Andrea Bocco, Enrica Bodrato, Antonella Perin*

ARCHIWORDNET, A BILINGUAL THESAURUS FOR ARCHITECTURE AND BUILDING: COMPILATION AND APPLICATION TO HYBRID ARCHIVES

Abstract. Linguistic resources with domain-specific coverage are crucial for the development of application systems, especially when integrated with domain-independent resources. In this paper we discuss our experience in the creation of ArchiWordNet (AWN), an English-Italian thesaurus for architecture and building which is being created according to the WordNet model and integrated with multilingual Princeton WordNet itself.

The project was born out of the cooperation between the Bruno Kessler Foundation (FBK), which offers the computational linguistics competences, and the Politecnico di Torino, which brings in the domain-specific knowledge. The original aim was to support the management of a Still-Image Database and some archival collections at the School of Architecture. Both include analog as well as digital documents, carrying both visual and textual information.

The WordNet model was chosen because of its structure, conceptually and relationally more rigorous than traditional thesauri. In fact, concepts in WordNet are represented by sets of synonymous terms ("synsets"), and relations between them are explicit and homogenously codified. This allows for more coherent and meaningful results from retrieval operations. AWN makes maximum use of:

1) information contained in WordNet, when acceptable or adaptable by domain specialists;
2) reference sources in Italian and English, among which *AAT*, CI|SfB's *Construction Indexing Manual*, international standards, technical dictionaries.

AWN will include some 10,000 "synsets". The "Materials" and "Single buildings and building complexes" hierarchies have been populated (with some 3,500 synsets), while the "Components of buildings" one is being developed now.

As the project went on, we had the chance to develop AWN for two Regione Piemonte programs – *Guarini Archivi* database for architectural archives, and *Guarini Censimento* databank of architectural heritage –, allowing us to test in progress its completeness and efficacy. In the validation of *Guarini Censimento* records, we revised 6.622 terms used, and designed automatic procedures to align them with parts of AWN: one for every applicable short-text, controlled-vocabulary field.

In the future, *Guarini* and AWN will be interoperable: terms used in the database will be uniform as they're imported from the thesaurus in indexing and retrieval operations; meanings and semantic relations will be easily checked.

Finally, the *KX* software, developed by FBK, will permit to automatically extract key-concepts from free-text description fields, abstracting over morphological and lexical variance, and will complement the manual operation of indexing.

* Politecnico di Torino – Dipartimento Casa Città (DICAS), viale Mattioli 39 - 10125 Torino, Italy

1 Introduction

This paper constitutes an update of what we delivered at the Second International WordNet Conference held in 2004 at the Masaryk University in Brno (Czech Republic) [17] and at the Italian Association for Terminology (AssITerm) Conference last year at the Università della Calabria (Italy) [22].

The ArchiWordNet (AWN) project was born out of the co-operation between the Bruno Kessler Foundation (FBK, formerly Istituto Trentino di Cultura) and the Politecnico di Torino, with the aim of creating a thesaurus for architecture and building, to be used within a databank of architectural photographs (Still Image Server – SIS), and some archival fonds managed by the "History and Cultural Heritage Laboratory" (LSBC).

These constitute two separate systems. SIS is a project created for educational purposes by four different departments of our two Schools of Architecture. The section our department is in charge of manages some 17.000 digital-born as well as digitised analogue photographs. SIS is accessible through the Politecnico Intranet: therefore, unfortunately, not open to Internet consultation. This is due to the fact that albeit many photos are copyright of the Politecnico or Politecnico professors/alumni, some were reproduced from various external sources. The contents of the archive are catalogued and organised in a purpose-tailored relational database developed on an *Oracle*™ platform; the cataloguing card was designed on the basis of VRA Core Categories scheme. A special effort is made to compile quite extensively the content description (Subject) field: a descriptive text associated with the image where remarks by the author of the photograph as well as the curators and even the users may be recorded.

LSBC, another part of DICAS department, conserves ten architectural archives (Bonamico, Brayda, Collettivo di Architettura, Lange, Melano, Melis de Villa, Mosca, Musso Clemente, Verzone, DICAS). Some were produced by the Department itself, others were gifted or bought. In total, some 17.000 drawings and 4.000 photographs plus 6 cubic meters and 90 linear meters of documents still to be catalogued. Basic information about such fonds is accessible through the web [http://www.regione.piemonte.it/guaw/MenuAction.do], while the documents themselves are now accessible through on-site consultation only. They primarily consist of drawings and photographs, although many fonds include also correspondence, notebooks, contracts, specifications and invoices. In 1998, LSBC decided to adopt two software produced by the Regione Piemonte: *Guarini Patrimonio culturale* and *Guarini Archivi*. The fonds are described according to standards defined by the international councils for archives – ISAD(G) and ISAAR – and the Italian Central Institute for Cataloguing and Documentation – ICCD –.

