The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study

Roxana Girju*
University of Illinois at
Urbana-Champaign

In this article we explore the syntactic and semantic properties of prepositions in the context of the semantic interpretation of nominal phrases and compounds. We investigate the problem based on cross-linguistic evidence from a set of six languages: English, Spanish, Italian, French, Portuguese, and Romanian. The focus on English and Romance languages is well motivated. Most of the time, English nominal phrases and compounds translate into constructions of the form N P N in Romance languages, where the P *(preposition) may vary in ways that correlate with the semantics. Thus, we present empirical observations on the distribution of nominal phrases and compounds and the distribution of their meanings on two different corpora, based on two state-of-the-art classification tag sets: Lauer's set of eight prepositions and our list of 22 semantic relations. A mapping between the two tag sets is also provided. Furthermore, given a training set of English nominal phrases and compounds along with their translations in the five Romance languages, our algorithm automatically learns classification rules and applies them to unseen test instances for semantic interpretation. Experimental results are compared against two state-of-the-art models reported in the literature.*

1. Introduction

Prepositions are an important and frequently used category in both English and Romance languages. In a corpus study of one million English words, Fang (2000) shows that one in ten words is a preposition. Moreover, about 10% of the 175 most frequent words in a corpus of 20 million Spanish words were found to be prepositions (Almela et al. 2005). Studies on language acquisition (Romaine 1995; Celce-Murcia and Larsen-Freeman 1999) have shown that the acquisition and understanding of prepositions in languages such as English and Romance is a difficult task for native speakers, and even more difficult for second language learners. For example, together with articles, prepositions represent the primary source of grammatical errors for learners of English as a foreign language (Gocsik 2004).

* Linguistics and Computer Science Departments, University of Illinois at Urbana-Champaign, Urbana, IL 61801. E-mail: girju@illinois.edu.

Submission received: 1 August 2006; revised submission received: 20 January 2008; accepted for publication: 17 March 2008.

Although the complexity of preposition usage has been argued for and documented by various scholars in linguistics, psycholinguistics, and computational linguistics, very few studies have been done on the function of prepositions in natural language processing (NLP) applications. The reason is that prepositions are probably the most polysemous category and thus, their linguistic realizations are difficult to predict and their cross-linguistic regularities difficult to identify (Saint-Dizier 2005a).

In this article we investigate the role of prepositions in the task of automatic semantic interpretation of English nominal phrases and compounds. The problem is simple to define: Given a compositional noun phrase (the meaning of the phrase derives from the meaning of the constituents) constructed out of a pair of nouns, N_1 N_2, one representing the head and the other the modifier, determine the semantic relationship between the two nouns. For example, the noun–noun compound *family estate* encodes a POSSESSION relation, while the nominal phrase *the faces of the children* refers to PART-WHOLE. The problem, although simple to state, is difficult for automatic semantic interpretation. The reason is that the meaning of these constructions is most of the time implicit (it cannot be easily recovered from morphological analysis). Interpreting nominal phrases and compounds correctly requires various types of information, from world knowledge to lexico-syntactic and discourse information.

This article focuses on nominal phrases of the type N P N and noun compounds (N N) and investigates the problem based on cross-linguistic evidence from a set of six languages: English, Spanish, Italian, French, Portuguese, and Romanian. The choice of these constructions is empirically motivated. In a study of 6,200 (Europarl[1]) and 2,100 (CLUVI[2]) English token nominal phrase and compound instances randomly chosen from two English–Romance parallel text collections of different genres, we show that over 80% of their Romance noun phrase translations are encoded by N P N and N N constructions. For instance, *beer glass*, an English compound of the form N_1 N_2, translates into N_2 P N_1 instances in Romance: *tarro de cerveza* ('glass of beer') in Spanish, *bicchiere da birra* ('glass for beer') in Italian, *verre à bière* ('glass at/to beer') in French, *copo de cerveja* ('glass of beer') in Portuguese, and *pahar de bere* ('glass of beer') in Romanian. In this article, in addition to the sense translation (in italics), when relevant we also provide the word-by-word gloss (in 'parentheses'). Moreover, we use N_1, N_2 to denote the two lexical nouns that encode a semantic relation (where N_1 is the syntactic modifier and N_2 is the syntactic head), and Arg_1, Arg_2 to denote the semantic arguments of the relation encoded by the two nouns. For example, *beer glass* encodes a PURPOSE relation where Arg_1 (*beer*) is the purpose of Arg_2 ('glass'; thus 'glass (used) for beer').

We argue here that the syntactic directionality given by the head-modifier relation (N_1 N_2 in noun compounds and N_2 P N_1 in nominal phrases) is not always the same as the semantic directionality given by the semantic argument frame of the semantic relation. Otherwise said, N_1 does not always map to Arg_1 and N_2 to Arg_2 for any given relation.

Languages choose different nominal phrases and compounds to encode relationships between nouns. For example, English nominal phrases and compounds of the

1 http://www.isi.edu/koehn/europarl/.
This corpus contains over 20 million words in eleven official languages of the European Union covering the proceedings of the European Parliament from 1996 to 2001.

2 CLUVI - Linguistic Corpus of the University of Vigo Parallel Corpus 2.1; http://sli.uvigo.es/CLUVI/.
CLUVI is an open text repository of parallel corpora of contemporary oral and written texts in some of the Romance languages (such as Galician, French, Spanish, and Portuguese) and Basque parallel text collections.

form N_1 N_2 (e.g., *wood stove*) and N_2 P_1 N_1 (e.g., *book on the table*) usually translate in Romance languages as N_2 P_2 N_1 (e.g., *four à bois* in French – 'stove at/to wood', and *livre sur la table* – 'book on the table'). Romance languages have very few N N compounds and they are of limited semantic categories, such as TYPE (e.g., *legge quadro* in Italian – 'law framework' – translates as *framework law*). Besides the unproductive N N and the productive N P N phrases, Romanian also uses another productive construction: the genitive-marked noun–noun compounds (e.g., *frumuseţea fetei* – beauty-the girl-GEN – translated as *the beauty of the girl*). Whereas English N N compounds are right-headed (e.g., *framework*/Modifier *law*/Head), Romance compounds are left-headed (e.g., *legge*/Head *quadro*/Modifier). Moreover, the Romance preposition used in the translations of English nominal phrase instances of the type N P N is one that comes closest to having overlapping semantic range as intended in the English instance, but may not be the exact counterpart for the whole semantic range. For example, *Committee* **on** *Culture* translates as *Comisión* **de** *la Cultura* (Spanish) ('Committee of the Culture'), *Commission* **de** *la Culture* (French) ('Committee of the Culture'), *Commissione* **per** *la Cultura* (Italian) ('Committee for the Culture'), *Comissão* **para** *Cultura* (Portuguese) ('Committee for Culture'), and *Comitet* **pentru** *Cultură* (Romanian) ('Committee for Culture'). Even those Romance prepositions that are spelled "de" are pronounced differently in different Romance languages.

Thus, the focus on nominal phrases and compounds in English and Romance languages is also motivated linguistically. The extension of this task to natural languages other than English brings forth both new insights and new challenges. The Romance prepositions used in the translations of English nominal phrases and compounds, may vary in ways that correlate with the semantics. Thus, Romance language prepositions will give us another source of evidence for disambiguating the semantic relations in English nominal phrases and compounds. We argue that, in languages with multiple syntactic options such as English (N N and N P N) and Romanian (N N, genitive-marked N N, and N P N), the choice between such constructions in context is governed in part by semantic factors. For example, the set of semantic relations that can be encoded by pairs of nouns such as *tea–cup* and *sailor–suit* varies with the syntactic construction used. In English, while the noun–noun compounds *tea cup* and *sailor suit* encode only PURPOSE, the N P N constructions *cup of tea* and *suit of the sailor* encode CONTENT-CONTAINER (a subtype of LOCATION) and MEASURE relations and POSSESSION, respectively. Similarly, in Romanian both *tea cup* and *cup of tea* translate only as the N P N instance *ceaşcă de ceai* ('cup of tea'), while *sailor suit* translates as *costum de marinar* ('suit of sailor') and the *suit of the sailor* as the genitive-marked N N *costumul marinarului* ('suit-the sailor-GEN'). Thus, we study the distribution of semantic relations across different nominal phrases and compounds in one language and across all six languages, and analyze the resulting similarities and differences. This distribution is evaluated over the two different corpora based on two state-of-the-art classification tag sets: Lauer's set of eight prepositions (Lauer 1995) and our list of 22 semantic relations. A mapping between the two tag sets is also provided.

In order to test their contribution to the task of semantic interpretation, prepositions and other linguistic clues are employed as features in a supervised, knowledge-intensive model. Furthermore, given a training set of English nominal phrases and compounds along with their translations in the five Romance languages, our algorithm automatically learns classification rules and applies them to unseen test instances for semantic interpretation. As training and test data we used 3,124 Europarl and 2,023 CLUVI token instances. These instances were annotated with semantic relations and analyzed for inter-annotator agreement. The results are compared against two

state-of-the-art approaches: a supervised machine learning model, semantic scattering (Moldovan and Badulescu 2005), and a Web-based unsupervised model (Lapata and Keller 2005). Moreover, we show that the Romanian linguistic features contribute more substantially to the overall performance than the features obtained for the other Romance languages. This is explained by the fact that the choice of the linguistic constructions (either genitive-marked N N or N P N) in Romanian is highly correlated with their meaning.

The article is organized as follows. Section 2 presents a summary of related work. In Section 3 we describe the general approach to the interpretation of nominal phrases and compounds and list the syntactic and semantic interpretation categories used along with observations regarding their distribution in the two different cross-linguistic corpora. Sections 4 and 5 present a learning model and experimental results. Section 6 presents linguistic observations on the behavior of English and Romanian N N and N P N constructions. Finally, in Section 7 we provide an error analysis and in Section 8 we offer some discussion and conclusions.

2. Previous Work

2.1 Noun Phrase Semantic Interpretation

The semantic interpretation of nominal phrases and compounds in particular and noun phrases (NPs) in general has been a long-term research topic in linguistics, computational linguistics,[3] and artificial intelligence.

Noun–noun compounds in linguistics
Early studies in linguistics (Lees 1963) classified noun–noun compounds on purely grammatical criteria using a transformational approach, criteria which failed to account for the large variety of constraints needed to interpret these constructions. Later on, Levi (1978) attempted to give a tight account of noun–noun interpretation, distinguishing two types of noun–noun compounds: (a) compounds interpreted as involving one of nine predicates (CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM, ABOUT) (e.g., *onion tears* encodes CAUSE) and (b) those involving nominalizations, namely, compounds whose heads are nouns derived from a verb, and whose modifiers are interpreted as arguments of the related verb (e.g., a *music lover* loves music). Levi's theory was cast in terms of the more general theory of Generative Semantics. In that theory it was assumed that the interpretation of compounds was available because the examples were derived from underlying relative clauses that had the same meanings. Thus, *honey bee*, expressing the relation MAKE, was taken to be derived from a headed relative *a bee that makes honey*. Levi was committed to the view that a very limited set of predicates constituted all of the relations that could hold between nouns in simple noun–noun compounds. This reductionist approach has been criticized in studies of language use by psycholinguists (Gleitman and Gleitman 1970; Downing 1977) who claim that noun–noun compounds, which are frequent in languages like English, encode in principle an

3 In the past few years at many workshops, tutorials, and competitions this research topic has received considerable interest from the computational linguistics community: the Workshops on Multiword Expressions at ACL 2003, ACL 2004 and COLING/ACL 2006; the Computational Lexical Semantics Workshop at ACL 2004; the Tutorial on Knowledge Discovery from Text at ACL 2003; the Shared Task on Semantic Role Labeling at CONLL 2004 and 2005 and at SemEval 2007.

unbounded number of possible relations. One such example is *apple juice seat*—"a seat in front of which an apple juice [is] placed" (Downing 1977, page 818)—which can only be interpreted in the current discourse context.

In this article we tackle the problem using a unified framework. Although we agree with Downing (1977) that pragmatics plays an important factor in noun–noun interpretation, a large variety of noun–noun meanings can be captured with a well-chosen set of semantic relations. Our proposed semantic classification set differs from that of Levi (1978) in the sense that it contains more homogenous categories. Levi's categories, instead, are more heterogeneous, including both prepositions and verbs, some of which are too general (e.g., the prepositions *for, in* and the verb *to have*), and thus, too ambiguous. Moreover, in our approach to automatic semantic interpretation we focus on both N N and N P N constructions and exploit a set of five Romance languages.

Noun–noun compounds in computational linguistics

The automatic interpretation of nominal phrases and compounds is a difficult task for both unsupervised and supervised approaches. Currently, the best-performing noun–noun interpretation methods in computational linguistics focus mostly on two or three-word noun–noun compounds and rely either on ad hoc, domain-specific, hand-coded semantic taxonomies, or statistical models on large collections of unlabeled data. Recent results have shown that symbolic noun–noun compound interpretation systems using machine learning techniques coupled with a large lexical hierarchy perform with very good accuracy, but they are most of the time tailored to a specific domain (Rosario and Hearst 2001; Rosario, Hearst, and Fillmore 2002), or are general purpose (Turney 2006) but rely on semantic similarity metrics on WordNet (Fellbaum 1998). On the other hand, the majority of corpus statistics approaches to noun–noun compound interpretation collect statistics on the occurrence frequency of the noun constituents and use them in a probabilistic model (Lauer 1995). The problem is that most noun–noun compounds are rare and thus, statistics on such infrequent instances lead in general to unreliable estimates of probabilities. More recently, Lapata and Keller (2005) showed that simple unsupervised models applied to the noun–noun compound interpretation task perform significantly better when the n-gram frequencies are obtained from the Web (55.71% accuracy[4]), rather than from a large standard corpus. Nakov and Hearst (2005) improve over Lapata and Keller's method through the use of surface features and paraphrases only for the task of noun–noun compound bracketing (syntactic parsing of three-word noun compounds) without their interpretation. Other researchers (Pantel and Ravichandran 2004; Pantel and Pennacchiotti 2006; Pennacchiotti and Pantel 2006) use clustering techniques coupled with syntactic dependency features to identify IS-A relations in large text collections. Kim and Baldwin (2005) propose a general-purpose method that computes the lexical similarity of unseen noun–noun compounds with those found in training. More recently Kim and Baldwin (2006) developed an automatic method for interpreting noun–noun compounds based on a set of 20 semantic relations. The relations are detected based on a fixed set of constructions involving the constituent nouns and a set of seed verbs denoting the semantic relation (e.g., *to own* denotes POSSESSION). Then all noun–noun instances

4 These results were obtained on AltaVista on a general and abstract set of eight prepositions (Lauer 1995) as semantic classification categories: *of, for, with, in, on, at, about*, and *from*.

in transitive sentential contexts (i.e., those sentences containing a transitive verb) are mapped onto the selected set of constructions based on lexical similarity over the verbs.

