Representing Italian Complex Nominals: a Pilot Study

Valeria Quochi

Linguistics Department, University of Pisa
Via S. Maria, 36, 56126 Pisa; ILC-CNR, Area della Ricerca, Via Moruzzi, 1, 56100 Pisa.
valeria.quochi@ilc.cnr.it

Abstract

A corpus-based investigation of Italian Complex Nominals (CNs), of the form N+PP, which aims at clarifying their syntactic and semantic constitution, is presented. The main goal is to find out useful parameters for their representation in a computational lexicon. As a reference model we have taken an implementation of Pustejovsky's Generative Lexicon Theory (1995), the SIMPLE Italian Lexicon, and in particular the Extended Qualia Structure. Italian CN formation mainly exploits post-modification; of particular interest here are CNs of the kind N+PP since this syntactic pattern is highly productive in Italian and such CNs very often translate compound nouns of other languages. One of the major problems posed by CNs for interpretation is the retrieval or identification of the semantic relation linking their components, which is (at least partially) implicit on the surface. Studying a small sample, we observed some interesting facts that could be useful when setting up a larger experiment to identify semantic relations and/or automatically learn the syntactic peculiarities of given semantic paradigms. Finally, a set of representational features exploiting the results from our corpus is proposed.

1. Introduction

This paper will describe a corpus-based investigation of Italian Complex Nominals (CNs), of the form N+PP, aimed at clarifying their syntactic and semantic constitution. The main goal is to identify useful parameters that can be used when representing these multiword expressions (MWEs) in a computational lexicon. Studies and projects such as the LinGo MultiWord Expression Project[1] and Xmellt[2] drew the attention of the computational linguistics community to problematic issues concerning MWEs, especially in natural language processing (NLP). The present investigation is based on the work done by the Computational Lexicon Working Group (CLWG) within ISLE[3] (Calzolari et al., 2002a). CNs are a particularly difficult case as they exhibit internal cohesion together with a high degree of variability in lexicalization and language-dependent variation.

One of the major difficulties in dealing with Italian CNs from a monolingual perspective is that they instantiate well-formed syntactic patterns, which are, nonetheless, not totally predictable or, if so, highly ambiguous w.r.t. "regular" syntactic constructions. The claim is, thus, that some sort of semantic information is needed in order to adequately characterise them; syntactic evidence is not sufficient.

The main goal of the present study is to find out some semantic paradigms and their morpho-syntactic features that can be exploited for representational purposes. As a reference model we have taken an implementation of Pustejovsky's Generative Lexicon Theory (1995), the SIMPLE Italian Lexicon, and in particular its Extension of the Qualia Structure, the Extended Qualia Structure (EQS hereafter) (see for details and explanations Lenci et al. 2000). Theoretical inspiration for this pilot study comes also from an experiment on CNs within Generative Lexicon by Johnston and Busa (1999).

In the following section, we briefly present the Qualia Theory of Generative Lexicon. Section 3 outlines the issues involved in representing CNs, and in particular Italian Complex Nominals and Section 4 describes our data and methodology and discusses the parameters we have derived from the data. Section 5 suggests a representation of the syntactic and semantic constitution of our structures, following the proposals made by the ISLE CLWG.

2. The Qualia Structure and SIMPLE[4]

The Qualia Structure is one of the most interesting parts of the Generative Lexicon Theory (Pustejovsky, 1995) in that it decomposes the internal constitution of lexical items into 4 basic roles, thus allowing to systematically structure and specify the relationships among lexical items both paradigmatically and syntagmatically. The SIMPLE lexicon project implements this structure, further specifying for each role its possible relations in the EQS. In this study we have exploited the information contained in the Italian SIMPLE Lexicon: in particular the semantic type ascribed to the nouns occurring in our CNs (which in SIMPLE is coded, for our purposes, as their Template Type) and the relations specified in the EQS, i.e. the semantic type of both head and modifier nouns has been used to infer the semantic relation underlying CNs.

3. General Issues about Complex Nominals

Complex nominals are expressions with a strong lexical-like behaviour (both at the syntactic and semantic level), whose interpretation can be (at least partially) compositional, in the sense that it relies heavily on the particular semantic relation existing between the component elements. This relation is, however, covert and therefore difficult to retrieve. Despite the high degree of variation, the semantic relations between the elements are taken to be a function of the interaction of the semantics of both the head and modifier nouns. For this reason, our interest lies not so much in the "idiom-like" CNs (i.e. with

[1] A subproject within LinGo (Linguistic Grammars Online). http://lingo.stanford.edu/mwe/reading-group.html

[2] XMELLT stands for "Cross-lingual Multi-Word Expression Lexicons".