(A couple of items, one from each archive, are exemplified in section 6 of this paper.)

The two archives envisaged the advantage of sharing some indexing standards and procedures not provided by institutional bodies. Both, in fact, needed to index and describe their holdings (and in particular, their content). To facilitate image browsing/retrieval, the words used by cataloguers and final users should be systematised through a thesaurus. This should include both technical and general language given the fact that the architecture and construction sector uses as any other discipline a lot of jargon, but also inherits a lot of words from the laymen's language. This is one of the reasons why we decided to create AWN, a bilingual thesaurus (English/Italian), integrated in the general language WordNet.

In this text we present our experience in the creation of AWN and its application within the two archives above mentioned, as well as within *Guarini* regional databases. In paragraph 2 the principles which guided our choice to build a "WordNet-like" thesaurus and its characteristics are described; in paragraph 3 we deal with some of the main problems connected with the creation of a domain-specific linguistic resource integrated in a general-language one and the solutions adopted; in paragraph 4 are given some remarks about the compilation of the {building} hierarchy. In paragraph 5 we describe the work on *Guarini Censimento* regional database, while in paragraphs 6-7 two examples, one from each of our archives, are briefly described, using them to highlight a few features and problems . Finally, in paragraph 8, we briefly mention possible future developments.

2 ArchiWordNet: a WordNet-like thesaurus

The main feature of AWN is that, while referring as much as possible to existing architecture and building thesauri and other specialised sources, it is structured according to the WordNet psycholinguistic model created at the Princeton University [11], and fully integrated into it.

AWN is different from traditional thesauri with respect both to concepts and relations [13]. In fact, thesauri usually represent concepts through a controlled lexicon where synonyms and local variations may be missing. Also, they include few relations, the most relevant being hypernymy/hyponymy: i.e. the so-called 'ISA' relation. On the contrary, concepts in WordNet are represented by sets (called 'synsets') including those words which are used as synonyms in current (general or domain-specific) language, and relations are numerous (between nouns: IS PART OF, IS MADE OF/HAS SUBSTANCE, HAS ROLE, HAS FORM, IS MEMBER OF) and explicit, therefore enabling inheritance and a certain semantic richness (if the thesaurus is compiled fully exploiting such potential). Moreover, the accent is on the logical structure, and on the concept definitions, much more than the words, thus creating a virtually trans-lingual organisation of knowledge. Given such differences, we decided to adopt the WordNet because its structure allows for a powerful retrieval mechanism, and is particularly suitable for educational as well as research purposes, as it provides conceptual frameworks which can support learning. Its well-structured hierarchies can be in fact browsed to form both a general idea of the architecture and building domain, and a logically organised knowledge of some specific topic.

AWN and traditional thesauri differ not only in their structure, but also in the fact that the first fully integrates a domain-specific linguistic resource a general language thesaurus (WordNet). WordNet offers a general and multilingual framework for the specialised knowledge we are inputting through the compilation of the section called AWN. Moreover, given the huge cost in terms of human effort, involved in the construction of such a linguistic resource, the integration is particularly favourable, as information already available in general-language WordNet is directly imported in the specialised AWN.

AWN is being compiled with a constant tension between the diverging methodologies and inherent ends of the two disciplines involved in the project: computational linguistics and architecture. More specifically, we had to find a trade off between compiling a logically well-formalised linguistic resource, suitable for Natural Language

Processing applications, and creating a tool geared to meet the practical needs of "domain specialists" (all kinds of people involved in the architecture and building sector). This interdisciplinary research turned out to be very fruitful.

In the creation of AWN we had to face a number of problematic issues related both to the adoption of the WordNet model, and to the integration within WordNet itself. In the following section we discuss the steps we undertook in order to build such a resource.

3 Adopting and adapting the WordNet model

Two basic criteria are being followed in the construction of AWN. First, we refer as much as possible to acknowledged specialised sources for the architecture and building domain. Second, WordNet information is exploited whenever possible.

Various domain-specific sources are being used to create both the synsets, and the hierarchies of AWN, among which the *Art and Architecture Thesaurus* (AAT) [5], the *Construction Indexing Manual* of CI|SfB [6], the international and national terminology standards [9; 10; 20], and other scientific literature, including technical dictionaries [1; 2; 3; 4; 7; 8; 12; 15; 16; 18; 19] and websites.

We try to give preference to bilingual sources, but, as these aren't many nor detailed, we cannot avoid using a great number of monolingual (English and Italian) sources to populate the thesaurus with synsets which include matching words from the two languages.