However, although the Web-based solution might overcome the data sparsity problem, current probabilistic models are limited because they do not take full advantage of the structure and the meaning of language.

From a cross-linguistic perspective, there hasn't been much work on the automatic interpretation of nominal phrases and compounds. Busa and Johnston (1996), Johnston and Busa (1996), and Calzolari et al. (2002), for example, focus on the differences between English and Italian noun–noun compounds. In their work they argue that a computational approach to the cross-linguistic interpretation of these compounds has to rely on a rich lexical representation model, such as those provided by FrameNet frames (Baker, Fillmore, and Lowe 1998) and qualia structure (Pustejovsky 1995). In the qualia structure representation, for example, the meaning of a lexical concept, such as the modifier in a noun–noun compound, is defined in terms of four elements representing concept attributes along with their use and purpose. Thus, qualia structure provides a relational structure that enables the compositional interpretation of the modifier in relation to the head noun. Two implementations of such representations are provided by the SIMPLE Project ontology (Lenci et al. 2000) and the OMB ontology (Pustejovsky et al. 2006). The SIMPLE ontology, for example, is developed for 12 European languages and defines entry words that are mapped onto high-level concepts in EuroWordNet (Vossen 1998), a version of WordNet developed for European languages.

In this article, we use a supervised semantic interpretation model employing rich linguistic features generated from corpus evidence coupled with word sense disambiguation and WordNet concept structure information. The results obtained are compared against two state-of-the-art approaches: a supervised machine learning model, semantic scattering (Moldovan and Badulescu 2005), and a Web-based unsupervised model (Lapata and Keller 2005). In this research we do not consider extra cross-linguistic information, such as semantic classes of Romance nouns (those provided by IS-A relations; e.g., *cat* belongs to the class of *animals*) made available, for example, by the SIMPLE ontology. However, such resources can be added at any time to further improve the performance of noun–noun interpretation systems.

2.2 Semantics of Prepositions

Although prepositions have been studied intensively in linguistics (Herskovits 1987; Zelinski-Wibbelt 1993; Linstromberg 1997; Tyler and Evans 2003; Evans and Chilton 2009, among others), they have only recently started to receive more attention in the computational linguistics community.[5] Moreover, the findings from these broad studies have not yet been fully integrated into NLP applications. For example, although information retrieval, and even question answering systems, would benefit from the incorporation of prepositions into their NLP techniques, they often discard them as stop words.

5 The first Workshop on the Syntax and Semantics of Prepositions, Toulouse, France, 2003; the second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Colchester, UK, 2005; the third ACL-SIGSEM Workshop on Prepositions, Trento, Italy, 2006.

Prepositions in linguistics

Considerable effort has been allocated to the investigation of spatial prepositions mainly based on a cognitive approach, not only in English (Herskovits 1987; Linstromberg 1997; Tyler and Evans 2003; Evans and Chilton 2009), but also in many of the Indo-European languages (Casadei 1991; Vandeloise 1993; Cadiot 1997; Melis 2002; Luraghi 2003). These studies provide a detailed analysis of such prepositions trying to give a methodological motivated account for the range of their polysemy. These works identify special constraints on various prepositional patterns, such as semantic restrictions on the noun phrases occurring as complements of the preposition. For example, in prepositional phrase constructions such as *in NP*, the head noun can be a container (*in a cup*), a geometrical area (*in a region*), a geo-political area (*in Paris*), an atmospheric condition (*in the rain*), and so on. These selectional restrictions imposed by the preposition on the noun phrases it combines with are presented in various formats from lists (Herskovits 1987; Linstromberg 1997) to semantic networks of cluster senses (Tyler and Evans 2003). In this article we also focus on the polysemy of such prepositions, but we identify the selectional restrictions automatically based on a specialization procedure on the WordNet IS-A hierarchy. However, unlike Herskovits, we do not consider pragmatic issues such as relevance and tolerance. These account for the difference that pragmatic motivations and context dependency make to how expressions are understood. Relevance has to do with communicative goals and choice of means and is evident, for example, in instances such as *cat on the mat* which is still relevant even when only the paws and not the whole cat are on the mat. Tolerance occurs in situations in which a book, for example, is described as *on the table* even though a set of files are placed between it and the table.

The use of spatial prepositions can also trigger various inferences. For example, *the man at his desk* (cf. Herskovits 1987) implies, besides a LOCATION relation, that the man is using the desk, thus an INSTRUMENT relation. Other inferences are more subtle, involving spatial reasoning about the actions that can be performed on the arguments of the preposition. One such instance is *infant in a playpen* (cf. Tyler and Evans 2003), where the movement of the playpen involves the movement of the infant. In order to identify such inferences the automatic interpretation system has to rely on pragmatic knowledge. In this research we do not deal with such inference issues, rather we identify the meaning of N P N constructions based on the local context of the sentence.

Prepositions in computational linguistics

In order to incorporate prepositions into various resources and applications, it is necessary to perform first a systematic investigation of their syntax and semantics. Various researchers (Dorr 1993; Litkowski and Hargraves 2005; Saint-Dizier 2005b; Lersundi and Aggire 2006) have already provided inventories of preposition senses in English and other languages. Others have focused on the analysis of verb particles (Baldwin 2006a, 2006b; Villavicencio 2006), the distributional similarity (Baldwin 2005) and the semantics of prepositions (Kordoni 2005) in a multilingual context, and the meaning of prepositions in applications such as prepositional phrase attachment (O'Hara and Wiebe 2003; Kordoni 2006; Volk 2006).

Moreover, although there is a large amount of work in linguistics and computational linguistics relating to contrastive analysis of prepositions (Busa and Johnston (1996); Johnston and Busa (1996); Jensen and Nilsson (2005); Kordoni (2005), inter alia), to our knowledge, there have not been any attempts to provide an investigation of the prepositions' role in the task of automatic noun phrase interpretation in a large cross-linguistic English–Romance framework.

3. Linguistic Considerations of Nominal Phrases and Compounds

The meaning of nominal phrases and compounds can be **compositional** (e.g., *spoon handle*—PART–WHOLE, *kiss in the morning*—TEMPORAL), or **idiosyncratic**, when the meaning is a matter of convention (e.g., *soap opera*, *sea lion*). These constructions can also encode metaphorical names (e.g., *ladyfinger*), proper names (e.g., *John Doe*), and dvandva compounds[6] in which neither noun is the head (e.g., *player–coach*).

Moreover, they can also be classified into **synthetic** (verbal, e.g., *truck driver*) and **root** (non-verbal, e.g., *tea cup*) constructions.[7] It is widely held (Levi 1978; Selkirk 1982b) that the modified noun of a synthetic noun–noun compound, for example, may be associated with a theta-role of the compound's head noun, which is derived from a verb. For instance, in *truck driver*, the noun *truck* satisfies the THEME relation associated with the direct object in the corresponding argument structure of the verb *to drive*.

In this article we address English–Romance compositional nominal phrases and compounds of the type N N (noun–noun compounds which can be either genitive-marked or not genitive-marked) and N P N, and disregard metaphorical names, proper names, and dvandva structures. In the following we present two state-of-the-art semantic classification sets used in automatic noun–noun interpretation and analyze their distribution in two different corpora.

3.1 Lists of Semantic Classification Relations

Although researchers (Jespersen 1954; Downing 1977) argued that noun–noun compounds, and noun phrases in general, encode an infinite set of semantic relations, many agree (Levi 1978; Finin 1980) there is a limited number of relations that occur with high frequency in these constructions. However, the number and the level of abstraction of these frequently used semantic categories are not agreed upon. They can vary from a few prepositions (Lauer 1995) to hundreds and even thousands of more specific semantic relations (Finin 1980). The more abstract the category, the more noun phrases are covered, but also the larger the variation as to which category a phrase should be assigned. Lauer, for example, classifies the relation between the head and the modifier nouns in a noun–noun compound by making use of a set of eight frequently used prepositions: *of, for, with, in, on, at, about*, and *from*. However, according to this classification, the noun–noun compound *love story*, for instance, can be classified both as *story* **of** *love* and *story* **about** *love*. The main problem with these abstract categories is that much of the meaning of individual compounds is lost, and sometimes there is no way to decide whether a form is derived from one category or another. On the other hand, lists of very specific semantic relations are difficult to build as they usually contain a very large number of predicates, such as the list of all possible verbs that can link the noun constituents. Finin, for example, uses semantic categories such as **dissolved in** to build interpretations of compounds such as *salt water* and *sugar water*.

In this article we experiment with two sets of semantic classification categories defined at different levels of abstraction. The first is a core set of 22 semantic relations (SRs), a set which was identified by us from the linguistics literature and from various experiments after many iterations over a period of time (Moldovan and Girju 2003).

6 The term *dvandva* comes from Sanskrit, translates literally as 'two-and-two' and means 'pair'.

7 In the linguistic literature the words "synthetic" and "root" have been coined for noun–noun compounds. Because these terms apply also to nominal phrases, we use them in relation to these constructions as well.

Moldovan and Girju proved empirically that this set is encoded by noun–noun pairs in noun phrases; the set is a subset of their larger list of 35 semantic relations used in a large set of semantics tasks. This list, presented in Table 1 along with examples and semantic argument frames, is general enough to cover a large majority of text semantics while keeping the semantic relations to a manageable number. A semantic argument frame is defined for each semantic relation and indicates the position of each semantic argument in the underlying relation. For example, "Arg_2 is part of (whole) Arg_1" identifies the part (Arg_2) and the whole (Arg_1) entities in this relation. This representation is important because it allows us to distinguish between different arrangements of the arguments for given relation instances. For example, most of the time, in N N compounds Arg_1 precedes Arg_2, whereas in N P N constructions the position is reversed (Arg_2 P Arg_1). However, this is not always the case as shown by N N instances such as *ham*/Arg_2 *sandwich*/Arg_1 and *spoon*/Arg_1 *handle*/Arg_2, both encoding PART–WHOLE. More details on subtypes of PART–WHOLE relations are presented in Section 6.2. A special relation here is KINSHIP, which is encoded only by N P N constructions and whose argument order is irrelevant. Thus, the labeling of the semantic arguments for each relation as Arg_1 and Arg_2 is just a matter of convention and they were introduced to provide a consistent guide to the annotators to easily test the goodness-of-fit of the relations. The examples in column 4 are presented with their WordNet senses identified in context from the CLUVI and Europarl text collections, where the specific sense is represented as the sense number preceded by a "#" sign.

The second set is Lauer's list of eight prepositions (exemplified in Table 2) and can be applied only to noun–noun compounds, because in N P N instances the preposition is explicit. We selected these two state-of-the-art sets as they are of different size and contain semantic classification categories at different levels of abstraction. Lauer's list is more abstract and thus capable of encoding a large number of noun–noun compound instances found in a corpus (e.g., many N_1 N_2 instances can be paraphrased as N_2 *of*

Table 1
The set of 22 semantic relations along with examples interpreted in context and the semantic argument frame.

No.	Semantic relations	Default argument frame	Examples
1	POSSESSION	Arg_1 POSSESSES Arg_2	*family#2*/Arg_1 *estate#2*/Arg_2
2	KINSHIP	Arg_1 IS IN KINSHIP REL. WITH Arg_2	*the sister#1*/Arg_2 *of the boy#1*/Arg_1
3	PROPERTY	Arg_2 IS PROPERTY OF Arg_1	*lubricant#1*/Arg_1 *viscosity#1*/Arg_2
4	AGENT	Arg_1 IS AGENT OF Arg_2	*investigation#2*/Arg_2 *of the police#1*/Arg_1
5	TEMPORAL	Arg_1 IS TEMPORAL LOCATION OF Arg_2	*morning#1*/Arg_1 *news#3*/Arg_2
6	DEPICTION-DEPICTED	Arg_2 DEPICTS Arg_1	*a picture#1*Arg_2 *of my nice#1*/Arg_1
7	PART-WHOLE	Arg_2 IS PART OF (whole) Arg_1	*faces#1*/Arg_2 *of children#1*/Arg_1
8	HYPERNYMY (IS-A)	Arg_1 IS A Arg_2	*daisy#1*/Arg_1 *flower#1*/Arg_2
9	CAUSE	Arg_1 CAUSES Arg_2	*scream#1*/Arg_2 *of pain#1*/Arg_1
10	MAKE/PRODUCE	Arg_1 PRODUCES Arg_2	*chocolate#2*/Arg_2 *factory#1*/Arg_1
11	INSTRUMENT	Arg_1 IS INSTRUMENT OF Arg_2	*laser#1*/Arg_1 *treatment#1*/Arg_2
12	LOCATION	Arg_2 IS LOCATED IN Arg_1	*castle#1*/Arg_2 *in the desert#1*/Arg_1
13	PURPOSE	Arg_1 IS PURPOSE OF Arg_2	*cough#1*/Arg_1 *syrup#1*/Arg_2
14	SOURCE	Arg_1 IS SOURCE OF Arg_2	*grapefruit#2*/Arg_1 *oil#3*/Arg_2
15	TOPIC	Arg_1 IS TOPIC OF Arg_2	*weather#1*/Arg_1 *report#2*/Arg_2
16	MANNER	Arg_1 IS MANNER OF Arg_2	*performance#3*/Arg_2 *with passion#1*/Arg_1
17	MEANS	Arg_1 IS MEANS OF Arg_2	*bus#1*/Arg_1 *service#1*/Arg_2
18	EXPERIENCER	Arg_1 IS EXPERIENCER OF Arg_2	*the fear#1*/Arg_2 of the girl#1/Arg_1
19	MEASURE	Arg_2 IS MEASURE OF Arg_1	*inches#1*/Arg_2 *of snow#2*/Arg_1
20	TYPE	Arg_2 IS A TYPE OF Arg_1	*framework#1*/Arg_1 *law#2*/Arg_2
21	THEME	Arg_1 IS THEME OF Arg_2	*acquisition#1*/Arg_2 *of stock#1*/Arg_1
22	BENEFICIARY	Arg_1 IS BENEFICIARY OF Arg_2	*reward#1*/Arg_2 *for the finder#1*/Arg_1
		OTHERS	*cry of death*

Table 2
Lauer's set of prepositions along with examples interpreted in context.

