examples show all the types of relations so far encoded for this subset:

Roma	*Belongs_ to_ class*	città (city, town)
Romano	P*ertains_ to*	Roma (Rome)
Roma	*Derivation*	romanità (Roman world)
Roma	*Eq synonym*	Rome
Lucca	*Eq_belongs_ to_ class*	town

1.1.2 Coding geographic names

As said above, in order to choose the first nucleus of proper names in IWN, we referred to the corpus selected for the semantic annotation, where we found that geographic names such as Italia, Roma, Milano, New York etc. were among the most frequent in the list of occurrences. Furthermore, geographic names denote 'entities' that have a kind of 'stability' and originate adjectives and nouns of common type which should be linked to their bases. For all these reasons we decided to start the coding from these kinds of proper names.

The geographic names were subdivided into many types of semantic classes (over 25): nation, city, region, sea, lake, river, etc., and all (more then 1300) were linked to the class they belong to by means of the semantic relation *Belongs_to_class*, for example:

Firenze	*Belongs_to_class*	città (city)
Cuba 1	*Belongs_to_class*	isola (island)
Cuba 2	*Belongs_to_class*	nazione (nazione)

When coding we noticed that homonymy among nouns denoting different objects occurs either for 'instances' belonging to different classes (e.g. Cuba, Washington, New York), or for 'instances' belonging to the same class (e.g. Hebron, Tripoli, Cambridge). In the first case, we created one entry for each class. In the second case we used only the definition to distinguish two identical entries, such as, for example, Hebron in Canada and Hebron in Israel; but, in the future, we shall extend to this kind of geographic names the possibility of their being encoded by means of the relation *Has_holo_location* / *Has_mero_location* which would make explicit (for automatic applications in NLP) the different places where the homonymous towns are situated. Thus we will have:

Hebron	*Has_holo_location*	Canada
Hebron	*Has_holo_location*	Israel

Another phenomenon occurring within geographic names is that they have sometimes changed with the passing of time. In these cases the present names have been included in the database, and the older, but better known ones, have been included as variants; see for example the case of {Byrmania, Myanmar} or {Persia, Iran}.

1.1.3 Other kinds of names

Moreover from the TAL corpus and from the DMI[3], a file of over 250 records has been created, made up of names of famous persons that have given origin to adjectives and/or common names (e.g. Ario, Machiavelli, Parkinson, etc.). More than 70 types or classes which these names belong to as instances have been identified: i.e. sculptor (Fidia), painter (Modigliani), character of a novel or a drama (Hamlet), writer (De Amicis), philosopher (Plato), etc.. In a few cases a proper name denoting a person have been codified as instance of more than one class, e.g. 'Michelangelo', belongs 'conjunctively' to the classes 'painter', 'sculptor', 'architect'. A definition has been given to all these personages (about 300) using the De Agostini 'Compact Encyclopedia' as a reference. Some files have been extracted also from sources of various type: atlases, web sites, several lists containing names of famous persons, divinities, celestial bodies, etc., useful in extending the lexical coverage of IWN. They have been checked, tidied and reorganised and then added to the set as new entries.

Up until now the set of Proper Names contains 3600 instances, belonging to 200 classes. In the table below it is possible to see a few of the most represented classes:

Belonging Class	**No. of "instances"**
Città (city, town)	556
Museo (museum)	240
Teatro (theatre)	172
Porto (harbour)	153
Nazione (nation, country)	161
Popolo (citizenry, people)	129
Fiume (river)	126
Comune (municipality)	124
Regione1 (territory, district)	106
Divinità (divinity)	104
Parco (national park)	80
Cometa (comet)	76
Ditta (company)	75
Opera lirica (opera)	74
Stella (star)	70
Monte (mountain)	60
Lago (lake)	54
Luna (moon)	53
Isola (island)	41
Valle (valley)	39
Scrittore (writer)	34
Poeta (poet)	33
Filosofo (philosopher)	32

Table 1: A few of the most represented classes.