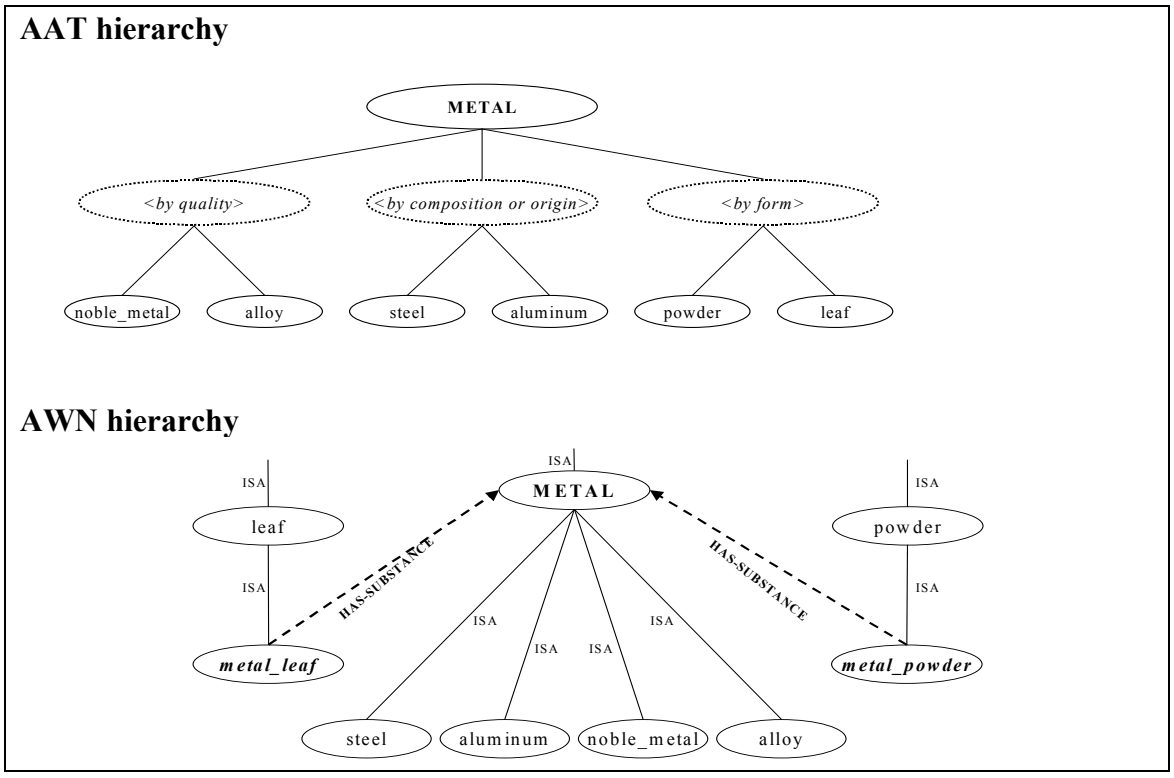

Fig. 1. Reorganisation of the AAT hierarchy for "metal" according to the WordNet model

Unfortunately very seldom the sources are directly compatible with the WordNet model, either because they're not structured on the basis of the ISA relation, or they

present mixed hierarchies where different levels are not homogenous and relations between concepts are underspecified and/or ambiguous. Thus, we usually need to reorganise the information to make it compatible with the WordNet model.

For instance, to retrieve information from AAT in the compilation of the AWN {material} hierarchy, we had to interpret its spurious relations by disambiguating the type of relation connecting superordinate and subordinate concepts and by deciding how to manage intermediate "artificial" nodes which are not relevant from the point of view of the ISA hierarchy. As it can be seen from the excerpt in Figure 1, when we reorganised the AAT hierarchy for the term "metal", the artificial nodes have been cancelled and only the ISA relations have been maintained. The concepts previously connected to "metal" by a "form" relation have been modified, put in their appropriate ISA hierarchy, and connected to "metal" with the HAS SUBSTANCE WordNet relation.

Another example of the same issue is furnished by the "diverted" logic which seems to guide some terminology standards. Typically, the upper levels are built according to ISA relations, but, as the hierarchies become more detailed, they tend to switch to meronymy ("IS PART OF") relations. When we tackled the {*building element*, *architectural element*} hierarchy, we found it correct to organise it according to ISO standards which break down the building in {*load-carrying structure, load-bearing structure*}, {*enclosure, building envelope*}, {*partition*}, etc., according to the main functional role of each part.

For instance, {*exterior wall, external wall*, *outer wall, outside wall, side wall*} and {*roof*} are kinds of {*enclosure*, *building envelope*}; going down and specifying, we can recognise that {*pitched roof, sloped roof*} is a kind of {*roof*} and that {*gabled roof, pitched roof, saddle roof*} is a particular case of {*pitched roof, sloped roof*}. However the parts which compose a {*roof*} – the {*roofing, roof covering, roof cladding*}, for instance – cannot be classified as hyponyms of {*roof*}: the statement "the {*roofing, roof covering, roof cladding*} IS A kind of {*roof*}" would be, in fact, false; the {*roofing, roof covering, roof cladding*} is, instead, a part of it, one of its layers.

