# Automatic identification of semantic relations in Italian complex nominals

Fabio Celli
CLIC-CIMeC
University of Trento
fabio.celli@email.unitn.it

Malvina Nissim
Dipartimento di Studi Linguistici e Orientali
University of Bologna
malvina.nissim@unibo.it

**Abstract**

This paper addresses the problem of the identification of the semantic relations in Italian complex nominals (CNs) of the type N+P+N. We exploit the fact that the semantic relation, which is underspecified in most cases, is partially made explicit by the preposition. We develop an annotation framework around five different semantic relations, which we use to create a corpus of 1700 Italian CNs, obtaining an inter-annotator agreement of K=.695. Exploiting this data, for each preposition $p$ we train a classifier to assign one of the five semantic relations to any CN of the type N+$p$+N, by using both string and supersense features. To obtain supersenses, we experiment with a sequential tagger as well as a plain lookup in MultiWordNet, and find that using information obtained from the former yields better results.

## 1 Introduction

Complex nominals are pervasive in language, and include noun-noun (N+N) and adjective-noun (A+N) combinations (Levi, 1978), as in Ex. 1 and 2.

(1) dessert fork

(2) medieval historian

A "dessert fork" is "a fork for eating dessert", and a "medieval historian" can be also described as "a historian who studies medieval times".[1] In both cases the relation is not overtly marked. Indeed, syntactically, there is nothing that tells us that the semantic relation between "dessert" and "fork" in Ex. 1 is different than the one binding "plastic" and "fork" in Ex. 3.

(3) plastic fork

However, it is well known that whereas English composes CNs of the type N+N, Romance languages must glue the two nouns by means of a preposition, thus yielding CNs of the form N+P+N, thereby partially making explicit the underlying semantic relation (Busa and Johnston, 1996). So, in Ex. 4, the "purpose" relation between dessert and fork is (partially) made explicit by the preposition "da". In contrast, the "property" relation binding plastic and fork (a fork made of plastic) is expressed using "di" (Ex. 5).

(4) forchetta *da* dessert (*en*: dessert fork)
(5) forchetta *di* plastica (*en*: plastic fork)

Recently, Girju (2007) has exploited this observation including cross-language information in a system for the automatic interpretation of NN compounds in English. However, whereas it is true that the overt preposition restricts the set of possible relations, it is also true that prepositions are still semantically ambiguous, since there is no one-to-one correspondence between prepositions and relations. So, "di", used in a "property" relation above, can also express a "part-whole" (Ex. 6), a "theme" (Ex. 7), and several other relations.

(6) dorso *della* mano (the back of the hand)
(7) suonatore *di* chitarra (guitar player)

In this work, we also exploit the presence of a preposition in Italian CNs as an aid to detect the semantic relation. We extract and annotate CNs in a corpus of written Italian, and develop a supervised system for determining the semantics of the CN, comparing the contribution of plain nouns with that of hypernym classes, and different ways in which such hypernyms can be obtained. In the next section, we discuss previous work on the semantics of complex nominals. In Section 3, we define a set of five semantic relations for the annotation of Italian CNs and the details of the annotation framework, and discuss the corpus distribution. In Section 4 we describe the experiments for the automatic identification of semantic relations, and discuss the results. We conclude with ideas for future work in Section 5.

[1] In this work we will only consider N+N CNs, thereby excluding A+N CNs.

## 2 Previous work

Given their underspecified nature, CNs, especially in English, have received a large amount of attention in the linguistic and computational linguistic literature (Downing, 1977; Levi, 1978; Warren, 1978; Lauer, 1995; Johnston and Busa, 1996; Rosario and Hearst, 2001; Lapata, 2002; Girju, 2007, among others). Current interest in NLP is also shown in the organisation of a SemEval task especially dedicated to noun-noun compound interpretation (Task 4, (Girju et al., 2007)). Indeed, NLP systems which aim at full text understanding for higher NLP tasks, such as question answering, recognising textual entailment and machine translation, need to grasp the semantic relation which noun compounds mostly leave underspecified.