No.	Preposition	Examples
1	of	sea bottom (bottom *of* the sea)
2	for	leisure boat (boat *for* leisure)
3	with	spoon feeding (feeding *with* a spoon)
4	in	London house (house *in* London)
5	on	Saturday snowstorm (snowstorm *on* Saturday)
6	at	night flight (flight *at* night)
7	about	war story (story *about* war)
8	from	almond butter (butter *from* almonds)

N_1), whereas our list contains finer grained semantic categories (e.g., only some N_1 N_2 instances encode a CAUSE relation).

In the next section, we present the coverage of these semantic lists on two different corpora, how well they solve the interpretation problem of noun phrases, and the mapping from one list to another.

3.2 Corpus Analysis

For a better understanding of the semantic relations encoded by N N and N P N instances, we analyzed the semantic behavior of these constructions on two large cross-linguistic corpora of examples. Our intention is to answer questions like:

(1) *What syntactic constructions are used to translate the English instances to the target Romance languages and vice versa?* (cross-linguistic syntactic mapping)

(2) *What semantic relations do these constructions encode?* (cross-linguistic semantic mapping)

(3) *What is the corpus distribution of the semantic relations per each syntactic construction?*

(4) *What is the role of English and Romance prepositions in the semantic interpretation of nominal phrases and compounds?*

For questions (1) and (2), we expand the work of Selkirk (1982b), Grimshaw (1990), Giorgi and Longobardi (1991), and Alexiadou, Haegeman, and Stavrou (2007) on the syntax of noun phrases in English and Romance languages by providing cross-linguistic empirical evidence for in-context instances on two different corpora based on the set of 22 semantic tags. Following a configurational approach, Giorgi and Longobardi, for example, focus only on synthetic nominal phrases, such as *the capture of the soldier* (THEME), where the noun *capture* is derived through nominalization from the verb *to capture*. Besides synthetic constructions, we also consider root nominal phrases and compounds, such as *family estate* (POSSESSION).

The data
In order to perform empirical investigations of the semantics of nominal phrases and compounds, and to train and test a learning model for the interpretation of noun–noun

instances encoded by these constructions, we collected data from two text collections with different distributions and of different genres, Europarl and CLUVI.

The Europarl data were assembled by combining four of the bilingual sentence-aligned corpora made public as part of the freely available Europarl corpus. Specifically, the Spanish–English, Italian–English, French–English and Portuguese–English corpora were automatically aligned based on exact matches of English translations.[8] Then, only those English sentences which appeared verbatim in all four language pairs were considered. The resulting English corpus contained 10,000 sentences which were syntactically parsed using Charniak's parser (Charniak 2000). From these we extracted 6,200 token instances of N N (49.62%) and N P N (50.38%) constructions.

CLUVI (Linguistic Corpus of the University of Vigo) is an open text repository of parallel corpora of contemporary oral and written languages, a resource that besides Galician also contains literary text collections in other Romance languages. Because the collection provides translations into only two of the Romance languages considered here, Spanish and Portuguese, we focused only on the English–Portuguese and English–Spanish literary parallel texts from the works of Agatha Christie, James Joyce, and H. G. Wells, among others. Using the CLUVI search interfaces we created a sentence-aligned parallel corpus of 4,800 unique English–Portuguese–Spanish sentences. The English version was syntactically parsed using Charniak's parser (Charniak 2000) after which each N N and N P N instance was manually mapped to the corresponding translations. The resulting corpus contains 2,310 English token instances with a distribution of 25.97% N N and 74.03% N P N.

Corpus annotation and inter-annotator agreement

For each corpus, each nominal phrase and compound instance was presented separately to two experienced annotators[9] in a Web interface in context along with the English sentence and its translations. Because the corpora do not cover some of the languages (Romanian in Europarl, and Romanian, Italian, and French in CLUVI), three other native speakers of these languages who were fluent in English provided the translations, which were added to the list. The two computational semantics annotators had to tag each English constituent noun with its corresponding WordNet sense.[10] If the word was not found in WordNet the instance was not considered. The annotators were also asked to identify the translation phrases, tag each instance with the corresponding semantic relation, and identify the semantic arguments Arg_1 and Arg_2 in the semantic argument frame of the corresponding relation. Whenever the annotators found an example encoding a semantic relation or a preposition paraphrase other than those provided, or if they did not know what interpretation to give, they had to tag it as OTHER-SR (e.g., *melody of the pearl*: here the context of the sentence did not indicate the association between the two nouns; *cry of death*: the cry announcing death), and OTHER-PP (e.g., *box* **by** *the wall*, *searches* **after** *knowledge*) respectively.

Tagging each noun constituent with the corresponding WordNet sense in context is important not only as a feature employed in the training models, but also as guidance for the annotators to select the right semantic relation. For instance, in the following sentences, *daisy flower* expresses a PART–WHOLE relation in Example (1) and an

8 This version of the Europarl text collection does not include Romanian.

9 The annotators have extensive expertise in computational semantics and are fluent in at least three of the Romance languages considered for this task.

10 We used version 2.1 of WordNet.

IS-A relation in Example (2) depending on the sense of the noun *flower* (cf. WordNet 2.1: *flower#2* is a "reproductive organ of angiosperm plants especially one having showy or colorful parts," whereas *flower#1* is "a plant cultivated for its blooms or blossoms").

(1) Usually, more than one *daisy#1 flower#2* grows on top of a single stem.

(2) Try them with orange or yellow flowers of red-hot poker, solidago, or other late *daisy#1 flowers#1*, such as rudbeckias and heliopsis.

In cases where noun senses were not enough for relation selection, the annotators had to rely on a larger context provided by the sentence and its translations.

Moreover, because the order of the semantic arguments in a nominal phrase or noun–noun compound is not fixed (Girju et al. 2005), the annotators were presented with the semantic argument frame for each of the 22 semantic relations and were asked to tag the instances accordingly. For example, in PART–WHOLE instances such as *chair*/Arg_1 *arm*/Arg_2 the part *arm* follows the whole *chair*, whereas in *spoon*/Arg_1 *handle*/Arg_2 the order is reversed. In the annotation process the translators also used the five corresponding translations as additional information in selecting the semantic relation. For instance, the context provided by the Europarl English sentence in Example (3) does not give enough information for the disambiguation of the English nominal phrase *judgment of the presidency*, where the modifier noun *presidency* can be either AGENT or THEME in relation to the nominalized noun head *judgment*. The annotators had to rely on the Romance translations in order to identify the correct meaning in context (THEME): *valoración sobre la Presidencia* (Sp. – Spanish), *avis sur la présidence* (Fr. – French), *giudizio sulla Presidenza* (It. – Italian), *veredicto sobre a Presidência* (Port. – Portuguese), *evaluarea Preşendinţiei* (Ro. – Romanian).

Most of the time, one instance was tagged with one semantic relation, and one preposition paraphrase (in case of noun–noun compounds), but there were also situations in which an example could belong to more than one category in the same context. For example, *Texas city* is tagged as PART–WHOLE, but also as a LOCATION relation using the 22-SR classification set, and as *of, from, in* based on the 8-PP set (e.g., *city of Texas*, *city from Texas*, and *city in Texas*). Overall, 8.2% CLUVI and 4.8% Europarl instances were tagged with more than one semantic relation, and almost half of the noun–noun compound instances were tagged with more than one preposition.

(3)
En.:	If you do , *the final judgment of the Spanish presidency* will be even more positive than it has been so far.
Sp.:	Si se hace, la valoración sobre la Presidencia española del Consejo será aún mucho más positiva de lo que es hasta ahora.
Fr.:	Si cela arrive, notre avis sur la présidence espagnole du Conseil sera encore beaucoup plus positif que ce n'est déjà le cas.
It.:	Se ci riuscirà, il nostro giudizio sulla Presidenza spagnola sarà ancora più positivo di quanto non sia stato finora.
Port.:	Se isso acontecer, o nosso veredicto sobre a Presidência espanhola será ainda muito mais positivo do que o actual.
Ro.:	Dacă are loc, evaluarea Preşedinţiei spaniole va fi încă mai positiva decât până acum.

Thus, the corpus instances used in the corpus analysis phase have the following format: ⟨NP_{En}; NP_{Es}; NP_{It}; NP_{Fr}; NP_{Port}; NP_{Ro}; target⟩. The word *target* is one of the 23 (22 + OTHER-SR) semantic relations and one of the eight prepositions considered for noun compound instances, and one of the 23 semantic relations for N P N instances. For example, ⟨*development cooperation; cooperación para el desarrollo; cooperazione allo sviluppo; coopération au développement; cooperação para o desenvolvimento; cooperare de dezvoltare;* PURPOSE / FOR⟩.

Inter-annotator agreement was measured using kappa, one of the most frequently used measures of inter-annotator agreement for classification tasks: $K = \frac{Pr(A)-Pr(E)}{1-Pr(E)}$, where $Pr(A)$ is the proportion of times the annotators agree and $Pr(E)$ is the probability of agreement by chance. The K coefficient is 1 if there is a total agreement among the annotators, and 0 if there is no agreement other than that expected to occur by chance.

The kappa values along with percentage agreements obtained on each corpus are shown in Table 3. We also computed the number of instances that were tagged with OTHER by both annotators for each semantic relation and preposition paraphrase, over the number of examples classified in that category by at least one of the judges. For the instances that encoded more than one classification category, the agreement was measured on the first relation on which the annotators agreed.

The agreement obtained for the Europarl corpus is higher than that for CLUVI on both classification sets. Overall, the K coefficient shows a fair to good level of agreement for the corpus data on the set of 22 relations, with a higher agreement for the preposition paraphrases. However, according to Artstein (2007), kappa values can drop significantly if the frequency distribution of the annotation categories in the text corpus is skewed. This is the case here, as will be shown in the next section. Thus, for a better understanding of the annotation results we also computed the percentage agreement, which is indicated for each classification set in parentheses in Table 3.

7.8% of Europarl and 5.7% of CLUVI instances that could not be tagged with Lauer's prepositions were included in the OTHER-PP category. From these, 2.1% and 2.3%, respectively, could be paraphrased with prepositions other than those considered by Lauer (e.g., *bus service*: *service by bus*), and 5.7% and 3.4%, respectively, could not be paraphrased with prepositions (e.g., *daisy flower*).

In the next section we discuss the distribution of the syntactic and semantic interpretation categories on the two different cross-linguistic corpora.

Table 3
The inter-annotator agreement on the annotation of the nominal phrases and compounds in the two corpora. For the instances that encoded more than one classification category, the agreement was measured on the first relation on which the annotators agreed. N/A = not applicable.

		Kappa Agreement (% agreement)		
Corpus	Classification tag sets	**N N**	**N P N**	OTHER
Europarl	**8 PPs**	0.80 (85.4%)	N/A	91%
	22 SRs	0.61 (76.1%)	0.67 (80.8%)	78%
CLUVI	**8 PPs**	0.77 (84.7%)	N/A	86%
	22 SRs	0.56 (73.8%)	0.58 (75.1%)	69%

3.3 Distribution of Syntactic Constructions and Semantic Relations

A. Cross-linguistic distribution and mapping of nominal phrases and compounds

Table 4 shows the distribution of various syntactic constructions used for the translation of the 6,200 (3,076 N N and 3,124 N P N) Europarl and 2,310 (600 N N and 1,710 N P N) CLUVI English token instances in each of the five target languages considered. The data show that N N and N P N constructions cover over 83% of the translation patterns for both text corpora. However, whereas the distribution of both constructions is balanced in the Europarl corpus (about 45%, with the exception of Romanian for which N P N constructions are less frequent), in CLUVI the N P N constructions occur in more than 85% of the cases (again, with the exception of Romanian where they represent about 56% of the data). The high percentage obtained for N P N instances in CLUVI is explained by the fact that Romance languages have very few N N compounds which are of limited semantic types, such as TYPE. Moreover, it is interesting to note here that some of the English instances are translated into both noun–noun (N N) and noun–adjective (N A) compounds in the target languages. For example, *love affair* translates into either the N A construction *enredo amoroso* (Spanish), *aventure amoureuse* (French), *relazione amorosa* (Italian), *relaçao amorosa* (Portuguese), and *aventură amoroasă* (Romanian), or using the more common N *de* N pattern *aventura de amor* (Spanish), *aventure d'amour* (French), *storia d'amore* (Italian), *estoria de amor* (Portuguese), and *aventură de dragoste* (Romanian). There are also instances which translate as one word in the target language, shown in Table 4, column 6. For example,

Table 4
The distribution of syntactic constructions used in the translation of 6,200 Europarl and 2,310 English NN and N P N instances. N A = noun–adjective; pph = other syntactic paraphrase.