The same logic applies to the analysis of other primary building elements: a {*leaf*} IS a PART OF, not A kind of {*wall*}, even if it obviously is a {*building element, architectural element*}, as the {*wall*} itself. Concepts like {*roofing, roof covering, roof cladding*} and {*wall covering, facing*, *revetment*} are therefore classified in the branch of the {*building element, architectural element*} hierarchy which includes the {*covering, outer layer*}, and each of them is linked by meronymy relations to the set(s) to which they usually belong in an actual edifice.

Such a problem does not exist for those words which stand for parts coinciding with single building products: a {*roof tile, roofing tile*}, for instance, is categorised as a {*building product, construction product*} (a particular class of {*artifact, artefact*}), not as a {*building element, architectural element*}; even if obviously may be used to construct a {*roof covering, roof cladding*}. You guessed it, {*roof tile, roofing tile*} is linked to {*tiled roof*} by a meronymy relation.

The second main source for the creation of AWN, mainly used when a complete and structured domain-specific terminology is not available, is WordNet itself. Synsets already existing in WordNet, which are considered appropriate by the domain experts, are included into AWN. However, this method cannot be applied straightforwardly. In fact, as WordNet synsets represent general language while AWN must represent a

specialised language, it is possible that WordNet synonyms and/or relations are not always completely suitable for representing the architecture and building domain.

When included into AWN, WordNet synsets can undergo three different kinds of modification:

1) in those cases where the criterion for synonymy suitable for WordNet is inadequate for AWN, it is possible to add or delete synonyms. This can happen as words that are considered synonyms in everyday usage may not be synonyms in the architecture domain.
2) when a general language definition is not compatible with a technical one, it is possible to modify it.
3) relations between synsets may be deleted and/or added. When included into AWN, a synset can maintain all or some or none of its original WordNet relations, depending on their appropriateness to the architecture domain. On the other hand, new relations can be added to encode further information relevant to the specialised domain.

To integrate AWN with WordNet, a first list of 10,000 nouns has been created extracting them from the specialized sources above mentioned, as well as the direct knowledge of the domain experts. The majority of such terms has been grouped in 13 semantic areas: Architectural styles, Materials, Construction products, Techniques, Tools, Components of buildings (including Spaces as well as Building elements), Single buildings and building complexes, Physical properties, Conditions, Disciplines, People, Documents, Drawings and representations. These semantic areas correspond to the main hierarchies which are being developed in the AWN project (one more advantage of the WordNet model is that it is structured as to include not only nouns, but also adjectives and verbs; however, these have not been taken into consideration for AWN so far).

After the identification of the WordNet nodes where to plug the AWN hierarchies, the integration procedure requires the actual inclusion of AWN hierarchies in WordNet, and the handling of the overlapping between terms present in both WordNet and AWN. The latter requirement is due to the fact that, unlike other domains characterised by a very specialised terminology, the architecture and building domain includes a significant amount of terms commonly used in the general language.

As regards the population of AWN, up to now around 4,000 synsets have been compiled, containing in most cases both Italian and English synonyms along with an accurate bilingual definition.

4. An example of AWN hierarchies: The "Building" hierarchy

The first hierarchical node we faced was {*building, edifice*}. As we already mentioned, to edit this hierarchy we referred as much as possible to specialised language resources, always putting into question the pre-existing WordNet synsets structure.

The *building* node derives from the hierarchy:

Entity, Something
Object, Physical object
Artifact, Artefact
Structure, construction
<Structure (CEAr)>,

where the latter, which is the direct hyperonym of {building, edifice}, is an artificial node we created as a hyponym of the pre-existing {structure, construction} in order to distinguish those constructions which are relevant to the Civil Engineering and Architecture domain.

The dominant criterion we adopted in the creation of the {building, edifice} hierarchy was "use (function)" because less ambiguous than "typology". Defining typologically an object means to describe it essentially on the basis of the morphology (organisation of space, volume, etc.) or even on the basis of cultural/historical *parti pris* rather than of its functional role. Collocations like {*central-plan building*} or {*longitudinal plan building*} inform us about the morphology of a constructed object and not about its function, which, indeed, may have changed in time. Think to churches converted into auditoriums or industrial buildings and convents turned into museums and schools. However, when typology or other criteria of building classification (e.g. form, dedication, possession, etc.) are widely used in architectural studies, relevant synsets were included in AWN. Consider, for example, words like {*skyscraper*} or {*heraion*}.

