Annotating Qualia Relations in Italian and French Complex Nominals

Pierrette Bouillon[1], Elisabetta Jezek[2], Chiara Melloni[3], Aurélie Picton[1]
[1]TIM/ISSCO, Faculté de Traduction et d'Interprétation
Université de Genève
Bd du Pont-D'Arve 40
CH-1211 Genève

[2]Dipartimento di Linguistica Teorica e Applicata
Università di Pavia
Stada Nuova 65
27100 Pavia

[3]Dipartimento di Filologia, Letteratura e Linguistica
Università di Verona
Viale dell'Università 2
37129 Verona

E-mail: pierrette.bouillon@unige.ch, jezek@unipv.it, chiara.melloni@univr.it, aurelie.picton@unige.ch

Abstract

The goal of this paper is to provide an annotation scheme for compounds based on generative lexicon theory (GL, Pustejovsky, 1995; Bassac and Bouillon, 2001). This scheme has been tested on a set of compounds automatically extracted from the Europarl corpus (Koehn, 2005) both in Italian and French. The motivation is twofold. On the one hand, it should help refine existing compound classifications and better explain lexicalization in both languages. On the other hand, we hope that the extracted generalizations can be used in NLP, for example for improving MT systems or for query reformulation (Claveau, 2003). In this paper, we focus on the annotation scheme and its on going evaluation.

Keywords: compounds, annotation, Generative Lexicon, Italian, French

1. Introduction

The goal of this paper is to provide an annotation scheme for compounds based on generative lexicon theory (GL, Pustejovsky, 1995; Bassac and Bouillon, 2001). This scheme has been tested on a set of compounds automatically extracted from the Europarl corpus (Koehn, 2005) both in Italian and French. The motivation is twofold. On the one hand, it should help refine existing compound classifications and better explain lexicalization in both languages. On the other hand, we hope that the extracted generalizations can be used in NLP, for example for improving MT systems or for query reformulation (Claveau, 2003). The originality of the work is primarily justified by the proposed task: as in Seaghdha (2007), we annotate compounds in context, but for two less-studied languages in a comparable corpus. GL also provides us with a rich representation formalism that allows us to annotate the composition derivation i.e. how the qualia of the head is activated/exploited by the modifier (Pustejovsky, *et al.* 2008). This rich representation could help to obtain better interjudge agreement and simplify the task of automatic classification (Tratz and Hovy, 2010). In the rest of the paper, we first explain how we refine the existing compound classifications, then focus on the annotation scheme and its on going evaluation.

2. Compound Classification

2.1 Dataset

Our dataset comprises two classes of Italian complex nominals and their French equivalents: Noun-Noun structures (NN), usually dubbed as primary compounds, and prepositional compounds (NPN), largely attested in the Romance languages; see the table 1 below.

NN	It. *uomo rana*	Fr. *homme grenouille*	'frog man'
NPN	It. *bicchiere da vino*	Fr. *verre à vin*	'wine glass'

Table 1: NN and NPN structures

The set of NPN compounds in Italian is restricted to those including semantically light prepositions (i.e. *di* / *a* / *da*)

that are not followed by determiners (we adopt a broad definition of compound and include in our dataset complex nominals that do not comply with standard lexical integrity tests). French translations might include fully-fledged phrases if no corresponding compound is attested.

2.2 Compound classification

In the domain of theoretical studies, the existing taxonomic accounts of compound structures usually take into consideration (at least) the following factors:
a. presence or lack of formal/semantic head;
b. grammatical/semantic relation between the compound members. For instance, Bisetto and Scalise (2005) propose a six-output classificatory system based on the hierarchical structuring of the aforementioned criteria. Specifically, they identify three (grammatical) relations in compound structures, i.e. coordination, subordination, attribution, and each of these relations can be further specified along the endo-exocentricity criterion (i.e. presence/lack of formal and semantic head). Italian NN compounds, for instance, can be classified as subordinative (e.g. It. *centrotavola* 'centerpiece'), coordinative (*pub-pizzeria* 'pizzeria pub') and attributive (*stato cuscinetto* 'buffer state'); and they can be either endo- or exocentric (*terra-aria* 'ground(-to-)air' in an expression like *missile terra-aria* 'ground-to-air missile'). Baroni, Guevara and Pirrelli (2009) have further refined this classification and distinguished two subclasses of subordinate (NN) compounds: argumental, where a deverbal head holds an argumental/thematic relation with the nonhead (It. *raccolta rifiuti* 'collection+rubbish') and grounding, where the head does not license any argumental interpretation of the nonhead (*stanza server* 'room+server'). However, many subtle distinctions in the meaning conveyed by nominal compounds cannot be captured easily by these classificatory systems.