One main issue in noun-noun compound interpretation is the lack of general agreement on a well-defined set of semantic relations. Nastase and Szpakowicz (2003), for instance, propose a two-level taxonomy, in which fifteen fine-grained relations are subsumed into five general classes (causal, participant, spatial, temporal, quality). An example of a causal relation (with subtype "purpose") is "concert hall", and an example of a participant relation (with subtype "beneficiary") is "student discount".

Girju et al. (2007) propose the smaller set reported in Table 1, which was tested on English N+N complex nominals within the SemEval 2007 task. They specifically spell out semantic relations as two-poles relationships: for example an effect is an effect always with respect to a cause.

Table 1: The set of 7 semantic relations from Girju et al. (2007)

| Semantic relation | Examples |
|---|---|
| Cause-Effect | laugh (cause) wrinkles (effect) |
| Instrument-Agency | laser (instrument) printer (agency) |
| Product-Producer | honey (product) bee (producer) |
| Origin-Entity | message (entity) from outer-space (origin) |
| Theme-Tool | news (theme) conference(tool) |
| Part-Whole | the door (part) of the car (whole) |
| Content-Container | apples (content) in the basket (container) |

As far as relation detection is concerned, Johnston and Busa (1996), working specifically on Italian, have suggested using information included in *qualia structures* (Pustejovsky, 1995) for deriving the compound's interpretation. The use of qualia structures for this task is appropriate and semantically sound but absolutely not straightforward to implement, since there does not exist an electronic repository of qualias, so that the structures

would need to be constructed by hand, thereby involving a large amount of manual work. Recent work has shown that the automatic acquisition of qualias can be performed with reasonable success exploiting information obtained using lexico-syntactic patterns over the Web (Cimiano and Wenderoth, 2005). For our purposes, though, if lexico-syntactic patterns can be used successfully to induce qualia roles, we could directly use the information we obtain from them, thus bypassing the qualia structure representation. We plan to include features based on such kinds of patterns in future development of this work (see also (Nakov and Hearst, 2008)).

More purely computational approaches include both supervised (Lauer, 1995) as well as unsupervised models, such as (Lapata and Keller, 2005), who use frequencies obtained over the Web. Some researchers also suggest solutions to the data sparsness problem, which affects our approach as well, by using lexical similarity (Turney, 2006) or clustering techniques (Pantel and Pennacchiotti, 2006).

Finally, there exists specific work on compound nouns whose head is derived from a verb (Lapata, 2002), and information about verbs deverbal nouns are linked to has proved a useful feature in previous approaches (Girju, 2007). Whereas we have exploited this information in the annotation phase, we have not included corresponding features yet in the statistical model we use, but we plan to do so in future extensions.

# 3 Annotation Framework and Data

For developing an annotation framework, we built on Italian grammars, existing classifications (see Section 2), and a preliminary study of corpus data.

## 3.1 Annotation framework

In determining the set of relations to be annotated, following (Girju, 2007), we also define two-pole relations between the involved nominals.

We assume that relations can be extracted and subsumed in general classes starting from $\theta$-roles, which are partially made explicit by the prepositional phrase. Since there is no general agreement on a complete list of $\theta$-roles we chose to work with types of complements, which are provided by traditional Italian grammars and can be found in almost every Italian dictionary. In (Zingarelli, 2008), we found 33 different types of prepositional phrases (PPs), which we grouped into 21 classes (for instance, all of the

location-related PPs were grouped under a single LOC class). This information was included in the annotation scheme (Celli, 2008), although is not used in the current relation identification model.

Following (Langacker, 1987), the nouns within each CNs were also revisited within a *trajector* (Tr) and *landmark* (Lm) approach mirroring the two-pole interpretation of the semantic relations.

The set of five semantic relations we arrived at is given in Table 2. These five relations are the target of our classification experiments (Section 4).