Sixteen synsets are direct hyponyms of {*building, edifice*}: nine ({*pile building*}, {*tower*}, {*tholos*}, {*loggia*}, {*shed*}, {*labyrinth, maze*}, {*mole*}, {*skyscraper*}, {*rotunda*}) are concepts associated with certain building forms, while the remainder are artificial nodes grouping buildings according to their functional role ({*agricultural building*, *building for the primary sector*}, {*industrial building*, *manufacturing building*, *building for the secondary sector*}, {*<services building, building for the tertiary sector>*}, {*religious building*}, {*multipurpose building*}, {*outbuilding*}, {*residential building*}). Many of these categories derive from economic sectors. From each node descends a number of synsets, more and more detailed as the hierarchy levels proceed down, e.g.:

Building, Edifice
<Services building, Building for the tertiary sector>
Medical building, Health facility
Hospital, Infirmary
Children's hospital, Paediatric hospital, Pediatric hospital

(For the sake of the argument, a possible further hyponym, say, "children's orthopaedic hospital", would be a hyponym of both {children's hospital, paediatric hospital, pediatric hospital} and {orthopaedic hospital, othopedic hospital}, a possibility inherent with the WordNet model).

Of course some concepts do not fit easily into the {*<structure (CEAr)>*} vs. {*building, edifice*} levels dichotomy, ex. *tollgate* and *subway station* are quite borderline entities. In many cases, the function housed is cognitively stronger than the container where it is performed: take the example of *florist's* which may indicate alternatively a single, mono-functional, perhaps small *building* and a shop which is part of a larger edifice, i.e. a *building space*. We chose to enter all such terms in the {*building, edifice*} hierarchy to have it as complete as possible; but, it seems to us, there are concepts which a "MAY BE A" (in *lieu* of "IS A") relation would better apply to.

The "Building" hierarchy, which includes 671 synsets, is made public in the occasion of this conference. This part of AWN is now finally on line under the MultiWordNet portal at [http://multiwordnet.itc.it/english/home.php]: still a work in progress we invite all of you to consult. Remarks and criticisms will be most welcome.

5. Testing ArchiWordNet: an application to *Guarini Censimento* databank

As the project went on, we had the chance to develop AWN for two Regione Piemonte programs – *Guarini Archivi* database for architectural archives, and *Guarini Censimento* databank of architectural heritage –, allowing us to use it for the indexing of external databanks, while testing in progress its completeness and efficacy.

In 2007, the Directorate for Cultural heritage of the Regione Piemonte commissioned us to validate the records of regional architectural heritage census (R.L. no. 35/1995). Here we revised the 6.622 terms used in the short-text, controlled-vocabulary fields.

Many of the problems we found derive from the fact that authority files were not provided by the system, therefore the compilers freely used the words they felt appropriate according to their culture, which means the filing cards were full of imprecise lexicon. There were also too case-specific, non-significant words (e.g. "squeeze tubes factory" – if such a concept is acceptable, then it may fit into AWN; but this does not automatically mean that such a detail level is appropriate in the regional architectural heritage census); misspelled words ("bowindows"); incorrect interpretations on the compilers' side ("most holy rosary recited in May"); unnecessary duplicates ("house" ≠ "House", "disused" ≠ "into disuse", etc.).

The validation operations have been typically:

- verifying the words used both through formal and logic operations, also referring to discipline-specific literature;
- ascribing synonyms and inappropriate words to the correct synsets;
- validating the terms and creating authority files (where the relevant part of the hierarchy did not yet exist in AWN). Anyway, the lists will stay open in order to accommodate additions and changes in time;
- cross-checking the validated nouns with AWN, to verify the correctness of the gloss, the semantic correspondence of the set of synonyms, the appropriateness of the hierarchic structure, and possibly to add and/or emend words. This feedback enriched AWN with 102 missing synsets, or emendations to existing ones;
- designing automatic procedures to substitute the words used in the records aligning them with parts of AWN;
- reviewing the *Guarini user's guide* incorporating, among other things, the new authority files (in form of synset hierarchies complete with glosses) and flux diagrams to facilitate the choice of the appropriate term(s).

Sometimes, during the validation process, we needed to examine a few records and the attached photographs. In any case, our intervention was limited to formal correction, and could not check the appropriateness to each particular case of every term used. Moreover, the problem of imprecise lexicon applies not only to the short-text, controlled-vocabulary fields, but all the free-text cataloguing fields.

For instance, the 338 terms used to classify roof forms have been reduced to 39. To begin, we selected the acceptable terms, and then we cross-checked them to recognise possible synonyms (e.g. "a due falde", "a capanna", "a doppia falda", "a schiena d'asino", all meaning "gabled" [roof]). As a rule, we preferred the synonym which expresses a geometrical form, and this was chosen as the first of the set, excepted where other terms are more widely used in current language and are felt acceptable by domain experts (e.g. "basilical" [roof]). To facilitate the compilers' work, a glossary of roof forms has been edited: synonyms and definitions are illustrated with schematic

drawings. To choose the correct term we advised to refer to the cross section of the roof, without regard to the building's shape in plan.

Finally, the Regione Piemonte has declared the intention of publishing on line the *Guarini Censimento* database. Such a task will require the migration of data to a new software, the present one having heavy limitations. For instance, no information browsing is currently possible, which makes this tool useless for valorisation policies on the architectural heritage.