		Syntactic distribution					
Corpus	**Language**	**N N**	**N P N**	**N A**	**word**	**pph**	**Total**
Europarl	**French**	2,747 (44.31%)	2,896 (46.71%)	372 (5.99%)	37 (0.6%)	148 (2.39%)	6,200
	Italian	2,896 (46.71%)	2,413 (38.92%)	520 (8.38%)	111 (1.8%)	260 (4.19%)	
	Spanish	2,896 (46.71%)	2,487 (40.12%)	483 (7.79%)	36 (0.58%)	298 (4.80%)	
	Portuguese	2,858 (46.1%)	2,301 (37.11%)	594 (9.58%)	75 (1.21%)	372 (6%)	
	Romanian	4,010 (64.68%)	1,596 (25.74%)	297 (4.79%)	74 (1.19%)	223 (3.6%)	
CLUVI	**French**	32 (1.39%)	1,967 (85.15%)	94 (4.07%)	154 (6.66%)	63 (2.73%)	2,310
	Italian	25 (1.08%)	2,046 (88.57%)	75 (3.25%)	113 (4.89%)	51 (2.21%)	
	Spanish	25 (1.08%)	1,959 (84.81%)	107 (4.63%)	163 (7.06%)	56 (2.42%)	
	Portuguese	25 (1.08%)	1,990 (86.15%)	163 (7.05%)	88 (3.81%)	44 (1.91%)	
	Romanian	758 (32.81%)	1,295 (56.06%)	88 (3.81%)	125 (5.41%)	44 (1.91%)	

ankle boot is translated into *bottine* in French and *stivaletto* in Italian. The rest of the data is encoded by other syntactic paraphrases, as shown in Table 4, column 7. For example, *bomb site* is translated into Italian as *luogo dove è esplosa la bomba* ('the place where the bomb has exploded'). Moreover, Table 5 shows the distribution of the prepositions present in the N P N translations.

Table 5
The distribution of N P N constructions used in the translation of the English noun phrase instances on both text corpora. The preposition *a* is used to denote *a, ad*, and *de* to denote simple and articulated prepositions (*de, di, du, de la, della, degli, d'*, etc.).

Corpus	Language	N P N distribution	Total
Europarl	English	of (81.15%); for (3.27%); in (4.61%); on (2.43%); at (1.22%); from (0.67%); with (2.85%); by (1.5%); against (0.42%); through (0.29%); under (0.42%); after (0.38%); before (0.85%)	3,124
	French	de (75.69%); à (2.93%); pour (6.42%); par (1.42%); en (1.62%); avec (1.6%) ; devant (1.6%); après (1.21%); dans (2.11%); sur (2.6%); contre (0.4%); avant (0.4%)	2,896
	Italian	de (71.78%); a (7%); su (1.29%); a (3.11%); da (6.59%); per (6.22%); via (0.79%); in (0.79%); con (1.41%); contra (0.62%); davanti (0.2%); dopo (0.2%)	2,413
	Spanish	de (83.39%); a (1.81%); en (1.41%); para (3.5%); por (2.61%); con (3.18%); sobre (3.3%); contra (0.4%); en materia de (0.4%)	2,487
	Portuguese	de (78.4%); a (0.8%); em (0.8%); para (3.5%); por (1.6%); com (0.8%); sobre (1.3%); antes de (0.4%)	2,301
	Romanian	de (82.2%); înainte de (1.82%); cu (1.82%); pentru (4.51%); despre (1.63%); la (0.38%); datorită (0.38%); pe (6.08%); pe calea (0.37%); în (0.81%)	1,596
CLUVI	English	of (83.80%); for (1.17%); in (5.90%); on (2.40%); at (0.76%); with (1.99%); against (1.17%); through (0.41%); over (0.41%); above (0.41%); beside (0.41%); about (0.41%); behind (0.76%)	1,710
	French	de (82.33%); à (6.2%); pour (1.42%); en (1.8%); sur (7.02%); contre (0.41%); près de (0.41%); à coté de (0.41%)	1,967
	Italian	de (75.42%); a (8.07%); su (1.32%); da (6.6%); per (6.21%); in (0.78%); con (0.4%); contra (0.4%); sopra (0.2%); accanto a (0.2%); dietro de (0.2%); via (0.2%)	2,046
	Spanish	de (85.96%); a (2.81%); en (3.89%); para (0.71%); por (1.74%); con (2.1%); sobre (1.38%); contra (0.36%); detrás (0.71%); encima (0.36%)	1,959
	Portuguese	de (78.4%); a (0.8%); em (0.82%); para (3.5%); por (1.6%); com (0.8%); sobre (1.3%); acima de (0.4%)	1,990
	Romanian	de (85.21%); cu (1.82%); pentru (4.5%); la (0.4%); datorită (0.4%); pe (5.08%); despre (1.58%); în (0.79%); lânga (0.2%)	1,295

For the purposes of this research, from the 6,200 Europarl and 2,310 CLUVI instances, we selected those which had all the translations encoded only by N N and N P N constructions. Columns 3 and 4 in Table 4 show the number of N N and N P N translation instances in each Romance language. Out of these, we considered only 3,124 Europarl and 2,023 CLUVI token instances representing the examples encoded by N N and N P N in all languages considered, after inter-annotator agreement.

B. Cross-linguistic distribution of semantic relations and their mapping to nominal phrases and compounds

A closer look at the N N and N P N translation instances in Table 4 shows that their syntactic distribution is influenced by the text genre and the semantics of the instances. For example, in Europarl most of the N N instances were naming noun–noun compounds referring to entities such as *member states* and *framework law* which were repeated in many sentences. Many of them encoded TYPE relations (e.g., *member state*, *framework law*) which, most of the time, are encoded by N N patterns in the target languages (*stato membro* and *legge quadro* in Italian, respectively). In the CLUVI corpus, on the other hand, the N N Romance translations represented only 1% of the data. A notable exception here is Romanian (64.68% of Europarl and 32.8% of CLUVI). This is explained by the fact that, in Romanian, many noun phrases are represented as genitive-marked noun compounds (N_1 N_2). In Romanian the genitive case is realized either as a suffix attached to the modifier noun N_2 or as one of the genitival articles *a/al/ale*. If the modifier noun N_2 is determined by an indefinite article then the genitive mark is applied to the article, not to the noun, for example *o fată – unei fete* ('a girl – of/to a girl') and *un băiat – unui băiat* ('a boy – of/to a boy'). Similarly, if the modifier noun is determined by the definite article (which is enclitic in Romanian), the genitive mark is added at the end of the noun together with the article. For example, *fata–fetei* (the girl – girl-GEN), *cartea–cărţii* (the book – book-GEN). Thus, the noun phrase *the beauty of the girl*, for instance, is translated as *frumuseţea fetei* ('beauty-the girl-GEN'), and *the beauty of a girl* as *frumuseţea unei fete* ('beauty-the of/to a girl').

In general, in Romanian the choice between the N *de* N and the genitive-marked N N constructions depends on the specificity of the instance. Some noun–noun instances refer to a specific entity (existential interpretation), in which case the construction preferred is the genitive-marked N N, or they can refer in general to the category of those entities (generic interpretation),[11] thus using N *de* N. For example, the instance *the bite of the scorpion* (AGENT) translates into *muşcătura scorpionului* ('bite-the scorpion-GEN'), whereas *a scorpion bite* (AGENT) translates into *muşcătură de scorpion* ('bite of scorpion').

Many semantic relations that allow both the generic and the existential interpretations can be encoded by both N P N and genitive-marked N N constructions as shown by the example above. However, there are situations when the generic and the existential interpretations change the meaning of the noun–noun pair. One such example is *the suit of the sailor* (POSSESSION) translated as *costumul marinarului* ('suit-the sailor-GEN'), and *sailor suit* (PURPOSE) translated as *costum de marinar* ('suit of sailor').

11 The words *existential* and *generic* are borrowed here from the vast linguistic literature on definite and indefinite descriptions. Here, nouns such as *firemen* can have different readings in various contexts: *Firemen are available* (existential reading), vs. *Firemen are altruistic* (generic reading).

At the other extreme there are relations which prefer either the generic or the existential interpretation. For example, some POSSESSION-encoding instances such as *the budget of the University* translate as '*bugetul Universităţii*' (budget-the University-GEN) and not as '*bugetul de Universitate*' (budget-the of University). Other relations such as PURPOSE and SOURCE identify generic instances. For example, (a) *olive oil* (SOURCE) translates as '*ulei de măsline*' (oil of olive), and not as '*uleiul măslinei*' (oil-the olive-GEN), and (b) *the milk glass* (PURPOSE) translates as '*paharul de lapte*' (glass-the of milk) and not as '*paharul laptelui*' (glass-the milk-GEN). Other examples include CAUSE and TOPIC. This observation is very valuable for the interpretation of nominal phrases and compounds and is used in the learning model to discriminate among the possible interpretations.

Tables 6 and 7 show the semantic distribution of the instances on both text corpora. This distribution is represented both in number of tokens (the total number of instances per relation) and types (the unique number of instances per relation). In Europarl, the most frequently occurring relations are TYPE and THEME that together represent about 50% of the data with an equal distribution. The next most frequent relations are TOPIC, PURPOSE, AGENT, and PROPERTY with an average coverage of about 8%. Moreover, eight relations of the 22-SR set (KINSHIP, DEPICTION, CAUSE, INSTRUMENT, SOURCE, MANNER, MEASURE, and BENEFICIARY) did not occur in this corpus. The 9.61% of the OTHER-SR relation represents the ratio of those instances that did not encode any of the 22 semantic relations. It is interesting to note here the large difference between the number of types versus tokens for the TYPE relation in Europarl. This is accounted for by various N N instances such as *member states* that repeat across the corpus.

This semantic distribution contrasts with the one in CLUVI. Here, the most frequent relation by far is PART–WHOLE (40.53%), followed by LOCATION (8.95%), AGENT (6.23%), and IS-A (5.93%). The missing relations are KINSHIP, MANNER and BENEFICIARY. A larger percentage of OTHER-SR instances (12.95%) did not encode any of the 22 semantic relations. Moreover, in CLUVI 256 instances were tagged with more than one semantic relation with the following distribution: 46.8% MEASURE/PART–WHOLE (e.g., *a couple of cigarettes*), 28.2% PART–WHOLE/LOCATION (e.g., *bottom of the sea*), 10.9% MEASURE/LOCATION (e.g., *cup of chocolate*), 8.2% PURPOSE/LOCATION (e.g., *waste garden*), and 5.9% THEME/MAKE-PRODUCE (e.g., *makers of songs*). In Europarl, on the other hand, there were only 97 such cases: 81.4% THEME/MAKE-PRODUCE (e.g., *bus manufacturers*) and 18.6% MEASURE/PART–WHOLE (e.g., *number of states*).

One way to study the contribution of both the English and Romance prepositions to the interpretation task is to look at their distribution over the set of semantic relations on two reasonably large text corpora of different genres. Of course, this approach does not provide an analysis that generates an exhaustive generalization over the properties of the language. However, as Tables 6 and 7 show, there are dependencies between the structure of the Romance language translations and the semantic relations encoded by the nominal phrases and compounds, although the most frequently occurring prepositions are *de* and its English equivalent *of*. Here we use the preposition *de* to represent a set of translation equivalents in Romance languages (e.g., the Italian counterpart is *di*). These prepositions are semantically underspecified, encoding a large set of semantic relations. The many-to-many mappings of the prepositions to the semantic classes adds to the complexity of the interpretation task. For example, in the Europarl corpus LOCATION is encoded in French by *de, sur, devant*, and *à près de*, while TOPIC is encoded in English by *of*, *for*, *on*, *about* and noun compounds, and in Spanish by *de*, *sobre*, *en materia de*.

Table 6
Mapping between the set of 22 semantic classification categories and the set of English and Romance syntactic constructions on the Europarl corpus. The preposition *de* is used here to denote simple and articulated prepositions (*de, di, du, de la, della, degli, d'*, etc.). Also, the dash "–" refers to noun–noun compounds where there is no connecting preposition. The mapping was obtained on the 3,124 Europarl instance corpus. En. = English; Sp. = Spanish; It. = Italian; Fr. = French; Port. = Portuguese; Ro. = Romanian.

								Total		
Nr.	**SRs**	**En.**	**Sp.**	**It.**	**Fr.**	**Port.**	**Ro.**	Token [%]	Type [%]	**Example**
1	POSSESSION	of, –	de, –	de, –	de		de, –	2.85	2.4	*Union resources* '*resursele uniunii*' (Ro.) (resource-the union-GEN)
2	KINSHIP							0	0	
3	PROPERTY	of, for, in, –	de	de	de	de	de, –	6.05	6.05	*traffic density* '*densità del traffico*' (It.) (density of traffic)
4	AGENT	of, for, in, by, –	de	de,–	de	de	de	7.47	7.08	*request of a member* '*richiesta di uno membro*' (It.) (request of a member)
5	TEMPORAL	of, in, on, at, –	de, con a	de	de, avant	de acima de	de, înainte de	0.04	0.04	*year before the constitution* '*año anterior a la constitución*' (Sp.) (the year previous of the) constitution
6	DEPICTION							0	0	
7	PART–WHOLE	of, in, with, –	de, con	de, a, a, –	de, à	de,	de con	3.20	2.75	*Union citizen* '*citoyen de l' Union*' (Fr.) (citizen of the Union)
8	IS–A (HYPERNYMY)	of, – with	de, –	de, –	de, –	–	–	0.8	0.8	*process of decay* '*proces de descompunere*' (Ro.) (process of decay)
9	CAUSE							0	0	
10	MAKE/ PRODUCE	of, for, – in, from	de	de	de	de	de	1.43	1.43	*paper plant* '*fábrica de papel*' (Sp.) (plant of paper)
11	INSTRUMENT							0	0	
12	LOCATION	of, in, on, – at	de, en, sobre	de, su, a, in	de, sur, à, près de, devant	de	de, pe, la, în	2.14	2.14	*place of the meeting* '*lieu de la réunion*' (Fr.) (place of the meeting)
13	PURPOSE	of, – for	de, por, para, contra	de, da, per, a, –	contre, à, de, – pour	de, a	de, pentru	7.48	7.23	*building stone* '*pedras de construção*' (Port.) (stones of building)
14	SOURCE							0	0	
15	TOPIC	of, for, on, – about	de, sobre, en materia de	de, a, su	de	de, sobre	de, despre	11.03	11.03	*policy on asylum* '*política en materia de asilo*' (Sp.) (policy in regard to asylum)
16	MANNER							0	0	
17	MEANS	by de, –	por, en,	per, in, a, via	en, à, par	por	pe, cu, pe calea	0.07	0.07	*travel by train* '*calatorie cu tenul*' (Ro.) (travel with train-the)
18	EXPERIENCER	of,– in in	de	de	de	de	de, –	0.04	0.04	*suffering of the people* '*sofrimento das pessoas*' (Port.) (suffering of the people)
19	MEASURE							0	0	
20	TYPE	–	–	–	–	–	–	24.47	1.7	*framework law* '*legge quadro*' (It.) (law framework)
21	THEME	of, for, in, –	de	de, a	de	de	de	23.13	19.2	*conflict prevention* '*prevenire de conflict*' (Ro.) (prevention of conflict)
22	BENEFICIARY							0	0	
23	OTHER–SR	of, by	de	a, de	de, à	de, a, com	de, pentru	9.61	8.13	*tobacco addiction* '*adicción a tabaco*' (Sp.) (addiction to tobacco)
Total no. of examples								3,124	2,190	

Moreover, in the Europarl corpus, 31.64% of the instances are synthetic phrases encoding AGENT, MEANS, LOCATION, THEME, and EXPERIENCER. Out of these instances, 98.7% use the preposition *of* and its Romance equivalent *de*. In the CLUVI corpus, 14.1% of the examples were verbal, from which the preposition *of/de* has a coverage of 77.66%.