[3] The DMI is a dictionary realized in the seventies and contains about 106,000 lemmas, more than a million of inflected word-forms and 187,000 definitions for the three main parts of speech.

In the future we are aimed at increasing the lexical coverage in order to achieve a well reasoned extension of the database; in fact, the set of proper names already represented in IWN could be further increased either by adding new proper names classes, or widening the set of entries of the classes already represented, for instance the class of 'painters' or the class of 'writers'. All derivates from the proper names (common nouns, adjectives), as well the relations already identified but not yet included, will be added in order to make the semantic database richer and more complete.

2. Proper names and sense extension

On the basis of our experience in the IWN context, we shall illustrate a few semantic features, characterizing this part of the lexicon, which are worthy of some remarks.

In IWN, the 'synonymy' relation applies to the variants of a synset allowing to interchange the synonyms (or variants) in at least one proposition, and this kind of relation is valid also for the set of proper names as formalized below:

$$a = b \Leftrightarrow \{a.f(a)\} = \{b.f(b)\}$$

As far as proper names are concerned, in many cases we see that the variants differ from common nouns variants. Proper names may have variants (polilexical or not) showing relations more complicated and richer than common nouns, because the proper name and its variant/s are often linked by various type of semantic procedures, in particular 'antonomasia' and 'metaphoric periphrasis', v. 'the Aquinate' , or 'the Holy Father', or 'the Big Apple', to indicate San Tommaso d'Aquino, the Pope and New York respectively (Cohen, 1993).

In the case of antonomasia we see that a proper name is substituted by an epithet or by a periphrasis which evidences one peculiar quality or character as for example in: "il sommo poeta" (the great poet) or "il ghibellin fuggiasco" (the fugitive Ghibelline) to indicate (to denote) Dante, 'il flagello di Dio' (the God plague) to denote Attila, etc.

Other frequent figures are metaphor or metonymy which also regard the proper names. Many examples of this phenomenon have been evidenced by textual corpora. Particularly rich in metaphorical uses are newspaper articles, which employ an increasingly impressive language to capture the reader's attention. By means of these semantic procedures discourse is given total enrichment, a semantic 'surplus'. A metaphorical expression is the result of three phases: an association, an abbreviation (in fact the term 'as' is understood), a substitution, because the literal term is replaced by a term coming from another semantic field (different semantic fields); metaphor is a substitution on the basis of similarity, and, like similitude, it is not reversible: 'he is a true Adonis'.

We have often found sense shifting (or more properly reference shifting) by means of 'metonymy': whilst metaphor is based on apparent unrelatedness, metonymy is a function which involves use of one term instead of another which is directly related to or closely associated with it in a logical-causal or/and material-spatial contiguity. Whilst metaphor indicates an external relationship, because the two terms belong to different semantic fields, metonymy consists of a syntagmatic inherent relation, for example that between the author and the opera: 'to read Dante'; or between place and institution: 'Montecitorio made no comment'.

Sometimes the reference is 'extended' on the basis of an extensional contiguity i.e. by a synecdoche: 'the Talibans bomb the Buddha'. We observed various types of these reference displacements concerning the proper names which can be assimilated to typical cases of regular polysemy or regular sense shifting:

- The name of the commissioner becomes the name of the work of art/palace, so this one is well known by the name of the commissioner and not by the name given by the author or by the name of the author himself: (Tondo Doni, Palazzo Strozzi);
- The name of the discoverer of a physical phenomenon is transferred to the phenomenon and then to the tools, to the medical-biological analyses based on that phenomenon (Doppler effect, femoral Doppler);
- The initial words of a sequence become the title of that sequence (Dies irae, Magnificat, Requiem) and this title is then used metaphorically: 'they have sung the requiem for the Juventus';
- An object is given a name on the basis of an analogy (of matching of colours) with an event that is yet famous/well known… (Bloody Mary);
- The proper name (of a person, of a place, etc.) is transferred to the situation, to the quality that the name represents: 'This office is a true Babele', 'He is a little Dante', 'Sgarbi made an 'Amarcord'';
- From an emblematic word to the people represented by that word: 'the Islam approves..';
- The name of the person who planned/proposed/ etc. becomes the name of the prize/law/plan: 'Rubbia has been Nobel for physics', 'He won the prestigious Goncourt';
- From the geographic name to the sports style originated in that region (Telemark, Cristiania);
- The name of the place substitutes the institution that seats in that place and the human group: 'Palazzo Madama' to say 'Senato', 'Strasburgo' to indicate the Parlamento Europeo;
- The name of the town is the same as the basket or football team 'the Cagliari scored...';
- The name of a town is used to indicate also the set of its citizens;
- The name of a nation may denote its inhabitants;
- The name of the place where something is produced is transferred to the product (Cognac, Chianti, Gorgonzola);
- The name of the inventor/builder/creator becomes the name of the industrial unit and then of the manufacture itself (Goodyear, Lancia, Cartier);
- The name of the artisan or of the artist instead of the name of the opera (Stradivari, Caravaggio, Picasso).

When there is a shift of reference, also a change of the ontological 'value' may take place as a consequence: some proper names belonging to the 1° order entities can shift towards 2° or 3° order entities, for example the proper names indicating methodologies of various types: techniques of biomedical analyses (Doppler, Paul Bunnel, etc.), mathematical transformations (Fourier), methods to teach dancing (Cecchetti), biochemical methods (Western), surgical operating techniques (Fontan, Blalock-Taussig), playing strategies (Karpov),

architectural techniques for fortifications (Vauban), methods employed for practising sports (Telemark).

A problem arises about the ways of considering the proper names used in these figurative ways: it would seem that, when an antonomasia is used, the proper name may be considered as a variant ('The great poet' to say Dante), on the basis of the definition of antonomasia; on the other hand the proper names used metaphorically (for example Babele, Caporetto) are considered different senses of the same name: Caporetto 1 *Belongs_ to_ class* place, Caporetto 2 *Belongs_ to_ class* defeat, therefore Caporetto 2 can be considered also as a variant of the noun 'defeat'. In the case of antonomasia the proper name has polilexical units with common nouns among its variants (The fugitive Ghibelline); in the case of metaphor the common noun has a proper name among its variants, used frequently in a peculiar metaphorical manner.

Following the semantic chain again may be a method of testing whether there is a metonymic transposition: if Caravaggio links with the class 'painter' there is no transposition; if Caravaggio connects to the class 'painting', there has been a transposition. Making these reference shiftings explicit could be a useful improvement for the IWN database. This would be worthwhile, because proper names used in this particular way of sense extension are many and they are a very large number in percentage.

2.1 Corpus evidence

Considering the proper names in the database, 49 classes out of 194, that is 1920 proper names out of 3600 may have a regular polysemy with reference shifting.

A research has been carried out on the PAROLE corpus (Marinelli et al. 2002, forthcoming), containing over 20,000,000 occurrences, starting from a set of representative samples of proper names: each proper name that has been considered in the corpus presents more transpositions than expected, in most cases meaning extensions of various types were found. This is what happens to proper names like Maecenas, Cicero or Marathon: in this last case, for example, there has been an extension of the reference from the place of Greece to the athletic specialty and then to the true metaphorical sense, used in expressions like: 'Maratona di leggi' (Marathon of laws), 'maratona di beneficenza' (benefit marathon), etc. Moreover, it has been verified that the frequent use of a proper name in metaphorical sense makes it become a common noun: this phenomenon was called "frozen or dead metaphor": "once a metaphor freezes (or dies) it becomes an ordinary part of our literal vocabulary" (Grey, 2000).