Table 2: Relations for Italian prepositions.

| Relation(Tag) | Description | Examples |
|---|---|---|
| cause-effect (CE) | *tr.* causes *lm.* | death$^{Lm}$ by privations$^{Tr}$ |
| located-location (LL) | *lm.* localizes *tr.* | window$^{Lm}$ passage$^{Tr}$ |
| owner-property (OP) | *tr.* possess *lm.* | stone$^{Lm}$ statue$^{Tr}$ |
| included-set (IS) | *lm.* includes *tr.* | thousands$^{Tr}$ of men$^{Lm}$ |
| bound-bounded (RR) | *lm.* undergoes *tr.* | city$^{Lm}$ destruction$^{Tr}$ |

In the cause-effect (CE) relation the trajector is the cause or the agent and the landmark is the product or the effect produced by the agent/causer, as in "morte per stenti" (*en*: death by privations). In located-location (LL), a trajector is located in space or time with respect to a landmark, as "casa in montagna" (*en*: mountain house). The owner-property (OP) relation associates a trajector (owner) with its property, part, or characteristic, which is the landmark. Examples are "statua di pietra" (*en*: stone statue) and "cane da caccia" (*en*: hunting dog). In included-set (IS) the trajector is the included object and the landmark is the set: in "migliaia di uomini" (*en*: thousands of men), "migliaia" (*en*: thousands) is the subset and "uomini" (*en*: men) is the set. The bound-bounded (RR) relation is a direct relationship between an event, usually a deverbal (trajector), and its undergoer (landmark), as "distruzione della città" (*en*: destruction of the city). Classic relations such as part-whole, producer-product, and is-a are covered in this account by the owner-property, cause-effect and included-set relations, respectively.

**Annotation categories** Each extracted CN (see Section 3.2) was annotated with the following information:

- the lemma (A, CON, DI, DA, IN, PER, SU, TRA)[2]

[2]The preposition "tra" can also be written as "fra". They are semantically equivalent. Occurrences of both variants were extracted, but we refer to them always as "tra".

- the relation (CE, OP, LL, IS, RR)
- the type of prepositional phrase (21 tags)
- the semantic type of `n1`/`n2` (natural, abstract, artifact, metaphorical usage)
- the position of *trajector* and *landmark* in the CN (TL, LT)
- the order of the head and the modifier in the CN (HM, MH)

The following CN types were to be excluded from annotation:

- CNs including proper nouns, such as "problema di Marco" (*en*: Mark's problem);
- CNs involving complex prepositions, such as "hotel nel mezzo del deserto" (*en*: hotel in the middle of the desert);
- CNs involving `n1` and/or `n2` of categories other than noun, due to POS-tagging errors;
- CNs containing bisyllabic prepositions, such as "macchina senza benzina" (*en*: car without fuel);[3]
- CNs used as adverbs, e.g. "accordo di massima" (*en*: generally agreed with)

### 3.2 Data

**Corpus Selection** We used CORISsmall, a reduced version of CORIS, a 100M-word, balanced corpus of written Italian (Rossini Favretti, 2000). CORISsmall was sampled by randomly extracting sentences with a length between 2 and 40 words. We discarded a few domain-specific subcorpora which were likely to contain prepositions used in ways different from common usage, as the legal subcorpus. The resulting corpus, henceforth CORISnominals, contains 75,000 words. The corpus was then automatically tagged with part-of-speech information, using TreeTagger (Schmid, 1994).

**CN detection** We chose to annotate monosyllabic prepositions only, namely *a* (to), *con* (with), *di* (of), *da* (from), *in* (in),*per* (for), *su* (on) and *tra* (within), because they are more frequent in CNs, more polysemous and not occurring as any other grammatical category, differently from bisyllabic prepositions which can be used adverbially. In any case, bisyllabic prepositions occurr in less than 2% of all the extracted CNs (42 out of 2298).

Exploiting part-of-speech information, we extracted all the N+P+N combinations with a context window of 10 words left and right. The frequency of the CNs found in CORISnominals is reported in Table 3.

[3]Prepositions which incorporate the determiner, such as "della" (di+la, *en*: of the) or "sulla" (su+la, *en*: on the), although possible bisyllabic, are definite variants of their corresponding monosyllabic prepositions, and are therefore included in the dataset.