Based on the literature on prepositions (Lyons 1986; Barker 1998; Ionin, Matushansky, and Ruys 2006) and our own observations, the preposition *of/de* in both root and synthetic nominal phrases may have a functional or a semantic role, acting as a linking device with no apparent semantic content, or with a meaning of its own. Thus, for the interpretation of these constructions a system must rely on the meaning of preposition and the meaning of the two constituent nouns in particular, and on context

Table 7
Mapping between the set of 22 semantic classification categories and the set of English and Romance syntactic constructions on the CLUVI corpus. The preposition *de* is used here to denote simple and articulated prepositions (*de*, *di*, *du*, *de la*, *della*, *degli*, *d'*, etc.). Also, the dash "–" refers to noun–noun compounds where there is no connecting preposition. The mapping was obtained on the 2,023 CLUVI instance corpus. En. = English; Sp. = Spanish; It. = Italian; Fr. = French; Port. = Portuguese; Ro. = Romanian.

								Total		
Nr.	**SRs**	**En.**	**Sp.**	**It.**	**Fr.**	**Port.**	**Ro.**	Token [%]	Type [%]	**Example**
1	POSSESSION	of, –	de, –	de, –	de	de	de, –	1.35	1.21	*police car* '*coche de polizia*' (Sp.) (car of police)
2	KINSHIP							0	0	
3	PROPERTY	of, for, in, –	de	de	de	de	de, –	2.97	2.76	*beauty of the buildings* '*belleza de los edificios*' (Sp.) (beauty of the buildings)
4	AGENT	of, for, in, by, –	de	de, –	de	–	de, –	6.23	5.78	*return of the family* '*regresso da família*' (Port.) (return of the family)
5	TEMPORAL	of, in, on, at, –	de, con	de	de	de	de	2.97	2.97	*spring rain* '*pluie de printemps*' (Fr.) (rain of spring)
6	DEPICTION–DEPICTED	of	de	de	de	de	de	0.3	0.3	*picture of a girl* '*retrato de uma rapariga*' (Port.) (picture of a girl)
7	PART–WHOLE	of, in, with, –	de, con	de, a, –	de, à	de com	de, –	40.53	34.35	*ruins of granite* '*ruinas de granito*' (Sp.) (ruins of granite)
8	IS–A (HYPERNYMY)	of, – with	de, –	de, –	de, –	–	de	5.93	5.4	*sensation of fear* '*sensaçáo de medo*' (Port.) (sensation of fear)
9	CAUSE	from, –	de	de, da	de	de	de, datorită	2.72	2.72	*cries of delight* '*cri de joie*' (Fr.) (cries of delight)
10	MAKE/ PRODUCE	of, for, in, from, –	de	de	de	de	de	0.29	0.29	*noise of the machinery* '*ruido de la maquinaria*' (Sp.) (noise of the machinery)
11	INSTRUMENT	for, with	de, –	de, a, con	de, à	de	de, cu	0.29	0.29	*a finger scratch* '*o zgârietură de unghie*' (Ro.) (a scratch of finger)
12	LOCATION	of, in, on, at, –	de, en, sobre,	de, su, a, in, dietro de, accanto a, sopra	de, sur, à, près de, à coté de	de, em acima de	de, pe, la, în, lânga	8.65	8.01	*book on the table* '*livre sur la table*' (Fr.) (book on the table)
13	PURPOSE	of, – for	de, por, para, contra	de, da, per, a, – contra	contre, a, de, – pour	de	de, pentru	4.45	4.45	*nail brush* '*spazzolino per le unghie*' (It.) (brush for the nails)
14	SOURCE	of, from	de	de	de	de	de	0.94	0.15	*oil of cloves* '*óleo de cravinho*' (Port.) (oil of cloves)
15	TOPIC	of, for, on, about, –	de, sobre	de, a, su	de	de, sobre	de, despre	0.79	0.79	*love story* '*histoire d'amour*' (Fr.) (story of love)
16	MANNER							0	0	
17	MEANS	of, by	por	via	à	por	pe	0.15	0.15	*travel by car* '*călătorie cu maşina*' (Ro.) (travel by car)
18	EXPERIENCER	of, in, –	de	de	de	de	de, –	0.64	0.64	*the agony of the prisoners* '*l'agonia dei prigionieri*' (It.) (the agony of.the prisoners)
19	MEASURE	of	por	de	à	de	de, pentru	3.81	2.72	*a cup of sugar* '*o ceaşcă de zahăr*' (Ro.) (a cup of sugar)
20	TYPE							0	0	
21	THEME	for, – of, in	de	de, a	de	de	de, a, –	4.05	3.94	*lack of intelligence* '*manque d'intelligence*' (Fr.) (lack of intelligence)
22	BENEFICIARY							0	0	
23	OTHER–SR	of, by	de	de, a	de	de, a,	de	12.95	8.81	*cry of death* '*cri de mort*' (Fr.) (cry of death)
Total no. of examples								2,023	1,734	

in general. Because the two corpora used in this paper contain both root and synthetic instances, we employed two semantic resources for this task: WordNet noun semantic classes and a collection of verb classes in English that correspond to special types of nominalizations. These resources are defined in Section 4.2. Moreover, in Section 6 we present a detailed linguistic analysis of the prepositions *of* in English and *de* in Romance languages, and show how their selection correlates with the meaning of the construction.

4. Model

4.1 Mathematical Formulation

Given the syntactic constructions considered, the goal is to develop a procedure for the automatic annotation of the semantic relations they encode. The semantic relations derive from various lexical and semantic features of each instance.

The semantic classification of instances of nominal phrases and compounds can be formulated as a learning problem, and thus benefits from the theoretical foundation and experience gained with various learning paradigms. The task is a multi-class classification problem since the output can be one of the semantic relations in the set. We cast this as a supervised learning problem where input/output pairs are available as training data.

An important first step is to map the characteristics of each instance (i.e., list of properties that describe the instance, usually not numerical) into feature vectors. Let us define x_i as the feature vector of an instance i and let X be the space of all instances; that is, $x_i \in X$.

The multi-class classification is performed by a function that maps the feature space X into a semantic space $S, f : X \rightarrow S$, where S is the set of semantic relations from Table 1, namely, $r_j \in S$, where r_j is a semantic relation.

Let T be the training set of examples or instances $T = (x_1 r_1 \,..\, x_l r_l) \subseteq (X \text{ x } S)^l$ where l is the number of examples x each accompanied by its semantic relation label r. The problem is to decide which semantic relation to assign to a new, unseen example x_{l+1}. In order to classify a given set of examples (members of X), one needs some kind of measure of the similarity (or the difference) between any two given members of X.

Thus, the system receives as input an English nominal phrase and compound instances along with their translations in the Romance languages, plus a set of extra-linguistic features. The output is a set of learning rules that classify the data based on the set of 22 semantic target categories. The learning procedure is supervised and takes into consideration the cross-linguistic lexico-syntactic information gathered for each instance.

4.2 Feature Space

The set of features allows a supervised machine learning algorithm to induce a function that can be applied to accurately classify unseen instances. Based on the study of the instances and their semantic distribution presented in Section 3, we have identified and experimented with the following features presented subsequently for each language involved. Features F1–F5 have been employed by us in our previous research (Moldovan et al. 2004; Girju et al. 2005; Girju, Badulescu, and Moldovan 2006). All the other features are novel.

A. English features

F1 and **F2.** *Semantic class of noun* specifies the WordNet sense of the head noun (F1), and the modifier noun (F2) and implicitly points to all its hypernyms. The semantics of the instances of nominal phrases and compounds is heavily influenced by the meaning of the noun constituents. One such example is *family#2 car#1*, which encodes a POSSESSION relation. The hypernyms of the head noun *car#1* are: {*motor vehicle*}, {*self-propelled*

vehicle} ... {*entity*} (cf. WordNet 2.1). These features will help generalize over the semantic classes of the two nouns in the instance corpus.

F3 and **F4.** *WordNet derivationally related form* specifies if the head noun (F3), and the modifier noun (F4) are related to a corresponding verb in WordNet. WordNet contains information about nouns derived from verbs (e.g., *statement* derived from *to state*; *cry* from *to cry*; *death* from *to die*).

F5. *Prepositional cues* link the two nouns in a nominal phrase. These can be either simple or complex prepositions such as *of* or *according to*. In case of N N instances (e.g., *member state*), this feature is "–".

F6 and F7. *Type of nominalized noun* indicates the specific class of nouns the head (F6) or modifier (F7) belongs to depending on the verb from which it derives. First, we check if the noun is a nominalization or not. For English we used the NomLex-Plus dictionary of nominalizations (Meyers at al. 2004) to map nouns to corresponding verbs.[12] One such example is *the destruction of the city*, where *destruction* is a nominalization. F6 and F7 may overlap with features F3 and F4 which are used in case the noun to be checked has no entry in the NomLex-Plus dictionary.

These features are of particular importance because they impose some constraints on the possible set of relations the instance can encode. They take the following values: a) active form nouns, b) unaccusative nouns, c) unergative nouns, and d) inherently passive nouns. We present them in more detail subsequently.

a. Active form nouns are derived through nominalization from psych verbs and represent states of emotion, such as *love, fear, desire*, and so forth. They have an intrinsic active voice predicate–argument structure and, thus, resist passivisation. For example, we can say *the desire of Anna*, but not *the desire by Anna*. This is also explained by the fact that in English the AGENT or EXPERIENCER relations are mostly expressed by the clitic genitive *'s* (e.g., *Anna's desire*) and less or never by N P N constructions. Citing Anderson (1983), Giorgi and Longobardi (1991) mention that with such nouns that resist passivisation, the preposition introducing the internal argument, even if it is *of*, has always a semantic content, and is not a bare case-marker realizing the genitive case. Moreover, they argue that the meaning of these nouns might pattern differently in different languages. Consider for example the Italian sentences (4) and (5) below and their English equivalents (see Giorgi and Longobardi 1991, pages 121–122). In English the instance *Anna's desire* identifies the subject of desire (and thus encodes an EXPERIENCER relation), whereas in Italian it can identify either the subject (EXPERIENCER) as in Example (4), or the object of desire (THEME) as in Example (5), the disambiguation being done at the discourse level. In Example (6) the prenominal construction *il suo desiderio* encodes only EXPERIENCER.

(4) *Il desiderio di Anna* fu esaudito. (EXPERIENCER)
(*The desire of Anna* was fulfilled.)
'*Anna's desire* was fulfilled.'

(5) *Il desiderio di Anna* lo porterà alla rovina. (THEME)
(*The desire of Anna* him will ruin.)
'*The desire for Anna* will ruin him.'

12 NomLex-Plus is a hand-coded database of 5,000 verb nominalizations, de-adjectival, and de-adverbial nouns including the corresponding subcategorization frames (verb-argument structure information).

(6) *Il suo desiderio* fu esaudito. (EXPERIENCER)
(*The her desire* was fulfilled.)
'*Her desire* was fulfilled.'

However, our observations on the Romanian training instances in Europarl and CLUVI (captured by features F12 and F13 below) indicate that the choice of syntactic constructions can help in the disambiguation of instances that include such active nouns. Thus, whereas genitive-marked N N compounds identify only the subject (thus encoding EXPERIENCER), the N *de/pentru* N constructions identify only the object (thus encoding THEME). Such examples are *dorinţa Anei* ('desire-the Anna-GEN' – *Anna's desire*) (EXPERIENCER) and *dorinţa de/pentru Ana* ('desire-the of/for Anna' – *the desire for Anna*) (THEME).

Another example is *the love of children* and not *the love by the children*, where children are the recipients of love, not its experiencers. In Italian the instance translates as *l'amore per i bambini* ('the love for the children'), whereas in Romanian it translates as *dragostea pentru copii* ('love-the for children'). These nouns mark their internal argument through *of* in English and most of the time require prepositions such as *for* in Romance languages and vice versa.

b. Unaccusative nouns are derived from ergative verbs that take only internal arguments (e.g., those that indicate an object and not a subject grammatical role). For example, the transitive verb *to disband* allows the subject to be deleted as in the following sentences:

(7) The lead singer disbanded the group in 1991.

(8) The group disbanded.

Thus, the corresponding unaccusative nominalization of *to disband*, *the disbandment of the group*, encodes THEME and not AGENT.

c. Unergative nouns are derived from intransitive verbs. They can take only AGENT semantic relations. One such case is exemplified in the instance *l'arrivo della cavalleria* in Italian which translates in English as *the arrival of the cavalry* and in Romanian as *sorirea cavaleriei* ('arrival-the cavalry-GEN').

d. Inherently passive nouns. These nouns, like the verbs they are derived from, assume an implicit AGENT relation and, being transitive, associate to their internal argument the THEME relation. One such example is *the capture of the soldier* which translates in Italian as *la cattura del soldato* ('the capture of the soldier'), *la capture du soldat* in French ('the capture of soldier'), and *la captura de soldado* in Spanish and Portuguese ('the capture of soldier'), where the nominalization *capture* (*cattura, capture, captura* in Italian, French, and Spanish and Portuguese respectively) is derived from the verb *to capture*. Here, whereas English and Italian, Spanish, Portuguese, and French use the N *of/de* N construction (as shown in Examples (9) and (10) for English and Italian), Romanian uses genitive-marked noun compounds. In Romanian, however, nominalizations are formed through suffixation, where a suffix is added to the root of the verb it comes from. Different suffixes attached to the same verb may lead, however, to more than one nominalization, producing different meanings. The verb *to capture* (*a captura* in Romanian), for example, can result through suffixation in two nominalizations: *capturare* (with the

infinitive suffix *-are* and encoding an implicit AGENT relation) and *captură* (through zero derivation and encoding an implicit THEME relation) (Cornilescu 2001). Thus, the noun phrase *capturarea soldatului* ('capture-the soldier-GEN') encodes a THEME relation, while *captura soldatului* ('capture-the soldier-GEN') encodes an AGENT relation. In all the Romance languages with the exception of Romanian, this construction is ambiguous, unless the AGENT is explicitly stated or inferred as shown in Example (9) for Italian. The same ambiguity might occur sometimes in English, with the difference that besides the *of*-genitive, English also uses the *s*-genitive: *the soldier's capture* (AGENT is preferred if the context doesn't mention otherwise), *the soldier's capture by the enemy* (THEME), *the capture of the soldier* (THEME is preferred if the context doesn't mention otherwise), *the capture of the soldier by the enemy* (THEME).