Sometimes the proper name looses the capital initial letter, sometimes this remains. It is not possible to test the reasons why it happens, even in similar contexts.

In the following table some results are shown, obtained from our survey about proper names and their use in common sense expressions.

With regard to each proper name, a number represents the frequency in the corpus, in how many texts it occurs, how many times the literal sense has been used, and the extended/metaphorical one.

Name	Frequency	Texts	Proper use	Metaph.use
Siena	523	145	488	35
Nobel	249	111	12	237
Maratona	226	95	9	217
Strasburgo	221	79	171	50
Palazzo Madama	211	108	30	181
Caravaggio	75	34	16	59
Babele	70	48	52	18
Merlin	61	23	12	49
Chianti	61	27	52	9
Doppler	56	7	-	56
Caporetto	51	31	30	21
Amarcord	33	26	10	23
Stradivari	13	8	10	3
Telemark	3	2	-	3
Total	1853	744	892	961

Table 2: data from PAROLE corpus.

3. Final remarks

We do not consider finished the work we carried out on proper names in the framework of the SI-TAL project, but as we said in the paper, we will go on improving this subset in quantity and in quality. We will code new data, on the basis of corpora evidence, and we will complete the sense shifting encoding. In fact the regular polysemy phenomenon, was made explicit only for geographic names which indicate either the territory or the people living there. In this case a double link was created to the two belonging classes, corresponding to the double hyperonymy that occurs also for some common nouns. This type of encoding will be extended to other various kinds of regular reference shiftings, for instance: 'writer/written', 'philosopher / philosophy-theory', 'painter/ painting' and so on. Making these metaphorical and metonymic uses of proper names explicit, we think that our database will gain in coherence and in usability in NLP.

4. References

Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti T., Peters W. (1998). The Linguistic Design of the EuroWordNet Database. In: Ide N., Greenstein D., Vossen P. (eds.), Special Issue on EuroWordNet. Computers and the Humanities, Vol. 32, Nos. 2-3 1998, 91-115.

Blanché, R. (1968). Logica e assiomatica. Firenze, La Nuova Italia.

Cohen, G. (1993). The Origin of NYC's Nickname The Big Apple. In: Names, Vol. 41, No. 1, 23-28.

Enciclopedia Generale De Agostini, (1997). Novara, Officine Grafiche De Agostini.

Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

Gray, W. (2000). Metaphor and Meaning. In: Minerva-An Internet Journal of Philosophy, Vol. 4 .

Lyons, J. (1977) Semantics. London, Cambridge University Press.

Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A., (forthcoming): "Criteria and Methods for building the Italian PAROLE Corpus". In: Linguistica Computazionale, Giardini Editori, Pisa.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, Vol.3, No.4, 235-244.

Pamp, B. (1985). Ten Theses on Proper Names. In: Names, Vol. 33, No. 3 1985, 111-118.

Rodriguez H., Climent S., Vossen P., Bloksma L., Roventini A., Bertagna F., Alonge A., Peters W. (1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In: Ide N., Greenstein D., Vossen P. (eds.), Special Issue on EuroWordNet. Computers and the Humanities, Vol. 32, 2-3, 117-152.

Roventini A., Alonge A., Bertagna F., Calzolari N., Marinelli R., Magnini B., Speranza M. (2002). "ItalWordNet: a Large Semantic Database for the Automatic Treatment of the Italian Language" in: Proceedings of the First Global WordNet Conference, Central Institute of Indian Languages, Mysore, India, pp.1-11.

Roventini A., Alonge A., Bertagna F., Calzolari N., Cancila J., Marinelli R., Zampolli A., Magnini B., Girardi C., Speranza M., (forthcoming): "ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian", in Linguistica Computazionale, Giardini Editori, Pisa.

Russell, B. (1919). Introduction to Mathematical Philosophy. London, George Allen and Unwin.

Vossen P. (ed.) (1999). EuroWordNet General Document, http://www.hum.uva.nl/~ewn.