Table 3: Frequency of CN types in CORISnominals

| CNs extracted | #inst | example |
|---|---|---|
| N+P+N | 1125 | lampada *a* olio (oil lamp) |
| N+Pdet+N | 1044 | dorso *della* mano (back of the hand) |
| N+P+D+N | 129 | casa *per le* vacanze (holiday home) |
| total | 2298 | |

**Annotation procedure and evaluation** The annotation was performed by a native speaker of Italian, with experience in the semantic analysis of complex nominals. After discarding some CNs according to the rules defined in the annotation scheme, the final number of annotated instances is 1700. In order to assess the difficulty of the relation assignment task, a randomly extracted portion of the data (186 CNs) was further annotated by a second native speaker of Italian. The second annotator marked them up following specific guidelines and some training material composed of about 50 already annotated CNs as examples. We calculated inter-annotator agreement using Cohen's kappa statistics (Cohen, 1960), obtaining a kappa of .695. While this relatively not so high value can be considered satisfactory in the field of semantic annotation (this score is also in the same ballpark as the 70.3% agreement reported for the SemeEval Task 4 annotation (Girju et al., 2007)), it still indicates that the phenomenon involves a good amount of ambiguity thus making the classification task far from straightforward. Table 4 reports the confusion matrix for the annotated subset.

Table 4: Confusion matrix for annotator A and annotator B

| A/B | CE | IS | LL | OP | RR | total |
|---|---|---|---|---|---|---|
| CE | 2 | – | – | – | – | 2 |
| IS | 1 | 22 | 4 | 5 | 1 | 33 |
| LL | – | – | 12 | 4 | 1 | 17 |
| OP | 34 | 4 | 8 | 44 | 6 | 96 |
| RR | 5 | – | 1 | 3 | 29 | 38 |
| total | 42 | 26 | 25 | 56 | 37 | 186 |

The largest area of disagreement is in the opposition between CE and OP: annotator B assigned the type CE to a large number of CNs which annotator A had marked as OP. This might be due to the fact that CE relations can be triggered by parts of objects (or features of concepts), which are expressed by the OP relation. A prime example of such overlap is "fumo

di sigaretta" (*en*: cigarette smoke), which can be seen both as a cause-effect relation as well as a owner-property relation. Thus, future work will involve a reassessment of these two categories and a revision of the guidelines.

**Corpus Distribution** Table 5 illustrates the distribution of semantic relations across each preposition.

Table 5: Distribution of relations across prepositions in CORISnominals

| prep/rel | CE | IS | LL | OP | RR | total |
|---|---|---|---|---|---|---|
| a | 0 | 8 | 29 | 34 | 28 | 99 |
| con | 0 | 5 | 0 | 10 | 14 | 29 |
| di | 62 | 262 | 69 | 646 | 289 | 1328 |
| da | 2 | 0 | 7 | 18 | 8 | 35 |
| in | 0 | 5 | 50 | 31 | 14 | 100 |
| per | 8 | 2 | 2 | 29 | 7 | 48 |
| su | 3 | 0 | 18 | 12 | 11 | 44 |
| tra | 0 | 0 | 4 | 3 | 10 | 17 |
| total | 75 | 282 | 179 | 783 | 381 | 1700 |

The most striking figure is the overwhelming predominance of "di", which features in 78% of all CNs. This is in line with the extremely high overall frequency of "di" in Italian, which is ranked as the most frequent word in CoLFIS (an Italian frequency lexicon based on a 3M word corpus, Laudanna et al. (1995)), and also with Girju's 2007 observation that 77.7% of the English noun-noun compounds in her data can be rephrased as "of" phrases. We can also observe that some prepositions, namely "a" and "con", show more than one predominant relation usage in CNs. Overall, OP is by far the most frequent relation, occurring in nearly half of the CNs.

As an additional observation, for each preposition we compared its frequency of occurrence in CNs and in any other constructions. We found that while "di" and "su" are particularly CN-oriented prepositions, both with over 55% of their occurrences being in CNs, the others appear in CNs about 10% or less of their total occurrences.