(9) *La cattura del soldato* (da parte del nemigo) è cominciata come un atto terroristico. (THEME)
'*The capture of the soldier* (by the enemy) has started as a terrorist act.'

(10) *La sua cattura* è cominciata come un atto terroristico. (THEME)
'*His capture* has started as a terrorist act.'

These nouns have a different behavior than that of active form nouns. As shown previously, the object of inherently passive nouns can move to the subject position as in *the soldier's capture by the enemy*, whereas it cannot do so for active form nouns (e.g., **Anna's desire by John*). Similarly, in Italian, although active form nouns allow only the subject reading in prenominal constructions (e.g., *il suo desiderio* – 'her desire'), inherently passive nouns allow only the object reading (e.g., *la sua cattura* – 'his capture').

For Romanian, the nominalization suffixes were identified based on the morphological patterns presented in Cornilescu (2001).

We assembled a list of about 3,000 nouns that belong to classes a–d using the information on subcategorization frames and thematic roles of the verbs in VerbNet (Kipper, Dang, and Palmer 2000). VerbNet is a database which encodes rich lexical information for a large number of English verbs in the form of subcategorization information, selectional restrictions, thematic roles for each argument of the verb, and alternations (the syntactic constructions in which the verb participates).

B. Romance features

F8, F9, F10, F11, and **F12.** *Prepositional cues* that link the two nouns are extracted from each translation of the English instance: F8 (Sp.), F9 (Fr.), F10 (It.), F11 (Port.), and F12 (Ro.). These can be either simple or complex prepositions (e.g., *de*, *in materia de* [Sp.]) in all five Romance languages, or the Romanian genitival article *a/al/ale*. For N N instances, this feature is "–".

F13. *Noun inflection* is defined only for Romanian and shows if the modifier noun in N N instances is not inflected or is inflected and modifies the head noun which is or is not a nominalization. This feature is used to help differentiate between instances encoded by genitive-marked N N constructions and noun–noun compounds, when the choice of syntactic construction reflects different semantic content. Two such examples are the noun–noun compound *lege cadru* (law framework) (TYPE) which translates as *framework*

law and the genitive-marked N N instance *frumuseţea fetei* ('beauty-the girl-GEN') (PROPERTY) meaning *the beauty of the girl*. It also covers examples such as *capturarea soldatului* ('capture-the soldier-GEN'), where the modifier *soldatului* is inflected and the head noun *capturarea* is a nominalization derived through infinitive suffixation.

In the following Example we present the feature vector for the instance *the capture of the soldiers*.

(11) The instance *the capture of the soldiers* has the following Romance translations:

⟨*capture#4/Arg*$_2$ *of soldiers#1/Arg*$_1$*; captura de soldados; capture du soldats; cattura dei soldati; captura dos soldados; capturarea soldaţilor*; THEME⟩.
Its corresponding feature vector is:

⟨*entity#1/Arg*$_2$*; entity#1/Arg*$_1$*; capture; –; of; inherently passive noun; –; de; de; de; de; –; mod-inflected-inf-nom*; THEME⟩,

where *mod-inflected-inf-nom* indicates that the noun modifier *soldaţilor* in the Romanian translation *capturarea soldaţilor* ('capture-the soldiers-GEN') is inflected and that the head noun *capturarea* is an infinitive nominalization.

4.3 Learning Models

Several learning models can be used to provide the discriminating function f. We have experimented with the support vector machines model and compared the results against two state-of-the-art models: semantic scattering, a supervised model described in Moldovan et al. (2004), Girju et al. (2005), and Moldovan and Badulescu (2005), and Lapata and Keller's Web-based unsupervised model (Lapata and Keller 2005).

Each model was trained and tested on the Europarl and CLUVI corpora using a 7:3 training–testing ratio. All the test nouns were tagged with the corresponding sense in context using a state-of-the-art WSD tool (Mihalcea and Faruque 2004). The default semantic argument frame for each relation was used in the automatic identification of the argument positions.

A. Support vector machines
Support vector machines (SVMs) are a set of related supervised learning methods used for creating a learning function from a set of labeled training instances. The function can be either a classification function, where the output is binary (is the instance of category X?), or it can be a general regression function. For classification, SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface will attempt to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is similar but not identical to the training data.

In order to achieve classification in n semantic classes, $n > 2$, we built a binary classifier for each pair of classes (a total of C_n^2 classifiers), and then we used a voting procedure to establish the class of a new example. For the experiments with semantic relations, the simplest voting scheme has been chosen; each binary classifier has one vote, which is assigned to the class it chooses when it is run. Then the class with the largest number of votes is considered to be the answer. The software used in these experiments is the package LIBSVM (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`) which implements an SVM model. We tested with the radial-based kernel.

After the initial instances in the training and testing corpora were expanded with the corresponding features, we had to prepare them for the SVM model. The set-up procedure is now described.

Corpus set-up for the SVM model:
The processing method consists of a set of iterative procedures of specialization of the examples on the WordNet IS-A hierarchy. Thus, after a set of necessary specialization iterations, the method produces specialized examples which through supervised machine learning are transformed into sets of semantic rules for the semantic interpretation of nominal phrases and compounds. The specialization procedure is described subsequently.

Initially, the training corpus consists of examples that follow the format exemplified at the end of Section 4.2 (Example [11]). Note that for the English instances, each noun constituent was expanded with the corresponding WordNet top semantic class. At this point, the generalized training corpus contains two types of examples: unambiguous and ambiguous. The second situation occurs when the training corpus classifies the same noun–noun pair into more than one semantic category. For example, both relationships *chocolate#2 cake#3* (PART–WHOLE) and *chocolate#2 article#1* (TOPIC) are mapped into the more general type ⟨*entity#1, entity#1,* PART–WHOLE/TOPIC⟩.[13] We recursively specialize these examples to eliminate the ambiguity. By specialization, the semantic class is replaced with the corresponding hyponym for that particular sense, that is, the concept immediately below in the hierarchy. These steps are repeated until there are no more ambiguous examples. For this example, the specialization stops at the first hyponym of *entity*: *physical entity* (for *cake*) and *abstract entity* (for *article*). For the unambiguous examples in the generalized training corpus (those that are classified with a single semantic relation), constraints are determined using cross-validation on the SVM model.

B. Semantic scattering
The semantic scattering (SS) model was initially tested on the classification of genitive constructions, but it is also applicable to nominal phrases and compounds (Moldovan et al. 2004). SS is a supervised model which, like the SVM model described previously, relies on WordNet's IS-A semantic hierarchy to learn a function which separates positive and negative examples. Essentially, it consists of using a training data set to establish a boundary G^* on WordNet noun hierarchies such that each feature pair of noun–noun senses f_{ij} on this boundary maps uniquely into one of a predefined list of semantic relations. The algorithm starts with the most general boundary corresponding to the *entity* WordNet noun hierarchy and then specializes it based on the training data until a good approximation is reached.[14] Any feature pair above the boundary maps into more than one semantic relation. Due to the specialization property on noun hierarchies, feature pairs below the boundary also map into only one semantic relation. For any new pair of noun–noun senses, the model finds the closest boundary pair which maps to one semantic relation.

The authors define with $SC^m = \{f_i^m\}$ and $SC^h = \{f_j^h\}$ the sets of semantic class features for modifier noun and, respectively, head noun. A pair of <modifier, head> nouns maps

13 The specialization procedure applies only to features 1 and 2.
14 Moldovan et al. (2004) used a list of 35 semantic relations – actually only 22 of them proved to be encoded by nominal phrases and compounds.

uniquely into a semantic class feature pair $\langle f_i^m, f_j^h \rangle$ (henceforth f_{ij}). The probability of a semantic relation r given the feature pair f_{ij}, $P(r|f_{ij}) = \frac{n(r,f_{ij})}{n(f_{ij})}$ is defined as the ratio between the number of occurrences of a relation r in the presence of the feature pair f_{ij} over the number of occurrences of the feature pair f_{ij} in the corpus. The most probable semantic relation $\hat{r}$ is

$$\hat{r} = \arg\max_{r \in R} P(r|f_{ij}) = \arg\max_{r \in R} P(f_{ij}|r)P(r) \quad (1)$$

From the training corpus, one can measure the quantities $n(r,f_{ij})$ and $n(f_{ij})$. Depending on the level of abstraction of f_{ij} two cases are possible:

Case 1. The feature pair f_{ij} is specific enough such that there is only one semantic relation r for which $P(r|f_{ij}) = 1$ and 0 for all the other semantic relations.

Case 2. The feature pair f_{ij} is general enough such that there are at least two semantic relations for which $P(r|f_{ij}) \neq 0$. In this case Equation (1) is used to find the most appropriate $\hat{r}$.

Definition
A boundary G^* in the WordNet noun hierarchies is a set of synset pairs such that:

a) for any feature pair on the boundary, denoted $f_{ij}^{G^*} \in G^*$, $f_{ij}^{G^*}$ maps uniquely into only one relation r, and

b) for any $f_{ij}^u \succ f_{ij}^{G^*}$, f_{ij}^u maps into more than one relation r, and

c) for any $f_{ij}^l \prec f_{ij}^{G^*}$, f_{ij}^l maps uniquely into a semantic relation r.

Here relations $\succ$ and $\prec$ mean 'semantically more general' and 'semantically more specific', respectively.

As proven by observation, there are more concept pairs under the boundary G^* than above it, that is, $| \{f_{ij}^l\} | \gg | \{f_{ij}^u\} |$.

Boundary Detection Algorithm

Step 1. Create an initial boundary.
The initial boundary denoted G^1 is formed from combinations of the *entity#1 – entity#1* noun class pairs. For each training example a corresponding feature f_{ij} is first determined, after which it is replaced with the most general corresponding feature consisting of top WordNet hierarchy concepts. For example, both instances *family#2 estate#2* (POSSESSION) and *the sister#1 of the boy#1* (KINSHIP) are mapped into *entity#1 – entity#1*. At this level, the noun–noun feature encodes a number of semantic relations. For each feature, one can determine the most probable relation using Equation (1). For instance, the feature *entity#1 – entity#1* can be encoded by any of the 23 relations.

The next step is to construct a lower boundary by specializing the semantic classes of the ambiguous features. A feature f_{ij} is ambiguous if it corresponds to more than one relation and its most relevant relation has a conditional probability less than 0.9.

To eliminate irrelevant specializations, the algorithm specializes only the ambiguous classes that occur in more than 1% of the training examples.

The specialization procedure consists of first identifying the features f_{ij} to which correspond more than one semantic relation, then replacing these features with their hyponym synsets. Thus one feature breaks into several new specialized features. For example, the feature *entity#1 – entity#1* generated through generalization for the examples *family#2 estate#2* and *the sister#1 of the boy#1* is specialized now as *kin_group#1 – real_property#1* and *female_sibling#1 – male_person#1* corresponding to the direct hyponyms of the nouns in these instances. The net effect is that the semantic relations that were attached to f_{ij} will be 'scattered' across the new specialized features which form the second boundary. The probability of the semantic relations that are encoded by these specialized features is recalculated again using Equation (1). The number of relations encoded by each of this boundary's features is less than the one for the features defining the previous boundary. This process continues until each feature has only one semantic relation attached. Each iteration creates a new boundary.

Step 2. Test the new boundary.
The new boundary is more specific than the previous boundary and it is closer to the ideal boundary. One does not know how well it behaves on unseen examples, but the goal is to find a boundary that classifies these instances with high accuracy. Thus, the boundary is first tested on only 10% of the annotated examples (different from the 10% of the examples used for testing). If the accuracy is larger than the previous boundary's accuracy, the algorithm is converging toward the best approximation of the boundary and thus it repeats Step 2 for the new boundary. If the accuracy is lower than the previous boundary's accuracy, the new boundary is too specific and the previous boundary is a better approximation of the ideal boundary.

C. Lapata and Keller's Web-based unsupervised model
Lauer (1995) was the first to devise and test an unsupervised probabilistic model for noun–noun compound interpretation on *Grolier's Encyclopedia,* an eight million word corpus, based on a set of eight preposition paraphrases. His probabilistic model computes the probability of a preposition p given a noun–noun pair $n_1 - n_2$ and finds the most likely preposition paraphrase $p^* = argmax_p P(p|n_1, n_2)$. However, as Lauer noticed, this model requires a very large training corpus to estimate these probabilities. More recently, Lapata and Keller (2005) replicated the model using the Web as training corpus and showed that the best performance was obtained with the trigram model $f(n_1, p, n_2)$. In their approach, they used as the count for a given trigram the number of pages returned by using the trigram as a query. These co-occurrence frequencies were estimated using inflected queries which are obtained by expanding a noun–noun compound into all its morphological forms; then searching for N P N instances, for each of the eight prepositions P in Lauer's list. All queries are performed as exact matches using quotation marks. For example, for the test noun–noun compound instance *war stories*, all possible combinations of definite/indefinite articles and singular/plural noun forms are tried resulting in the queries *story about war*, *a/the story about war*, *story about a/the war*, *stories about war*, *stories about the wars*, *story about wars*, *story about the wars*, and so on. These forms are then submitted as literal queries, and the resulting hits are summed up. The query, and thus the preposition, with the largest number of hits is selected as the correct semantic interpretation category.