On these grounds, we propose to expand and refine the existing taxonomic accounts of (nominal) compounds by:

1) including NPN structures, often neglected in the literature on morphological compounds because of their unclear lexical vs. phrasal status. This class is widespread in Romance languages, where NPN compounds often realize their English NN-compound counterparts (e.g. En. *bread knife* corresponds to the It. *coltello DA pane*). NPN compounds are typically endocentric, with N1 acting as the formal/semantic head of the complex (exceptions can be found in the domain of metonymical/metaphoric compounds such as *testa di rapa* lit. 'head of turnip', 'meathead') and encode subordinative relations, of the Grounding and Argumental type, depending on whether the head is deverbal and accordingly projects argument structure (grounding: *coltello da cucina* 'kitchen knife' vs. argumental: *raccolta di frutta* 'fruit collection').

2) proposing a finer-grained taxonomy of nominal compounds based on the semantic relations encoded in Romance prepositional compounds (see also Celli and Nissim 2009, Girju *et al.*, 2009; Seaghdha, 2007). In particular, along the lines of Johnston and Busa (1999), Bassac and Bouillon (2001), Delfitto and Melloni (2009/forthcoming), we employ Qualia Structure (QS) as a heuristic tool to classify NN and NPN compounds that are non-argumental on the grounds of the semantic relation between N1 and N2, arguably restricted to the four relations expressed by *qualia* roles, as discussed in the next section.

3. Annotation framework and methodology

The annotation task involves tagging the semantic relation between elements in NN and NPN Italian compounds and their French translation equivalents. We assume two basic types of relations, i.e. Qualia and/or Argumental. In determining the set of Qualia relations to be annotated, we extend the annotation scheme proposed for nominal compounds in Pustejovsky *et al.* (2008) based on Bassac & Bouillon (2001). We distinguish four basic Qualia relations, i.e. Formal, Constitutive, Telic and Agentive. Each of them is distinguished with tags. The set of relations together with their interpretive correlates, the annotation tags and some target examples for each category is given in the table 2 below.

Relation	**Interpretive correlates**	**Tag**	**Example (It)**
Formal	N2 is a kind of N1 N1 has the shape of N2 N1 holds N2	is_a shape_of holds	*cane bassotto* 'dog+basset' *cuscino a cuore* 'pillow+A+heart' *bicchiere di vino* 'glass+DI+vino'

Constitutive	N1 is made of N2 N1 is a part of N2 N1 is spatially/temporally located in N2 N1 is member of N2 N1 has N2 as members	made_of part_of located_in member_of has members	*torta gelato* 'cake+ice cream' *centro città* 'centre+town' *casa di campagna* 'house+DI+countryside' *frutta di stagione* 'fruit+DI+season' *membro di partito* 'member+DI+party' *squadra di atleti* 'team+DI+athletes'
Telic	N1 has the purpose of (Predicate) N2 N1 is used for the activity N2 N1 has N2 as result/end goal N1 denotes the function which is N2.	Predicate used_for aims_at played_by	*treno merci* 'train+freight' *fucile da caccia* 'rifle+DA+hunting' *procedura di divorzio* 'procedure+DI+divorce' *ruolo da ministro* 'role+DA+minister'
Agentive	N1 is created/brought into existence/caused by N2 N1 is derived/extracted from N2	caused_by derived_from	*impronta di piede* 'print+DI+foot' *succo di frutta* 'juice+DI+fruit'
Argumental	N2 is an argument of N1	Argument	*raccolta di frutta* 'collection+DI+fruit'

Table 2: Relations and tags

Annotators will be asked to specify the semantic relation between the nouns and to tag the role played by the referent of each noun in the relation, choosing from the following list: ag=agent, cause, instr=instrument, source, loc=location, pt=patient, purpose, result, th=theme, time. For the Constitutive relation, the available relational roles will be part and whole. Also, we will ask the annotators to attach the broad semantic class associated with each noun (artifact, event, etc.) choosing from a revised version of the list of top types proposed in Pustejovsky *et al.* (2008): entity, abstract_entity, human, animate, organization, natural, artifact, substance, event, state, proposition, information, sensation, location, time period.