[3] ISLE stands for "International Standards for Language Engineering".

[4]"Semantic Information for Multipurpose Plurilingual Lexicons". A EU LE-Programme sponsored by DG-XIII.

almost completely idiosyncratic implicit relations), but in those CNs that appear to be quite regular both at the syntactic and at the semantic level, but which nevertheless pose problems if not recognised as units at some level of the linguistic analysis. Thus, we share here the broader view of MWEs adopted within the Xmellt project.

3.1. Italian CNs

Italian CNs of the form N+PP share most of the characteristics of noun compounds, (cfr. Lyons 1977, Downing 1977; Levi, 1978; Warren, 1978; Leonard, 1984; Quirk et al., 1985 among others). What makes them still more difficult to define and identify is that they are structurally similar, if not identical, to regular syntactic patterns, but constitute conceptual/semantic units. Some heuristics based on syntactic characteristics have been adopted to help the identification process, such as the absence of determiners in the PP, but none can be taken as a rule. Main discriminants seem to be semantic characteristics such as the reduced referentiality of the noun and the denotational function of the expression as a whole (Calzolari et al. forthcoming).

Italian CN formation mainly exploits post-modification; of particular interest here are CNs of the kind N+PP (like *coltello da pane* 'bread knife'); this syntactic pattern is, in fact, highly productive in Italian and this type of CN often translates compound nouns of other languages.

3.1.1. Morpho-syntactic description

Morphosyntactically, one significant peculiarity of N+PP complex nominals is that the modifier usually occurs either in the singular or in the plural form, depending on the type or even on the particular instance. On exclusively syntactic grounds, this appears to be a purely lexical choice.

From the syntactic point of view, only three prepositions generally occur: that is *a*, *di* and *da*, and no element can, normally, intervene between the elements of the construction, especially within the PP.

The noun in the modifier PP, moreover, tends to have no determiner; this, however, has proven a very weak criterion for identifying CNs in a text.

3.1.2. Semantic Properties

One of the major problems posed by CNs for their interpretation is the retrieval or identification of the semantic relation linking their components, which is (at least partially) implicit on the surface.

The presence of a preposition in Italian CNs, however, has been taken as an explicit mark of the underlying semantic relation (Johnston and Busa 1999:169). When confronted with corpus data, however, this assumption holds only at a general level. Our hypothesis is that we also need to take into account semantic information for the component nouns (which in SIMPLE is given by the TemplateType). This information, together with the preposition, can help identify the semantic relation that links the nouns, or, at least, we hope that it restricts the range of possible relations. Specific qualia relations, such as those included in the EQS, would then link the components of a CN in a principled way, i.e. through the qualia structure of the respective senses.

4. Data and Methodology

The present study is based on a collection of N+PP structures extracted from a representative corpus of contemporary Italian, the Italian PAROLE corpus, without any preprocessing.

To restrict the scope of the investigation we chose to ground the extractions on key nouns belonging to three subclasses of the Artifact semantic class of the SIMPLE Ontology: Instrument, Vehicles and Containers. We performed a preposition cooccurrence search, keeping only those expressions with the prepositions *a*, *di* and *da.*

The resulting data have been subsequently entered in a database manually adding syntactic and semantic features, i.e. PP_type, semantic relation, semantic type of nouns (see Figure 1).

Item	PP Type	Sem rel	H-Sem Type	M-Sem Type
scatola di vetro 'glass box'	PP_di	Made of	Container	Artif. Material
Barca a vela 'sailing boat'	PP_a	Has as parts	Vehicle	Part
Barattolo di miele 'honey pot'	PP_di	contains/obj.act	Container	Substance_ Food
Fucile da caccia 'hunting rifle'	PP_da	Used for (activity)	Instrument	activity

Figure 1: Organization of the data

Since we used untagged text and standard query tools the amount of data considered is not very large: about a hundred, semantically restricted, potential CNs. From this small sample, we observed some interesting facts that could be useful when setting up a larger experiment to identify semantic relations and/or automatically learn syntactic peculiarities of given semantic paradigms.

A few systematic syntactic-semantic paradigms can be identified: e.g. MADE OF, which is always realised with a PP*_di* and takes as modifier a noun belonging to the [Artifactual Material] or [Natural substance] class. This type, for example, allows no internal modification and no determiner within the PP. The MADE_OF pattern seems to apply to all semantic types of head nouns, combining with a 'material' or 'substance used as material'. This seems to be a fully regular and productive semantic paradigm; however, other more restricted patterns can also be detected. For example, for the Instrument class, in our data set, a HAS AS PART relation can be established when the modifier belongs to the [Part] semantic type, and is systematically realised syntactically as a PP*_a*.