For the Europarl and CLUVI test sets, we replicated Lapata and Keller's (2005) experiments using Google.[15] We formed inflected queries with the patterns they proposed and searched the Web.

5. Experimental Results

We performed various experiments on both the Europarl and CLUVI testing corpora using seven sets of supervised models. Table 8 shows the results obtained against SS and Lapata and Keller's model on both corpora and the contribution of the features exemplified in seven versions of the SVM model. Supervised models 1 and 2 are defined only for the English features. Here, features F1 and F2 measure the contribution of the WordNet IS-A lexical hierarchy specialization. However, supervised model 1, which is also the baseline, does not differentiate between unambiguous and ambiguous training examples and thus does not specialize those that are ambiguous. These models show the difference between SS and SVM and the contribution of the other English features, such as preposition and nominalization (F1–F7).

The table shows that overall the performance is better for the Europarl corpus than for CLUVI. For the supervised models 1 and 2, SS [F1 + F2] gives better results than SVM [F1 + F2]. The inclusion of the other English features (SVM [F1–F7]) adds more than 10% accuracy (with a higher increase in Europarl) for the supervised model 1.

The results obtained are presented using the standard measure of *accuracy* (the number of correctly labeled instances over the number of instances in the test set).

5.1 The Contribution of Romance Linguistic Features

Our intuition is that the more information we use from other languages for the interpretation of an English instance, the better the results. Thus, we wanted to see the impact of each Romance language on the overall performance. Supervised model 3 shows the results obtained for English and the Romance language that contributed the least to the performance (English and Spanish for the entire English feature subset F1–F8). Here we computed the performance on all five English–Romance language combinations and chose the Romance language that provided the best result. Thus, supervised models 3 through 7 add Spanish, French, Portuguese, Italian, and Romanian in this order and show the contribution of each Romance preposition and all English features.

The language ranking in Table 8 shows that Romance languages considered here have a different contribution to the overall performance. Whereas the addition of Portuguese in CLUVI decreases the performance, in Europarl it increases it, if only by a few points. However, a closer analysis of the data shows that this is mostly due to the distribution of the corpus instances. For example, French, Italian, Spanish, and Portuguese are consistent in the choice of preposition (e.g., if the preposition *de* [*of*] is used in French, then the corresponding preposition is used in the other four language translations). A notable exception here is Romanian which provides two possible constructions with almost equal distribution: the N P N and the genitive-marked N N. The table shows (in the increase in performance between supervised models 6 and 7) that

15 As Google limits the number of queries to 1,000 per day per computer, we repeated the experiment using 10 computers for a number of days. Although Keller and Lapata used AltaVista for the interpretation of two noun–noun compounds, they showed that there is almost no difference between the correlations achieved using Google and AltaVista counts.

Table 8
The performance obtained by five versions of the cross-linguistic SVM model compared against the baseline, an English SVM model, and the SS model. The results obtained are presented using the standard measure of *accuracy* (number of correctly labeled instances over the number of instances in the test set).

		Results [%]			
		CLUVI		**Europarl**	
Learning models		**8-PP**	**22-SR**	**8-PP**	**22-SR**
Supervised model 1: Baseline	SS (F1+F2)	42.01	46.03	35.8	36.2
(English nominal features only)	SVM (F1+F2)	34.17	38.11	30.02	33.01
(no WordNet specialization)	SVM (F1-F7)	–	50.1	–	43.33
Supervised model 2	SS (F1+F2)	55.20	61.02	54.12	57.01
(English features)	SVM (F1+F2)	41.8	46.18	41.03	41.3
	SVM (F1-F7)	–	61.04	–	67.63
Supervised model 3	SVM (F1-F7+F8)	–	63.11	–	68.04
(English and Spanish features)					
Supervised model 4	SVM	–	65.81	–	69.58
(English, Spanish, and French features)	(F1-F7+F8+F9)				
Supervised model 5	SVM	–	64.31	–	69.92
(English, Spanish, French, and Portuguese features)	(F1-F7+F8+F9) (+F11)				
Supervised model 6	SVM	–	66.05	–	71.25
(English, Spanish, French, Portuguese, and Italian features)	(F1-F7+F8+F9+ F10+F11)				
Supervised model 7 (SVM)	–	**72.82**	–	**76.34**	
(English and all Romance features: F1–F13)					
Lapata and Keller's Web-based unsupervised model (English)	41.10	–	42.12	–	

this choice is not random, but influenced by the meaning of the instances (features F12, F13). This observation is also supported by the contribution of each feature to the overall performance. For example, in Europarl, the WordNet verb and nominalization features of the head noun (F3, F6) have a contribution of 5.12%, whereas for the modifier nouns they decrease by about 2.7%. The English preposition (F5) contributes 6.11% (Europarl) and 4.82% (CLUVI) to the overall performance.

The most frequently occurring preposition in both corpora is the underspecified preposition *de* (*of*), encoding almost all of the 22 semantic relations. The many-to-many mappings of the preposition to the semantic classes adds to the complexity of the interpretation task. A closer look at the Europarl and CLUVI data shows that Lauer's set of eight prepositions represents 88.2% (Europarl) and 91.8% (CLUVI) of the N P N instances. From these, the most frequent preposition is *of* with a coverage of 79% (Europarl) and 88% (CLUVI). Because the polysemy of this preposition is very high, we wanted to analyze its behavior on the set of most representative semantic relations in both corpora. Moreover, we wanted to see what prepositions were used to translate the English nominal phrase and compound instances in the target Romance languages, and thus to capture the semantic (ir)regularities among these languages in the two corpora and their contribution to the semantic interpretation task.

For most of the N P N instances, we noticed consistent behavior of the target Romance languages in terms of the prepositions used. This behavior can be classified

roughly in four categories exemplified subsequently: Example (12) shows a combination of the preposition *of/de* and more specific prepositions; Example (13) shows different prepositions than the one corresponding to the English equivalent in the instance; and Examples (14) and (15) show corresponding translations of the equivalent preposition in English in all Romance languages with variations in Romanian (e.g., *de* for *of*, *para/pour/par/pentru* for *for*).

(12) Committee *on* Culture (En.) – Comisión *de* la Cultura (Sp.) – commission *de* la culture (Fr.) – commissione *per* la cultura (It.) – Comissão *para* Cultura (Port.) – comitet *pentru* cultură (Ro.) (PURPOSE)

(13) the supervision *of* the administration (En.) – control *sobre* la administración (Sp.) – contrôle *sur* l'administration (Fr.) – controllo *sull'*amministrazione (It.) – controlo *sobre* administração (Port.) – controlul *asupra* administraţiei (Ro.) (THEME)

(14) lack *of* protection (En.) – falta *de* protección (Sp.) – manque *de* protection (Fr.) – mancanza *di* tutela (It.) – falta *de* protecção (Port.) – lipsă *de* protecţie (Ro.) (THEME)

(15) the cry *of* a man (En.) – el llanto *de* un hombre (Sp.) – un cri *d'*homme (Fr.) – l'urlo *di* un uomo (It.) – o choro *de* um bêbado (Port.) – strigătul unui om (Ro.) (AGENT)

Because the last three categories are the most frequent in both corpora, we analyzed their instances. Most of the time Spanish, French, Italian, and Portuguese make use of specific prepositions such as those in Examples (12) and (13) to encode some semantic relations such as PURPOSE and LOCATION, but rely on N *de* N constructions for almost all the other relations. English and Romanian, however, can choose between N N and N P N constructions. In the next section we present in more detail an analysis of the semantic correlations between English and Romanian nominal phrases and compounds and their role in the semantic interpretation task.

6. Linguistic Observations

In this section we present some linguistic observations derived from the analysis of the system's performance on the CLUVI and Europarl corpora. More specifically, we present different types of ambiguity that can occur in the interpretation of nominal phrases and compounds when using more abstract interpretation categories such as Lauer's eight prepositions. We also show that the choice of syntactic constructions in English and Romanian can help in the identification of the correct position of the semantic arguments in test instances.

6.1 Observations on Lapata and Keller's Unsupervised Model

In this section we show some of the limitations of the unsupervised probabilistic approaches that rely on more abstract interpretation categories, such as Lauer's set of eight prepositions. For this, we used Lapata and Keller's approach, a state-of-the-art knowledge-poor Web-based unsupervised probabilistic model which provided a performance of 42.12% on Europarl and 41.10% on CLUVI. We manually checked the first

Table 9
Experimental results with Lapata and Keller's Web-based unsupervised interpretation model on different types of test sets from the Europarl corpus.

Noun–noun compound test set	Ambiguity of noun constituents	Accuracy [%]
Set#1	one part of speech, one WordNet sense	35.28%
Set#2	multiple parts of speech, one WordNet sense	31.22%
Set#3	one part of speech, multiple WordNet senses	50.63%
Set#4	multiple parts of speech, multiple WordNet senses	43.25%

five entries of the pages returned by Google for each most frequent N P N paraphrase for 100 CLUVI and Europarl instances and noticed that about 35% of them were wrong due to syntactic (e.g., part of speech) and/or semantic ambiguities. For example, *baby cry* generated instances such as "it will make moms *cry with the baby*," where *cry* is a verb, not a noun. This shows that many of the NP instances selected by Google as matching the N P N query are incorrect, and thus the number of hits returned for the query is overestimated. Thus, because we wanted to measure the impact of various types of noun–noun compound ambiguities on the interpretation performance, we further tested the probabilistic Web-based model on four distinct test sets selected from Europarl, each containing 30 noun–noun compounds encoding different types of ambiguity: In Set#1 the noun constituents had only one part of speech and one WordNet sense; in Set#2 the nouns had at least two possible parts of speech and were semantically unambiguous; in Set#3 the nouns were ambiguous only semantically; and in Set#4 they were ambiguous both syntactically and semantically. Table 9 shows that for Set#1, the model obtained an accuracy of 35.28%, while for more semantically ambiguous compounds it obtained an average accuracy of about 48% (50.63% [Set#3] and 43.25% [Set#4]. This shows that for more syntactically ambiguous instances, the Web-based probabilistic model introduces a significant number of false positives, thus decreasing the accuracy (cf. sets #1 vs. #2 and #3 vs. #4).

Moreover, further analyses of the results obtained with Lapata and Keller's model showed that about 30% of the noun–noun compounds in sets #3 and #4 were ambiguous with at least two possible readings. For example, *paper bag* can be interpreted out-of-context both as *bag of paper* (bag made of paper—STUFF–OBJECT, a subtype of PART–WHOLE) and as *bag for papers* (bag used for storing papers—PURPOSE). Similarly, *gingerbread bowl* can be correctly paraphrased both as *bowl of/with gingerbread* (CONTENT–CONTAINER) and as *bowl of gingerbread* (bowl made of gingerbread—STUFF–OBJECT). The following two examples show the two readings of the noun–noun compound *gingerbread bowl* as found on Google:

(16) Stir a *bowl of gingerbread*,
Smooth and spicy and brown,
Roll it with a rolling pin,
Up and up and down,
...[16]

16 An excerpt from the "Gingerbread Man" song.

(17) The gingerbread will take the shape of the glass bowl. Let it cool for a few minutes and then carefully loosen the foil and remove the gingerbread from the glass. And voilà: your *bowl of gingerbread*.

These ambiguities partially explain why the accuracy values obtained for sets #3 and #4 are higher then the ones obtained for the other two sets. The semantic ambiguity also explains why the accuracy obtained for set #2 is higher than that for set #4. For these sets of examples the syntactic ambiguity affected the accuracy much less than the semantic ambiguity (that is, more *N P N* combinations were possible due to various noun senses). This shows one more time that a large number of noun–noun compounds are covered by more abstract categories, such as prepositions. Moreover, these categories also allow for a large variation as to which category a compound should be assigned.

6.2 Observations on the Symmetry of Semantic Relations: A Study on English and Romanian

Nominal phrases and compounds in English, nominal phrases in the Romance languages considered here, and genitive-marked noun–noun compounds in Romanian have an inherent directionality imposed by their fixed syntactic structure. For example, in English noun–noun compounds the syntactic head always follows the syntactic modifier, whereas in English and Romance nominal phrases the order is reversed. Two such examples are *tea*/Modifier *cup*/Head and *glass*/Head *of wine*/Modifier.

The directionality of semantic relations (i.e., the order of the semantic arguments) however, is not fixed and thus it is not always the same as the inherent directionality imposed by the syntactic structure. Two such examples are *ham*/Modifier/Arg_2 *sandwich*/Head/Arg_1 and *spoon*/Modifier/Arg_1 *handle*/Head/Arg_2. Although both instances encode a PART–WHOLE relation (Arg_1 is the semantic argument identifying the whole and Arg_2 is the semantic argument identifying the part), their semantic arguments are not listed in the same order (Arg_1 Arg_2 for *spoon handle* and Arg_2 Arg_1 for *ham sandwich*). For a better understanding of this phenomenon, we performed a more thorough analysis of the training instances in both CLUVI and Europarl. Because the choice of syntactic constructions in context is governed in part by semantic factors, we focused on English and Romanian because they are the only languages from the set considered here with two productive syntactic options: N N and N P N (English) and genitive-marked N N and N P N (Romanian). Thus, we grouped the English–Romanian parallel instances per each semantic relation and each syntactic construction and checked if the relation was symmetric or not, according to the following definition.

Definition
We say that **a semantic relation is symmetric** relative to a particular syntactic construction if there is at least one relation instance whose arguments are in a different order than the order indicated by the relation's default argument frame for that construction.

For example, PART–WHOLE is symmetric with regard to nominal phrases because the semantic arguments of the instance *the building*/*Arg*1 *with parapets*/*Arg*2 are in a different order than the one imposed by the relation's default argument frame (*Arg*2 P *Arg*1) for nominal phrases (cf. Table 1).