According to these specifications, for *torta gelato* 'ice cream cake', *fucile da caccia* 'hunting rifle' and *impronta di piede* 'foot print', we will expect the following annotation:

(1) *torta*_1 *gelato*_2 'ice cream cake'
CONST[made _of]
1 whole / artifact
2 part / substance

(2) fucile_1 da caccia_2 'hunting rifle'
TELIC[used_for]
1 instr / artifact
2 event

(3) *impronta*_1 *di piede*_2 'foot print'
AG[caused_by]
1 result/artifact
2 cause/natural

If the value of the relation is implicit, as the TELIC in *treno merci* 'freight train' (=train which *transports* goods), we will ask the participants to specify the implicit predicate. So for example in (4) *merci* 'freight' is the theme of the telic value of *treno* 'train' [transport]:

(4) *treno*_1 *merci*_2 'freight train'
TELIC[transport]
1 instr/artifact
2 th/artifact

Finally, participants will also be required to annotate whether N2 is interpreted as modal or not (Bassac and Bouillon 2001), as in (5).

(5) *bicchiere*_1 *di vino*_2 'glass of wine'
FORM[hold]
1 th/artifact
2 th/substance/modal = no

*bicchiere*_1 *da vino*_2 'wine glass'
TELIC[ingest]
1 instr/artifact
2 th/substance/modal = yes

Summarising, the task involves tagging 1) the relation, 2) the implicit predicate in the relation, if there is one, 3) the role played by the referents of the nouns in the relation, 4) the broad semantic class associated with these referents, and 5) modality.

Drawing from the results of previous experiments (cf. Girju, 2009 :119), we will allow the annotator to annotate more than one relation for each compound. For example, in *pattini a rotelle* 'roller skates', the annotator might want to mark that both the CONST and the TELIC relation appear to be involved in the interpretation of the compound (the *rotelle* are a part of the artifact but also the means through which its function may be satisfied).

(6) *pattini*_1 *a rotelle*_2 'roller skates'
CONST[part_of]
1 whole/artifact
2 part/artifact

TELIC[move]
1 th/artifact
2 instr/artifact/modal = no

Finally, we will allow the annotators to say "I don't know" or encourage them to suggest a new relation when they think that none of the relations in the tag list applies to the compound under examination.

4. Evaluation

The annotation scheme presented in §3 was evaluated on a set of 80 Italian compounds and their translations in French. This dataset was extracted in the following way. The Italian compounds were first automatically aquired from the Europarl IT-EN corpus with a set of non-ambiguous heuristics, for example for the N da N:

[tag="NOM.*"][lemma="da"][tag="NOM.*"][tag="SENT"]
[tag="SENT"][tag="DET.*"][tag="NOM.*"][lemma="da"][tag="NOM.*"][tag="VER.*"]

In this way, we obtained a list of tokens for each category, i.e. N da N, N di N, N a N, and NN, sorted by frequency. For each list, we lemmatized the first 100 forms based on their head and manually extracted 5 IT-EN contexts for each of the 20 most frequent types from the corpus. In context extraction, we focused on instances where the compound is an argument of a predicate (*recarsi in altri Stati membri* 'travel to other Member States'), where it appears in modification constructions (*vecchi stati membri* 'old member states'), where it is part of a larger expression (*navi da carico e passeggeri* 'cargo and passenger ships') and where it is complement of a preposition (*di*/*con*, etc.).

The corresponding French translations were then manually extracted in context in the Europarl corresponding EN-FRE version. In that way, we obtained five translations for each of the Italian compounds, ie. 400 translations (20 x 4 x 5), as illustrated in Table 3.

IT	*Oggigiorno il **trasporto merci** a domicilio è possibile solo su strada.*
EN	*Door-to-door **goods transport** is only possible by lorry these days.*
FR	*Le **transport** porte à porte **de marchandises** n' est plus possible aujourd'hui que par camion.*

Table 3: Corpus examples

An initial version of the annotation scheme proposed in section 3 was tested by three experts who had to reach a consensus. The aim was to get a complete tested annotation scheme before applying it on a larger scale and with non-experts. In the following, we give a summary of the results of the annotation, focusing on relations and translations for N da N and N di N.