A problem is still represented by the PP*_di* class of CNs. This is the most heterogeneous subset, except for the MADE_OF pattern; and it thus deserves more in depth investigations. In the PP*_di* type we find that almost every EQS semantic relation is possible, i.e. almost every qualia role can be involved, and the modifier nouns belong to various semantic classes. In most cases, nevertheless, the modifier expresses some property of the entity denoted by the head noun.

We feel that this kind of analysis is useful to discover regularities and also to identify those non-compositional CNs that must be treated exclusively lexically.

Given all the problematic issues concerning CNs, a more detailed classification of CNs types has often been called for. However, such a classification would be no easy task, because various parameters can be taken into account. Moreover, an approach that combines different parameters, seems preferable.

5. Representation of Italian CNs

Assuming we can decide what expressions have to be included in a lexicon, a major problem is how to represent them. This is still a very controversial point, and is strongly dependent on the theoretical framework adopted, on the specific tasks to be performed, and on the overall design of the lexicon. It depends also on the degree of idiosyncrasy of the specific phrases. Given all the different application needs, it has been claimed that a modular representation would be most appropriate (Calzolari et al., 2002b). We refer, in particular, to the proposal presented within the ISLE project. A MultiWord Expression Lexicon could be seen as one layer building on others, in particular on the morphological, syntactic and semantic layers. We will use some of the above observations drawn from our data as useful parameters in the representation of CNs.

5.1. Syntactic Representation

An exhaustive account of the syntactic behavior of Italian CNs must include a description of their internal syntactic composition, which relates each component to the corresponding syntactic unit and the whole to the syntactic pattern it instantiates, thus satisfying the requirements expressed within ISLE. The syntactic head and modifier must be indicated explicitly.

It is, moreover, necessary to specify the possibility of internal modification. Like English noun compounds, Italian CNs do not normally permit internal modification, but this is not always true. The degree of syntactic variability is also bound to the degree of lexicalization of the expression. Thus, we propose a feature ‘Modif’ with 3 possible values: default (modification is free but preferably lacking), blocked, restricted. In the last case, there is also the possibility of indicating the type of modifier that is allowed (AdjP, AdvbP etc). At the syntactic level, restrictions on the use of articles, if any, have to be specified, and the values we postulate are: default (no determiner allowed), free (any determiner can occur), def/indef (specifies whether the definite or indefinite article is allowed).

Syntactic composition: synU= fucile + PP= da(P) caccia(N)
Syntactic pattern: N+P+N
Syntactic head = N =fucile
Modifier: PP= da caccia
Use of article: no.
Modification: default.

Figure 2: Syntactic Representation of *fucile da caccia*, ‘hunting rifle’

5.2. Semantic Representation

At the semantic level, the most important piece of information to be made explicit is the semantic relation that links the components of the CN[5]. The idiosyncratic part of meaning can be added (manually by the lexicographer) by further specifying the Qualia structure of the entry for the CN. Moreover, we believe it useful to indicate the semantic head of the construction: although, in our data, this always corresponds to the syntactic governor of the phrase, there may be cases in which the two do not match[6]. Another characteristic that is useful to represent is the status of the senses of the elements of CNs. Three classes of CNs can be identified, according to the three different statuses that can be given to the senses (Lyons, 1977): one class includes CNs whose components are all senses used in contemporary Italian also in other contexts: e.g. *macchina da scrivere* ‘typewriter’; another class contains CNs where one of the components is used with the same sense only in a restricted range of CNs, so that it can be considered idiosyncratic: e.g. *pistola a salve* ‘gun with blanks’; finally one class of CNs, where at least one component is used exclusively in that particular context: e.g. *coltello a serramanico* ‘flick-knife’.

Each item of the CN can then be linked to the respective semantic unit, in the semantic layer. If these items are linked to existing semantic units, then they are senses that are used freely in the language. If one or both senses are not used in the language, then no link will be established. This is strongly related to the degree of lexicalization, which must, therefore, be indicated explicitly. We consider the following four degrees of lexicalization to be relevant: 1. Compositional (that is CNs that instantiate regular syntactico-semantic paradigms, probably created on line); 2. Institutionalised (regular, but high frequency CNs); 3. Lexicalised (CNs that present syntactic or semantic anomalies/idiosyncrasies); 4. Frozen (i.e. completely fixed, idiomatic expressions).

Finally, some other “conceptual” details, such as hyperonym, co-hyponyms, or synonyms, can be added, to fully specify the semantic place of the CN in the ontology and the relationships it has with other single and complex lexical items. Figure 3 exemplifies the semantic representation.