6 Application examples from the Politecnico archives, Example no. 1 (LSBC)

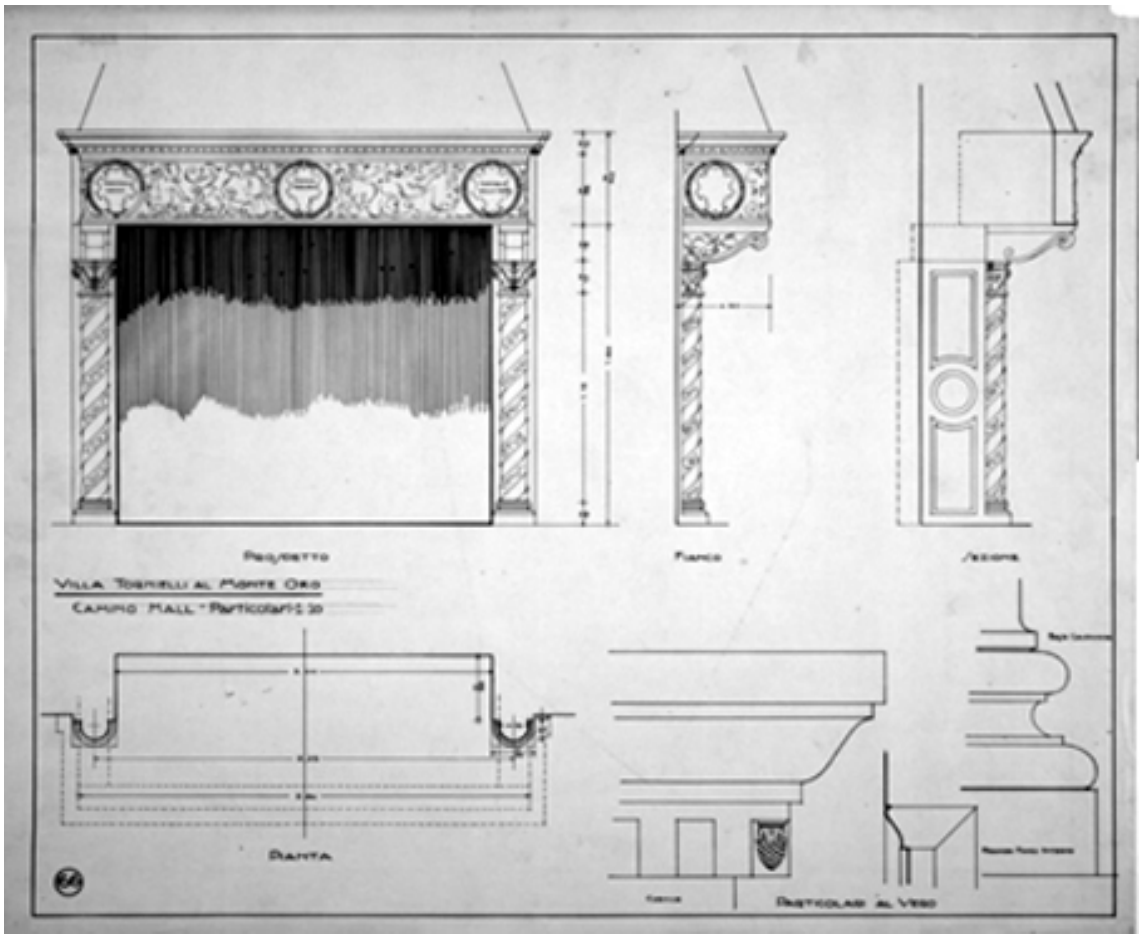

Fig. 2 – Politecnico di Torino, Dicas-LSBC, Melis fonds, Villa Tornielli al Monte Oro, 66

The drawing here reproduced belongs to the Melis de Villa fonds, which has been inventoried using *Guarini Archivi*. In this fonds various graphic documents created by architect and urban planner Armando Melis de Villa (active in Italy and its colonies between 1925-1956) are collected. The drawing refers to the project of villa Tornielli, which was built between 1924-1930 at Armeno (province of Verbania).

At present, *Guarini Archivi* does not include nor it is linked to an external thesaurus; the only terminology control means are vocabularies and word lists. Two fields of paragraph "subject indexing" are fed from a word list. The cataloguers have no other choice than separately consulting the AWN thesaurus, as a static source from which they extract the terms they will use to fill the cards, enriching the related word list. In

such operation the term chosen for indexing purposes loses its relations, both horizontal with synonyms, and vertical (ISA relation) with hyperonyms. We tried to save, albeit partially, the latter, exploiting the multilevel, hierarchic structure of the archival description: in high-level inventory cards (fonds, series) general terms pertaining to high positions in the thesaurus hierarchy are preferred, while more detailed terms are used in the descriptions of inferior level entities.