5. N da N

The table 4 below summarizes the results for the N da N. For this set, annotators reach agreement both for Italian and French. Table 4 partly confirms Johnston and Busa's analysis of Italian compounds (Johnston and Busa, 1999). In Italian, *da* often introduces a modal interpretation of its argument (N2): *animale da compagnia* 'pet animal' is an animal that can be used as a companion and *barbabietola da zucchero* 'sugar beet' is a vegetable from which we typically extract sugar. This modal interpretation is entailed by the telic relation, which is by far the most frequent (16/20). Data also confirm that in N da N, N1 and N2 can also be linked by an agentive relation (3/20). Among the different telic relations, **Telic[played_by]** is the less frequent: it refers to a special case where N1 refers to the qualia role itself (here the telic), as discussed by Busa et al. (2001). *Ruolo* 'role' *is what is played by somebody. Ruolo da protagonista* 'leading role' indicates that the TELIC of somebody is protagonist.

The analysis of the French examples shows that translation tends to preserve the qualia relation (all cases except one) and the syntax. In the majority of cases, N da N are translated by a NPN (80/100) and all French compounds of this category are at least translated once in that way. The most frequent alternative translations are N Adj (relational) (6/100) and simple nouns (11/100). Paraphrases are also possible as: *inquinamento da idrocarburi, "pollution occasionnée par le transport d'hydrocarbures"* or *animali da macello, "animaux destinés à l'élevage"*. The preposition can be "*à*" or "*de*", depending on the main relation between N1 and N2 and the semantic type/role of N2. In case of a telic relation, *da* is translated by the "*à*" if the relation is **Telic[Predicate]** (with a N2 of type **entity)** and by "*de*" if it is a **used_for** relation (with a N2 of type **event**), as already observed in

Bassac and Bouillon (2001); for the agentive relation, generalization is difficult at this stage but the preposition choice seems to depend either on the type of relation (**caused_by** or **derived_from**) and/or on the semantic type of N2 ("*de*" is used with an event - as in the telic relation - and "*par*" with an entity).

Relations (Italian)	**%**	**French translation**	**Examples**
Telic[Predicate]	7	N à N[entity]	*betterave à sucre* (It. *barbabietola da zucchero*)
Telic[used_for]	9	N de N[event]	*animal de compagnie* (It. *animale da compagnia*)
Telic[played_by]	1	N de N	*rôle de leader* (It. *ruolo da protagonista*)
Agentive[caused_by]	1	N par N[entity]	*pollution par hydrocarbure* (It. *inquinamento da idrocarburi*)
Agentive[derived_from]	2	N de N[event]	*revenu de l'épargne* (It. *reddito da risparmio*)

Table 4: N da N

6. N di N

For N di N, annotators reach less consensus at the beginning but at the end agree on the fact that *di* introduces either a non modal argument (relations of types **argumental**, **constitutive** and **agentive** in our set) (13/20) (e.g. *dichiarazioni di voto* 'explanations of votes' *turno di votazioni* 'vote', *dato di fatto* 'fact' respectively) or an expected result (tag of the Telic relation: **aims_at**) (e.g. *processo di pace* 'peace process') (7/20). In the latter case, *di* would then require that the event (specifically, a result state) will be achieved if the precondition is met (for an analysis along these lines, see also Johston and Busa 1999: 14 on similar constructions in Italian). As put forward by Johnston and Busa, the Telic relation can be lexicalized by *di* in Italian. (and *de* in French) when N2 is a (resultative) event, whilst (telic) *da* selects for both events and entities (see *animale da compagnia; barbabietola da zucchero*).

As regards the translation, the data show that when the N di N are translated by a NPN (88/100), the preposition is always "*de*". Alternative translations are most of the times simple nouns (*posto di lavoro, "emplois*; *dato di fatto, "preuves"*). In our set, not all the compounds have a translation in the form NPN as it was the case for the N da N.