[5] Relation that we have identified on the basis of the semantic class of both head noun and modifier noun plus the particular preposition occurring.

[6] We think in particular about metaphorical CNs, which have been deliberately excluded from the present investigation.

semantic relation: Telic:used_for: (cacciare ‘to hunt’
Semantic head: SemU= fucile ‘rifle’
Modifier: SemU= caccia (the hunting)
Lex: Lexicalised .
Hyperonym: ‘rifle’
Co-Hyponyms: carabina, fucile da sub. …
Idiosyncratic meaning: Qualia Structure of <Fucile da caccia>
Telic:**used by**: umani/ cacciatori ‘humans/hunters’;
Telic:**used for**: uccidere ‘to kill’
object of the activity: animali (selvatici) ‘(wild) animals’.

Figure 3: Semantic Representation of *fucile da caccia.*

6. Conclusions

We have shown that useful syntactic features can be derived from corpus data, once semantic paradigms are identified, and acquired using a representational format similar to the one proposed within ISLE. Future work in this area will be devoted to developing strategies for acquiring them (semi-)automatically. At the semantic level, where the identification of the underlying semantic relation is crucial, an automatic acquisition of the relevant semantic information seems more problematic, at least for the time being. However, some semantic paradigms can be identified exploiting the semantic/ontological information encoded in the SIMPLE entries for each head and modifier nouns. This information is useful to determine, to a certain degree of adequateness, the semantic relations linking the components.

Given that our data is quantitatively limited, we hope to be able to find more generalisations once we enlarge our sample, and to perform some reliable quantitative analyses. Identifying more semantic paradigms would be useful for classification, parsing and representation purposes, especially through the investigation of the semantic behavior and constituency of those CNs that should, but do not, belong to a given semantic paradigm. Such CNs seem to be good candidates of lexicalised, frozen or metaphorical CNs.

To be able to assign values to the representational features described above, a corpus-driven approach is obviously more reliable and efficient than intuition. We thus intend to automatise at least part of this task in order to acquire some of the syntactic parameters, such as internal modification, lack or presence of determiner etc., from the corpus. Once a semantic paradigm has been identified, one could learn and/or update (automatically) its syntactic properties from corpora, possibly with statistical methods.

References

Calzolari, N., Lenci, A., Quochi, V. (forthcoming) “Towards multiword and multilingual lexicons: between theory and practice.” Proceedings of LP2002. Japan.

Calzolari N., Zampolli A., Lenci A. (2002a) Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative - Lecture Notes in Computer Science, 2276. 264-279.

Calzolari N., Fillmore Charles J., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. (2002b). Towards Best Practice for Multiword Expressions in Computational Lexicons. In Proceedings of LREC Las Palmas, Canary Islands, Spain, 29th, 30th & 31 May 2002. Volume V, Paris, The European Languages Resources Association (ELRA).

Downing, P. (1977). "On the creation and Use of English Compound Nouns." Language 53: 810-842.

Xmellt: Cross-lingual Multi-word Expression Lexicons for Language Technology: Multilingual Information Access and Management. International Research Co.operation. http://www.cs.vassar.edu/~ide/ XMELLT.html

Johnston, Micheal and Federica Busa. (1999). "Qualia structures and the compositional interpretation of compound." in E.Viegas (ed.) Breadth and depth of semantic lexicons. Dordrecht: Kluwer Academic Publishers.

Lenci A., Calzolari N., Monachini M., Ruimy N., Zampolli A. et al. (2000). “SIMPLE: A General Framework for the Development of Multilingual Lexicons”, International Journal of Lexicography, XIII (4): 249-263.

Leonard, R. (1984). The Interpretation of English Noun Sequences on the Computer. Amsterdam: Elsevier Science Publishers.

Levi, Judith N. (1978). The syntax and semantics of Complex nominals. New York: Academic Press.

Lyons, J. (1977). "The Lexicon." Semantics. Cambridge: CUP, 512-569.

MultiWord Expression Project http://lingo.stanford.edu/mwe/ and http://lingo.stanford.edu/mwe/reading-group.html

PAROLE Corpus. Pisa: ILC-CNR.

Pustejovsky, J. (1995) The Generative Lexicon, Cambridge, MA, The MIT Press.

Quirk, R. ; Greenbaum, R. and Svartvik, P. (1985) A Comprehensive Grammar of the English Language. New York: Longman.

Warren, B. (?1978). "Semantic Pattern of Noun-Noun Compounds." Acta Gothenburgensis Studies in English. Gotheborg: University of Gotheborg.