# 4 Automatic identification of CN relations

We can see the problem of semantic relations in CNs from at least two (converging) points of view. From a more language understanding side, given a CN (two nouns connected by a preposition), we might want to know what the

Table 6: Accuracy for most frequent relation baseline and for basic system

| prep | #inst | most freq rel | baseline | basic system |
|---|---|---|---|---|
| a | 99 | OP (34) | 34.34 | 47.47 |
| con | 29 | RR (14) | 48.28 | 48.28 |
| da | 35 | OP (18) | 51.43 | 51.43 |
| di | 1328 | OP (646) | 48.64 | 56.40 |
| in | 100 | LL (50) | 50.00 | 52.00 |
| per | 48 | OP (29) | 60.42 | 60.42 |
| tra | 17 | RR (10) | 58.82 | 58.82 |
| su | 44 | LL (18) | 40.91 | 50.00 |

underlying semantic relation is. From a more language generation perspective, though, we might want to be able to select the appropriate preposition, given two nouns and a relation between the concepts they express.

This translates into two different classification tasks. One where the target categories are relations, the other where they are prepositions. In the work we describe in this paper we concentrate on the first task. For each preposition we build a supervised model where the target categories correspond to the annotation tags for the semantic relations: CE, IS, LL, OP, RR. As evaluation measures, we report accuracy and coverage. Coverage amounts to the portion of data for which supersenses could be found for both n1 and n2, thus providing insights in assessing the contribution of different supersense assignment methods (see Section 4.2 and Section 4.3).

For assessing the difficulty of the task, beside inter-annotator agreement, we take a simple baseline where we assign to each CN the semantic relation which is most frequently associated with the CN's preposition (Table 6).

In the learning experiments, we use the Weka implementation (Witten and Frank, 2000) of the sequential minimal optimization algorithm for training a support vector classifier, within a ten-fold cross-validation setting. Girju (2007) has shown SVMs to be most efficient for this task.

### 4.1 Basic system

The basic system uses as features only n1 and n2 as simple strings. Table 6 shows accuracy per preposition for the basic system and for the baseline.

The most evident limitation of this basic approach is data sparseness. Out of 1700 CNs, 1662 involve a combination of n1 and n2 which occurs only once, independently of the preposition used. The most frequent n1 ("parte", part) occurs 13 times with two different prepositions, and the most frequent

n2 ("lavoro", job/work) 16 across four different prepositions.

One intuitive way to alleviate the data sparseness problem without increasing the corpus size, is to cluster instances. Following Girju (2007), who uses hypernyms obtained from WordNet (Fellbaum, 1998) in place of strings, we reduce each noun in our data set to its hypernym. In this supersense assignment, we experimented with two procedures: a more sophisticated one involving sequential sense tagging, thus dealing with sense disambiguation, and a simpler one involving plain assignment of hypernyms.

### 4.2 Hypernym selection via sense tagging

Two major problems related to finding a hypernym for a word are *sense ambiguity* (one term can easily have more than one hypernym if it has more than one sense) and *coverage* (even large ontologies/databases might not include some of the encountered terms). A supersense tagger alleviates such limitations by tagging words in context, thus tackling the ambiguity issue, and by using a combination of features rather than just the lexical entry, thereby being able to classify also words that are not included in the dictionary. Picca et al. (2008) have developed such a tagger for Italian, building on an existing version for English (Ciaramita and Altun, 2006), retrained on MultiSemCor (Bentivogli and Pianta, 2005), a word-aligned English-Italian corpus which contains the translation of the English texts in SemCor. The set of 26 noun supersense labels come from MultiWordNet (Pianta et al., 2002), a multilingual lexical database in which the Italian WordNet is strictly aligned with Princeton WordNet 1.6, and which is linked to MultiSemCor.

The average reported performance of the tagger is about 60% (Picca et al., 2008). This relatively low accuracy introduces a large portion of errors in the classification, thus reducing the advantage of dealing with supersenses rather than words in the identification of semantic relations in CNs. Errors can be of three types: (i) the assignment of a wrong noun class, (ii) the assignment of a class of the wrong part-of-speech type (any non-noun tag), and (iii) the non-assignment of any class (tag "0"). Whereas errors of type (i) can only be spotted via manual investigation, mistakes of type (ii) and (iii) can be detected automatically and a backoff strategy can be deployed. In 228 CNs out of 1700 both nouns have been assigned a "0" tag. In a further 751 CNs, one of the two nouns is tagged as "0". Out of these, there are 33 cases where the other noun is assigned a non-noun tag (adj or verb). A non-noun tag for n1 or n2 is also found in a further 57 cases.