In the case of the drawing here reproduced, the Subject is indexed as follows: in the Fonds card the term *building* is used, among others; in the file card the terms *residential building* and *villa* are employed; finally, in the item card one finds terms which identify the 'content' of the drawing: *entrance hall, fireplace, mantelpiece, manteltree, column, frieze, cornice, modillion, capital, base, torus, scotia, listel*. This example shows that we stuck to an object description of the visual document, and left aside a stylistic one, which would have certainly called for further terms.

Besides, in the indexing of an archive document, one must also take into consideration those words which are written on it. Some of them may be archaic, uncommon or semantically diverged from the meaning they had when the document was originated. In this example three similar cases occur: *fianco* (lit., *side*) used to indicate a *side elevation*, *hall* to mean an *entrance hall* (which in Italian is usually called, *androne*) and *al vero* (lit., *life-scale*) to indicate the *scale of 1:1*. The words written on the drawing are ascribed to the appropriate AWN synsets, possibly adding them as synonyms.

Describing the visual document here presented doesn't obviously mean to identify its content only – it also involves to represent its extrinsic characteristics which, in further fields, call for the use of terms pertaining to other AWN hierarchies: for instance, the identification of support and medium (*butter paper, black China ink pen*), representation technique (*plan, elevation, section*); scale (*1:10, 1:1*).

7 Application examples from the Politecnico archives, Example no. 2 (ECO)

The image here proposed belongs to fonds no. 37 "Alpine Rural Architecture" (ARA) which includes more than 10.000 photographs shot in around 70 municipalities by professors, assistants and graduate students of our school. Its scope is to investigate vernacular architecture, in its environmental as well as technical and cultural context. This is one out of the 43 fonds, essentially consisting in photographic documents, which our sector DICAS-ECO manages through the SIS archive we mentioned in paragraph 1. The information is organised in "folders" (one for each set of photographs, preferably one for each building) and "cards" (one for each image), purpose-structured as described in [14].

In particular, our example is a house in Joussaud, a village of Pragelato (province of Torino). The "Description" field contains an analysis of the image content – the *south elevation* of a *rural building*, where three kinds of *enclosure* are recognisable (from bottom up):

- very thick *load-bearing wall*, made of mixed-sized *stones*, *plastered* with *lime mortar* which acts also as *binder*;
- *wooden frame structure*, with small stones *infilling*;
- *wooden frame structure*, with wooden *planks infilling*.

Fig. 3 – Politecnico di Torino, DICAS-Eco, fonds no. 37, series 2-Pragelato, eco_037_002_AC_1446. Photo by Diego Cappellazzo

The Description is intentionally left "open" to integrate subsequent remarks, introduced by different people. The choice of a quite informal, free text field – just calling the compilers for correct disciplinary terminology –, allows to contain possibly rich information in a small number of fields. This will have to be coupled with a system able to recognise, in retrieval operations, not only inflections but also logic equivalences (past participle with function of qualifier; saying that a part of the building is "infilled" within a frame equals mentioning an *infilling*; the adjective "wooden" means that element is made out of *wood*, etc.). The scope is to increase the, even casual, interrelation possibilities between images whose Descriptions contain terms having the same meaning, or anyway semantically related according to the AWN model: the fact that any significant term occurring in a description might be used as an "access point" is the most promising result expected with future implementations of the system.

A specific characteristic of fonds ARA is to include, whenever possible, local names, particularly those related to building elements. These not only constitute a lexical heritage menaced by extinction, but express a local lifestyle: dialectal names often correspond to specific ways to build or to perform other cultural activities [21]. So far, the vernacular terms we collected haven't been integrated in AWN (as well as Italian outdated terms). Such an activity would be very time-consuming; it'd be possible to face it systematically in the long run only.

8 Conclusions and future developments

In this paper we have presented our experience in creating AWN and two of its applications as a thesaurus for cataloguing images within architecture archives.

In the next years, we'll populate new AWN hierarchies and publish on-line those as they appear enough structured and complete in both languages, which is the hardest task for a work team composed by Italian mother-tongue speakers only. With respect to this point, we're discussing about switching to an international 'virtual community' AWN compilation process.

What is more important, AWN is going to become a useful integrated resource for Natural Language Processing applications. *Guarini* databases and AWN will be made interoperable: the terms used in the first will be uniform, as they're imported from the thesaurus in indexing and retrieval operations; meanings (glosses) and semantic relations will be easily checked, drastically reducing errors and offering new opportunities to browsing. Only think to automatic ascription of terms used in free-text, Description fields to AWN synsets (as mentioned in paragraph 7).

We hope that the *KX* software, recently developed by FBK, will permit to automatically extract key-concepts from free-text description fields, abstracting over morphological and lexical variance. This will partially alleviate the manual operation of indexing.