Relations (Italian)	**%**	**French translation**	**Examples**
Telic[aims_at]	7	N de N	*processus de paix* (It. *processo di pace*)
Const [part_of]	1/2	N	*vote* (It. *turno di votazioni*)
Agentive[caused_by]	1	N	*preuves* (It. *dato di fatto*)
Argumental	10/11	N de N	*explications de vote* (It. *dichiarazioni di voto*)

Table 5: N di N

7. Conclusion

The aim of this paper was to propose an annotation scheme for compounds based on GL. This work was done in two steps: a first scheme was done *a priori*, that was then tested on a set of examples by experts. This evaluation on real data was very useful to finalize the list of tags and types. For the types, we decided to dismiss *physical object* and to use *natural* as opposed to *artifact*. Moreover, we introduced *substance* to capture the mass vs. countable distinction, and *entity* as a general category for all kinds of objects (abstract, natural, artifactual etc.). Finally, we distinguished between *event* (dynamic eventuality) and *state*. New tags were also introduced, that were missing in the first scheme, for example we added the tags **used_for** and **aims_at**, in order to distinguish cases where N2 refers to a resulting state and an activity.

Next steps will involve tagging a set of new examples and verifying the inter-annotator agreement on the set described here.

8. References

Baroni, M., Guevara, E and Pirrelli, V. (2009). Sulla Tipologia dei Composti NN in Italiano: principi categoriali ed evidenza distribuzionale a confronto. In R. Benatti, G. Ferrari & M. Mosca (Eds.), Linguistica e modelli tecnologici di ricerca (Atti del 40esimo Congresso della Società di Linguistica Italiana), Roma: Bulzoni, pp.73-95.

Bassac, C. and Bouillon, P. (2001). The Telic Relationship in Compounds in French and Turkish. In P. Bouillon and K. Kanzaki (eds), Proceedings of the First International Workshop on Generative Approaches to the Lexicon, April 26-28 2001, Geneva, Switzerland.

Bisetto, A. and Scalise, S. (2005). The classification of compounds. Lingue e linguaggio, 4(2), pp. 319-332.

Busa, F., Calzolari, N. and Lenci, A. (2001). Generative Lexicon and the SIMPLE Model: Developing Resources for NLP. In P. Bouillon and F. Busa (Eds), The language of Word Meaning, CUP.

Celli, F. and Nissim, M. (2009). Automatic Identification of semantic relation in Italian complex nominals. In Proceedings of the Eight International Conference on Computational Semantics (IWCS-8), Tilburg, The Netherlands, January 7-9 2009, pp.45-60.

Claveau, V. (2003). Acquisition automatique de lexiques sémantiques pour la recherche d'information, PhD Thesis, Université de Rennes 1.

Delfitto, D. and Melloni, C. (2009). Compounds don't come easy. Lingue e Linguaggio, VIII (1), pp.75-104.

Delfittol, D. and Melloni, C. (forthcoming) Compounding as a symmetry breaking strategy. In P. M. Bertinetto et al. (Eds), Linguaggio e cervello (Semantica, Atti del XLII Convegno della Società di Linguistica Italiana, Pisa, Italy, September 25-27 2008). Volume I. Roma: Bulzoni.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P. and Yuret, D. (2009). Classification of semantic relations between nominals. Language Resources and Evaluation, 43, pp.105–121.

Johnston, M. and Busa, F. (1999). Qualia Structure and the Compositional Interpretation of Compounds. In E. Viegas (Ed.), Breadth and Depth of Semantics Lexicons, Dordrecht: Kluwer, pp.167-187.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the tenth Machine Translation Summit, Asia-Pacific Association for Machine Translation, Phuket, Thailand, September 12-16 2005, pp.79-86.

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

Pustejovsky, J., Rumshisky, A., Moszkowicz, J.L. and Batiukova, O. (2008). GLML: A Generative Lexicon Markup Language. In Proceedings of the Generative Lexicon Workshop, Instituto di Linguistica Computazionale (CNR), Pisa, Italy, September 2008.

Seaghdha, D. (2007). Annotating and Learning Compound Noun Semantics. In Proceedings of the ACL 2007 Student Research Workshop, Prague, Austria, June 2007, pp.73-78.

Tratz S. and Hovy, E. (2010). A taxonomy, dataset, and classifier for Automatic Noun Compound Annotation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July 11-16 2010, pp. 678–687.