As a backoff strategy for all cases that fall under (ii) and (iii), we searched

Table 7: Results using supersenses obtained via tagging, in combination with string features, and alone, and with and without backoff.

| | no backoff | | | | backoff | | | |
|---|---|---|---|---|---|---|---|---|
| prep | #inst | cov% | acc% | | #inst | cov% | acc% | |
| | | | string | nostring | | | string | nostring |
| a | 32 | 32.32 | 68.75 | 75.00 | 99 | 100 | 45.46 | 44.44 |
| con | 11 | 37.93 | 72.73 | 63.64 | 29 | 100 | 62.07 | 58.62 |
| da | 14 | 40.00 | 64.29 | 57.14 | 35 | 100 | 65.71 | 65.71 |
| di | 526 | 39.61 | 58.55 | 51.71 | 1328 | 100 | 59.71 | 50.75 |
| in | 36 | 36.00 | 63.89 | 61.11 | 100 | 100 | 64.00 | 62.00 |
| per | 16 | 33.33 | 68.75 | 56.25 | 48 | 100 | 56.25 | 54.17 |
| tra | 10 | 58.82 | 70.00 | 70.00 | 17 | 100 | 64.71 | 64.71 |
| su | 20 | 45.45 | 45.00 | 50.00 | 44 | 100 | 54.54 | 47.73 |

hypernyms directly in MultiWordNet (MWN). (The set of possible hypernyms is identical to the set of the 26 supersenses used by the tagger.) As a first step, we lemmatised the string using Morph-it!, an existing lemmatiser for Italian (Zanchetta and Baroni, 2005), since MWN contains lemmata but not their morphological variants. Whenever we found more than one synset associated to a term, a corresponding number of hypernyms was also found. If one of the hypernyms was recurring more than the others, this was selected. Otherwise, the hypernym associated to the first sense was selected.[4] Whenever the lemmatised noun was not in MWN (106 cases), we assigned the most frequent supersense in the dataset ("act" for both `n1` and `n2`).

We then ran classification experiments using the obtained supersenses for `n1` and `n2` as additional features, as well as on their own (thus ignoring the original string—this is reported as "nostring" in Tables 7–8), both with and without the backoff strategy. In the latter case, we excluded all CNs where at least one of the two nouns had been tagged as a non-noun or had no supersense assignment. Under these settings coverage was seriously affected, but accuracy was generally higher than when deploying the backoff strategy. Table 7 reports results.

[4]Optimally, we would select the hypernym for the most frequent sense (the one ranked first in Princeton WordNet). However, synsets for a given term are not ordered by frequency in MWN. One option would be to exploit frequencies from MultiSemCor, but the corpus is rather small and might not be very reliable.

Table 8: Results using supersenses obtained via plain assignment, in combination with string features, and alone, and with and without backoff.

| | no backoff | | | | backoff | | | |
|---|---|---|---|---|---|---|---|---|
| prep | #inst | cov% | acc% | | #inst | cov% | acc% | |
| | | | string | nostring | | | string | nostring |
| a | 90 | 90.91 | 47.78 | 42.22 | 99 | 100 | 39.39 | 34.34 |
| con | 26 | 89.65 | 61.54 | 57.69 | 29 | 100 | 55.17 | 55.17 |
| da | 30 | 85.71 | 60.00 | 63.33 | 35 | 100 | 62.86 | 65.71 |
| di | 1178 | 88.70 | 61.88 | 52.63 | 1328 | 100 | 60.54 | 51.13 |
| in | 88 | 88.00 | 50.00 | 52.27 | 100 | 100 | 56.00 | 53.00 |
| per | 41 | 85.42 | 65.85 | 60.98 | 48 | 100 | 56.25 | 47.92 |
| tra | 14 | 82.35 | 42.86 | 42.86 | 17 | 100 | 47.06 | 35.29 |
| su | 36 | 81.82 | 52.78 | 52.78 | 44 | 100 | 59.09 | 40.91 |

### 4.3 Hypernym selection via plain assignment

Given the large number of cases where we had to resort to a backoff strategy on the tagger's output, we tried to obtain hypernyms from MWN directly, thus bypassing the tagging stage. Whenever necessary, we employed the backoff strategies described above: most frequent hypernym found for an ambiguous term (or first sense's hypernym in case of equal frequency), and overall most frequent assigned hypernym in the corpus ("act" in this case as well) for all those nouns that were not found in MWN. This direct lookup approach should improve on coverage but suffer more from ambiguity-related problems. Table 8 summarises the results.