Of course, AWN might be rendered interoperable not only with closed systems such as *Guarini* databases, but potentially also with interacting open systems such as, e.g., *Wikipedia*.

Finally, it appears extremely inviting to us, that an illustrated WordNet is being discussed now inside the WordNet international community. The association of AWN synsets with images representing the concept would be a powerful feature with respect to other lexical resources.

Bibliography

1. Pevsner N., Fleming J., Honour H. (1966), *The Penguin Dictionary of Architecture,* Harmondsworth : Penguin = (1981) *Dizionario di architettura*, Torino : Einaudi
2. Harris C.M. (1977), *Illustrated Dictionary of Historic Architecture*, New York : McGraw-Hill
3. Harris C.M. (ed.) (1984), *Dictionary of Architecture and Construction*, New York : McGraw-Hill
4. Huber, R., Rieth, R. (1985-2003), *Glossarium Artis*, München : K.G. Saur
5. Petersen, T. (ed.) (1990), *Art and Architecture Thesaurus*, New York : Oxford University Press [www.getty.edu/research/conducting_research/vocabularies/aat/]
6. Ray-Jones, A., Clegg, D. (1991), *CI|SfB. Construction Indexing Manual 1976*, London : RIBA Publications = Vetriani G., Marolda M.C. (1983), *Piano di classificazione PC|SfB*, Milano : ITEC editrice
7. Neuteboom J.H., Francescato S. (1992), *Dizionario tecnico dell'edilizia Italiano-Inglese Inglese-Italiano*, Milano : BE-MA Editrice
8. Leva Pistoi M., Molino M., Piovesana M.M. (1993), *Il Nomenclatore di Architettura*, Torino : Rosenberg & Sellier

9. ISO (1994), *Technical Report 14177:1994. Classification of information in the construction industry*. Geneva : International Organization for Standardization
10. Ekholm A. (1996), *A Conceptual Framework for Classification of Construction Works*, Lund : Lund University. [www.itcon.org/1996/2/paper.htm]
11. Fellbaum, C. (ed.) (1998), *WordNet: an Electronic Lexical Database*, Cambridge, Mass. : The MIT Press [http://wordnet.princeton.edu/]
12. Gaddi P. (1999), *Architecture, Furnishing and Building Construction Dictionary English-Italian Italiano-Inglese*, Asmara : Italian School
13. Clark, P., Thompson, J., Holmback, H., Duncan, L. (2000), "Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search", *Proceedings of AAAI/IAAI 2000*, Menlo Park, Cal. / Cambridge, Ma. : AAAI Press / The MIT Press
14. Cavaglià, G. (2001), *L'analisi fotografica e la comprensione del costruito. Dalle patologie edilizie al progetto tecnologico*, Torino : Celid
15. Di Luciano A. (ed.) (2001), *Enciclopedia dell'Architettura*, Milano : Garzanti
16. Galliani G.V. (ed.) (2001), *Dizionario degli elementi costruttivi*, Torino : UTET
17. Bentivogli, L., Bocco, A., Pianta, E., (2004), "ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge", *Proceedings of the Second International WordNet Conference 2004*, Brno : Masaryk University
18. Portoghesi P. (ed.) (2005-07), *Dizionario enciclopedico di Architettura e di Urbanistica*, Roma : Gangemi
19. Bulletti P. (2006), *Le parole del progetto*, Milano : Il Sole 24 Ore
20. ISO (2007), *Standard 12006-3:2007. Building construction - Organization of information about construction works - Part 3: Framework for object-oriented information,* Geneva : International Organization for Standardization
21. Bocco, A., Cavaglià, G. (2008), *Flessibile come di pietra. Tattiche di sopravvivenza e pratiche di costruzione nei villaggi montani*, Torino : Celid
22. Bocco A., Bodrato E., Perin A. (2008), "Archiwordnet, un *thesaurus* di settore integrato nel *WordNet* della lingua generica: compilazione e applicazioni", *AIDA informazioni*, 26:1-2, gennaio-giugno 2008

Andrea Bocco [andrea.bocco@polito.it] is an architect and holds a Ph.D. degree in architecture and building design. He is assistant professor at the Politecnico di Torino, where he teaches *Fundamentals of Building Technology*. He has been working on ArchiWordNet since 1999, and coordinates the ECO architectural photographs archive.

Enrica Bodrato [enrica.bodrato@polito.it] is an architect and archivist. She is a technician at the Politecnico di Torino, responsible for LSBC and for Casa-Città Department architectural archives. She has been working on Guarini software since 1998 and on lexical resources since 1999, on ArchiWordNet since 2001.

Antonella Perin [antonella.perin@polito.it] is an architect and holds a Ph.D. degree in history of architecture. She has a contract as a researcher at LSBC at the Politecnico di Torino. She has been working on lexical resources since 1999, on ArchiWordNet since 2001.