### 4.4 Discussion

Under the best settings, at full coverage, our average performance is around 59% (using tagger-assigned supersenses, backoff, the string feature), with wide variation across prepositions. Given the currently limited set of features, results are in general promising, especially if compared to the inter-annotator agreement, and to previous work (see below).

When using supersenses obtained from the tagger, results are steadily better than when using hypernyms directly looked up in MWN (both with and without backoff) with the exception of "di" and "su". The low coverage but higher accuracy yielded when using the tagger's senses without resorting to a backoff strategy were both expected, as mentioned above.

Results suggest that the utility of a backoff strategy varies from one

preposition to another. For instance, for “a”, ”con”, and ”per”, backoff appears to lower performance, independently on how the supersenses were obtained. These three prepositions had the three lowest coverage scores when using the tagger, which suggests that if too large a proportion is left to the approximation of backoff, the benefits of accurate sense tagging are lost. This is not however true for the MWN lookup, where the coverage for these three prepositions is rather high.

Additionally, we can observe that in most cases, in the back-off settings, including the string as a feature helps improve the performance (both in the tagging and in the plain assignment). This is likely due to the fact that the approximation given by not having precise information about the supersense and needing to resort to a backoff strategy is (partially) compensated by taking into account the original noun. In contrast, using the string without the backoff strategy on the tagger's output yields a decrease in performance, proving supersenses useful.

For a better assessment of the actual contribution of using hypernyms for detecting the semantic relation without incurring in the noise introduced by wrong hypernym assignments or the backoff strategy, we manually corrected the tagger's output in 60% of the data. This allowed us to evaluate the tagger's performance on supersense assignment for this 60% portion as well as to compare on this subset, contaning 1024 CNs, an algorithm using ”gold” supersenses with that built on the tagger's output (using the backoff strategy, and including string features, see Section 4.2). We found that supersenses were assigned by the tagger with an accuracy of 63.9%, a result in line with previously reported performance (Picca et al., 2008). We also observed that using the manually assigned hypernyms yielded an average improvement of about seven percentage points over using the tagger's senses, although for some prepositions, instances in this smaller dataset were just too few to draw any solid conclusion. Although more accurate, the gold tags do not boost the performance as much as one might expect. On the one hand, this might suggest that hypernyms can contribute only to a certain extent to this task, and other more expressive features must be found. On the other, it is also possible that the chosen set of 26 supersenses is too large, especially for a dataset like ours which is rather small, thereby not really overcoming the data sparseness problem.

Comparison to previous work in terms of performance is not straightforward, because of the language difference, the relation sets used, and the evaluation settings. In the SemEval-2007 exercise, for example, for each of the seven semantic relations used (see Table 1), a system must decide whether a given instance expresses that relation or not within an ad hoc-

built dataset, so that the overall semantic relation identification of the task is actually split in seven different binary classification tasks, one per relation. The highest reported average accuracy is 76.3% (Girju et al., 2007).

Girju (2007) classifies noun-noun compounds in 22 different semantic relations. Best results on English are obtained when using a rich feature set including cross-linguistic information. Reported figures differ slightly according to the dataset used, with an average accuracy of 76.1%. When using only language-internal supersense features, the average accuracy is 44.15%. Girju (2007) also trains and tests another state-of-the-art supervised model for English, namely Semantic Scattering (Moldovan and Badulescu, 2005), reporting an average accuracy of 59.07%.

## 5 Conclusions and future work

We have presented a framework for the annotation of Italian complex nominals in a very high data sparseness condition, and supervised models for the identification of the underlying semantic relation for monosyllabic Italian prepositions. We exploited both string and supersense features, showing that the importance of including string information varies from one preposition to another and from whether we are using backoff strategies or not. We have also seen that for obtaining the supersenses, a sequential sense tagging approach yields better overall results than a simple lookup in MWN, although it dramatically cuts on coverage.

Future work will involve further classification experiments with additional features, including web counts obtained via lexico-syntactic patterns (Lapata and Keller, 2005; Nakov and Hearst, 2008). We will exploit part of the annotation which we have not considered in this study (see Section 3), namely the type of prepositional phrase (see Appendix), a very general conceptual clustering which also marks metaphorical usage, the position of *trajector* and *landmark* in the CN, and the order of the head and the modifier.

## References

Bentivogli, L. and E. Pianta (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering 11*(3), 247–261.

Busa, F. and M. Johnston (1996). Cross-linguistic semantics for complex nominals in the generative lexicon. In *AISB Workshop on Multilinguality in the Lexicon*.

Celli, F. (2008). La semantica delle preposizioni italiane nella combinazione concettuale. Master thesis in Linguistics, Università di Bologna.

Ciaramita, M. and Y. Altun (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP 2006*, pp. 594–602.

Cimiano, P. and J. Wenderoth (2005). Automatically learning qualia structures from the web. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, Ann Arbor, Michigan, pp. 28–37.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*, 37–46.

Downing, P. (1977). On the creation and use of English compound nouns. *Language 53*, 810–842.

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

Girju, R. (2007). Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of ACL'07*, pp. 568–575.

Girju, R., P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret (2007, June). SemEval-2007 Task 04: Classification of Semantic Relations between Nominals. In *Proceedings of SemEval-2007*, pp. 13–18.

Johnston, M. and F. Busa (1996). Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL Workshop on breadth and depth of semantic lexicons*.

Langacker, R.W. (1987). *Foundations of cognitive grammar*. Univ. Press.

Lapata, M. (2002). The disambiguation of nominalisations. *Computational Linguistics 28*(3), 357–388.

Lapata, M. and F. Keller (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing* 2.

Laudanna, A., A. Thornton, G. Brown, C. Burani, and L. Marconi (1995). Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. In S. Bolasco, L. Lebart, and A. Salem (Eds.), *III Giornate internazionali di Analisi Statistica dei Dati Testuali. Volume I*, pp. 103–109. Cisu.

Lauer, M. (1995). Corpus statistics meet the noun compound: some empirical results. In *Proceedings of ACL'95*.

Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. Academic Press.

Moldovan, D. and A. Badulescu (2005). A semantic scattering model for the automatic interpretation of genitives. In *Proceedings of HLT-EMNLP 2005*, pp. 891–898.

Nakov, P. and M. A. Hearst (2008). Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp. 452–460. Association for Computational Linguistics.

Nastase, V. and S. Szpakowicz (2003). Exploring noun-modifier semantic relations. In *Proceedings of IWCS-5*, pp. 285–301.

Pantel, P. and M. Pennacchiotti (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of ACL'06*, Sydney, Australia, pp. 113–120.

Pianta, E., L. Bentivogli, and C. Girardi (2002). MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pp. 293–302.

Picca, D., A. M. Gliozzo, and M. Ciaramita (2008). Supersense Tagger for Italian. In *Proceedings of LREC 2008*.

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

Rosario, B. and M. Hearst (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In L. Lee and D. Harman (Eds.), *Proceedings of EMNLP 2001*, pp. 82–90.

Rossini Favretti, R. (2000). Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In R. Rossini Favretti (Ed.), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Bulzoni.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of the Conference on New Methods in Language Processing*, 44-49.

Turney, P. D. (2006). Expressing implicit semantic relations without supervision. In *Proceedings of ACL'06*, Sydney, Australia, pp. 313–320.

Warren, B. (1978). Semantic patterns of noun-noun compounds. *Gothenburg Studies in English 41*.

Witten, I. H. and E. Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Zanchetta, E. and M. Baroni (2005). Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005 1*(1).

Zingarelli, N. (2008). *Lo Zingarelli 2008. Vocabolario della Lingua Italiana*. Zanichelli.