Because the relation distribution is skewed in both corpora, we focused only on those relations encoded by at least 50 instances in both Europarl and CLUVI. For example, in English the POSSESSION relation is symmetric when encoded by N P N and noun–noun compounds. For instance, we can say *the girl with three dogs* and *the resources of the Union*, but also *family estate* and *land proprietor*. The findings are summarized and presented in Table 10 along with examples. Some relations such as IS-A, PURPOSE, and MEASURE cannot be encoded by genitive-marked noun–noun compounds in Romanian (indicated by "–" in the table). A checkmark symbol indicates if the relation is symmetric ('✓') or not ('x') for a particular syntactic construction. It is interesting to note that not all the relations are symmetric and this behavior varies from one syntactic construction to another and from one language to another. Although some relations such as AGENT and THEME are not symmetric, others such as TEMPORAL, PART–WHOLE, and LOCATION are symmetric irrespective of the syntactic construction used.

Symmetric relations pose important challenges to the automatic interpretation of nominal phrases and compounds because the system has to know which of the nouns is the semantic modifier and which is the semantic head. In this research, the order of the semantic arguments has been manually identified and marked in the training corpora. However, this information is not provided for unseen test instances. So far, in our experiments with the test data the system used the order indicated by the default argument frames. Another solution is to build argument frames for clusters of prepositions which impose a particular order of the arguments in N P N constructions. For example, in the N_2 P N_1 phrases *the books on the table* (LOCATION) and *relaxation during the summer* (TEMPORAL), the semantic content of the prepositions *on* and *during* identifies the position of the physical and temporal location (e.g., that N_1 is the time or location). This approach works most of the time for relations such as LOCATION and TEMPORAL because in both English and Romance languages they rely mostly on prepositions indicating location and time and less on underspecified prepositions such as *of* or *de*. However, a closer look at these relations shows that some of the noun–noun pairs that encode them are not symmetric and this is true for both English and Romance. For instance, *cut on the chin* and *house in the city* cannot be reversed as *chin* P *cut* or *city* P *house*. One notable exception here is indicated by examples such as *box of/with matches – matches in/inside the box* and *vessels of/with blood – blood in vessels*[17] encoding CONTENT–CONTAINER. Another special case is when P_1 and P_2 are location antonyms (e.g., *the book under the folder* and *the folder on the book*). However, even here symmetry is not always possible, being influenced by pragmatic factors (Herskovits 1987) (e.g., we can say *the vase on the table*, but not *the table under the vase*—this has to do with the difference in size of the objects indicated by the head and modifier nouns. Thus, a larger object cannot be said to be placed *under* a smaller one).

It is important to stress here the fact that our definition of symmetry of semantic relations does not focus in particular on the symmetry of an instance noun–noun pair that encodes the relation, although it doesn't exclude such a case. We call this *lexical symmetry* and define it here.

Definition
We say that **a noun–noun pair (N_1 – N_2) is symmetric** relative to a particular syntactic construction and the semantic relation it encodes in that construction if the order of the nouns in the construction can be changed provided the semantic relation is preserved.

17 Here the noun *vessels* refers to a type of container.

Table 10
A summary of the symmetry properties of a set of the 12 most frequent semantic relations in CLUVI and Europarl. "–" means the semantic relation is not encoded by the syntactic construction, "✓" and "x" symbols indicate whether the relation is or is not symmetric.

		Symmetry				
		English		Romanian		
No.	Semantic relations	N N	N P N	genitive-marked N N	N P N	Examples
1	POSSESSION	✓	✓	✓	x	En.: *family#2/Arg$_1$ estate#2/Arg$_2$* vs. *land#1/Arg$_2$ proprietor#1/Arg$_1$* Ro.: *terenul/Arg$_2$ proprietarului/Arg$_1$* (land-the owner-GEN) ('the owner's land') *proprietarul/Arg$_1$ magazinului/Arg$_2$* (owner-the store-GEN) ('the owner of the store')
2	PROPERTY	x	✓	x	✓	En.: *calm#1/Arg$_2$ of evening#1/Arg$_1$* vs. *spots#4/Arg$_1$ of color#1/Arg$_2$* Ro.: *pete/Arg$_1$ de culoare/Arg$_2$* ('spots of color') *miros/Arg$_2$ de camfor/Arg$_1$* ('odour of camphor')
3	AGENT	x	x	x	x	*En.: the investigation#2/Arg$_2$ of the police#1/Arg$_1$* Ro.: *investigaţia/Arg$_2$ poliţiei/Arg$_1$* (investigation-the police-GEN)
4	TEMPORAL	✓	✓	✓	✓	En.: *news#3/Arg$_2$ in the morning#1/Arg$_1$* vs. *the evening#1/Arg$_1$ of her arrival#2/Arg$_2$* Ro.: *placinte/Arg$_2$ de dimineaţă/Arg$_1$* (cakes of morning) ('morning cakes') vs. *ani/Arg$_1$ de subjugare/Arg$_2$* ('years of subjugation')
5	PART–WHOLE	✓	✓	✓	✓	En: *faces#1/Arg$_2$ of children#1/Arg$_1$* vs. *the shell#5Arg$_2$ of the egg#2/Arg$_1$* Ro: *feţele/Arg$_2$ copiilor/Arg$_1$* (faces-the children-GEN) ('the faces of the children') vs. *coajă/Arg$_1$ de ou/Arg$_2$* (shell of egg) ('egg shell')
6	HYPERNYMY (is-a)	x	x	–	x	En.: *daisy#1/Arg$_1$ flower#1/Arg$_2$* Ro.: *meci/Arg$_2$ de fotbal/Arg$_1$* (match of football) ('football match')
7	LOCATION	✓	✓	✓	✓	En.: *castle#2/Arg$_2$ in the desert#1/Arg$_1$* vs. *point#2/Arg$_1$ of arrival#1/Arg$_2$* Ro.: *castel/Arg$_2$ in deşert/Arg$_1$* (castle in desert) ('castle in the desert') vs. *punct/Arg$_1$ de sosire/Arg$_2$* ('point of arrival')
8	PURPOSE	x	x	–	x	En.: *war#1/Arg$_1$ canoe#1/Arg$_2$* Ro.: *pirogă/Arg$_2$ de război/Arg$_1$* (canoe of war)
9	TOPIC	x	x	x	x	En.: *war#1/Arg$_1$ movie#1/Arg$_2$* Ro.: *film/Arg$_2$ despre război/Arg$_1$* ('movie about war')
10	MEASURE	–	x	–	x	En.: *inches#1/Arg$_2$ of snow#2/Arg$_1$* Ro.: *inci/Arg$_2$ de zapadă/Arg$_1$* (inches of snow)
11	TYPE	x	✓	x	x	En.: *framework#1/Arg$_1$ law#2/Arg$_2$* Ro.: *lege/Arg$_2$ cadru/Arg$_1$* (law framework)
12	THEME	x	x	x	x	En.: *examination#1/Arg$_2$ of machinery#1/Arg$_1$* Ro.: *verificarea/Arg$_2$ maşinii/Arg$_1$* (examination-the machinery-GEN) ('the examination of the machinery')

For instance, the pair *building–parapets* in the nominal phrases *the building/Arg_1 with parapets/Arg_2* and *the parapets/Arg_2 of the building/Arg_1* encodes a PART–WHOLE relation. Here, both the noun–noun pair and the semantic relation are symmetric relative to N P N. However, the situation is different for instances such as *the book/Arg_2 under the folder/Arg_1* and *the folder/Arg_2 on the book/Arg_1*, both encoding LOCATION. Here, the *book–folder* pair is symmetric in N P N constructions (in the first instance the *book* is the syntactic head and the *folder* is the modifier, whereas in the second instance the order is reversed). However, the LOCATION relation they encode is not symmetric (in both instances, the order of the semantic arguments matches the default argument frame for LOCATION). It is interesting to notice here that these two location instances are actually paraphrases of one another. This can be explained by the fact that both the *book* and the *folder* can act as a location with respect to the other, and that the prepositions *under* and *on* are location antonyms. In comparison, *the building with parapets* is not a paraphrase of *the parapets of the building*. Here, the nouns *building* and *parapets* cannot act as a whole/part with respect to each other (e.g., the only possible whole here is the noun *building*, and the only possible part here is the noun *parapets*). This is because *parts* and *wholes* have an inherent semantic directionality imposed by the inclusion operation on the set of things representing parts and wholes, respectively.

In this research we consider the identification and extraction of semantic relations in nominal phrases and compounds, but we do not focus in particular on the acquisition of paraphrases in these constructions. Our goal is to build an accurate semantic parser which will automatically annotate instances of nominal phrases and compounds with semantic relations in context. This approach promises to be very useful in applications that require semantic inference, such as textual entailment (Tatu and Moldovan 2005). However, a thorough analysis of the semantics of nominal phrases and compounds should focus on both semantic relations and paraphrases. We leave this topic for future research.

Because we wanted to study in more detail the directionality of semantic relations, we focused on PART–WHOLE. These relations, and most of the semantic relations considered here, are encoded mostly by N *of/de* N constructions, genitive-marked N N (Romanian), and noun–noun compounds (English) and thus, the task of argument order identification becomes more challenging. For the purpose of this research we decided to take a closer look at the PART–WHOLE relation in both CLUVI and Europarl where together it accounted for 920 token and 636 type instances. We show subsequently a detailed analysis of the symmetry property on a classification of PART–WHOLE relations starting with a set of five PART–WHOLE subtypes identified by Winston, Chaffin, and Hermann (1987):[18] (1) *Component–Integral object*, (2) *Member–Collection*, (3) *Portion–Mass*, (4) *Stuff–Object*, and (5) *Place–Area*.

(1) Component–Integral object

This is a relation between components and the objects to which they belong. Integral objects have a structure with their components being separable and having a functional relation with their wholes. This type of PART–WHOLE relation can be encoded by N *of* N and less often by N N constructions. Moreover, here the existential interpretation is preferred over the generic one. Such examples are *the leg of the table* and *the table leg* which translate in Romanian as *piciorul mesei* ('leg-the table-GEN'). In Romanian a

18 Winston, Chaffin, and Hermann (1987) identified six subtypes of PART–WHOLE relations, one of which, (*Feature–Activity*), is not presented here because it is not frequently encoded by N N and N P N constructions.

generic interpretation is also possible, but with change of construction and most of the time of semantic relation (e.g., *picior de masă* – 'leg of table' encoding PURPOSE[19]).

This relation subtype is symmetric in English for both N N and N P N constructions. In Romanian, however, it is symmetric only when encoded by N P N. Moreover, it is interesting to note that Modifier/Arg_1 Head/Arg_2 noun–noun compound instances translate as genitive noun–noun compounds in Romanian, whereas Modifier/Arg_2 Head/Arg_1 instances translate as N P N, with P different from *of*. For example, *chair*/Arg_1 *arm*/Arg_2 and *ham*/Arg_2 *sandwich*/Arg_1 translate in Romanian as Head/Arg_2 Modifier/Arg_1 – *braţul scaunului* ('arm-the chair-GEN') and Head/Arg_1 P Modifier/Arg_2 – *sandwich cu şuncă* ('sandwich with ham').

For N P N instances in Romanian and English both Arg_1 P_1 Arg_2 and Arg_2 P_2 Arg_1 argument orderings are possible, but with a different choice of preposition (with P_1 different from *of/de*). For example, one can say *the parapets*/Arg_2 *of the building*/Arg_1, but also *the building*/Arg_1 *with parapets*/Arg_2. A closer look at such instances shows that symmetry is possible when the modifier (in this case the part) is not a mandatory part of the whole, but an optional part with special features (e.g., color, shape). For example, *the car with the door* is less preferred than *the car with the red door* which differentiates the car from other types of cars.

(2) Stuff–Object

This category encodes the relations between an object and the material of which it is partly or entirely made. The parts are not similar to the wholes which they compose, cannot be separated from the whole, and have no functional role. The relation can be encoded by both N *of* N and N N English and Romanian patterns and the choice between existential and generic interpretations correlates with the relation symmetry. For N N constructions this relation subtype is not symmetric, while for N P N it is symmetric only in English. Such examples are *brush*/Arg_2 *hut*/Arg_1 in English, and *metalul*/Arg_2 *scaunului*/Arg_1 ('metal-the seat-GEN' – *the metal of the seat*) and *scaun de metal* ('chair of metal' – *metal chair*) in Romanian.

N P N instances can only be encoded by *of* in English or *de/din* (of/from) in Romanian. If the position of the arguments is Arg_1 *of* Arg_2 and Arg_2 is an indefinite noun indicating the part then the instance interpretation is generic. For example, *seat of metal* translates as *scaun de/din metal* ('chair of/from metal') in Romanian. It is important to note here the possible choice of the preposition *from* in Romanian, a preposition which is rarely used in English for this type of relation.

When the position of the arguments changes (e.g., Arg_2 *of* Arg_1), the same preposition *of* is used and the semantic relation is still STUFF–OBJECT, but the instance is more specific having an existential interpretation. For instance, *the metal of the seat* translates in Romanian as *metalul scaunului* ('metal-the seat-GEN') and not as *metalul de scaun* ('metal-the of seat').

(3) Portion–Mass

According to Selkirk (1982a), Ionin, Matushansky, and Ruys (2006), and our own observations on the CLUVI and Europarl data, this type of PART–WHOLE relation can be further classified into **mass, measure,** and **fraction partitives.** Here the parts are separable and similar to each other and to the whole they are part of. An example of a mass partitive is *half*/Arg_2 *of the cake*/Arg_1 which translates in Romanian as *jumătate*/Arg_2

19 This reading is possible if the *leg* is separated from the *table*.

de/din prajitură/Arg_1 ('half of/from cake'). Note that here the noun *cake* is indefinite in Romanian, and thus the instance is generic. An existential interpretation is possible when the noun is modified by a possessive (e.g., *half of your cake*).

Measure partitives are also called vague PART–WHOLE relations (Selkirk 1982b) because they can express both PART–WHOLE and MEASURE depending on the context. They are encoded by N_1 *of* N_2 constructions, where N_2 is indefinite, and can indicate both existential and generic interpretations. Two such examples are *bottles/Arg_1 of wine/Arg_2* and *cup/Arg_1 of sugar/Arg_2*. In Romanian, the preposition used is either *de* (*of*), or *cu* (*with*). For example, *sticle/Arg_1 de/cu vin/Arg_2* ('bottles of/with wine') and *ceaşcă/Arg_1 de/cu zahăr/Arg_2* ('cup of/with sugar').

Fraction partitives indicate fractions of wholes, such as *three quarters/Arg_2 of a pie/Arg_1* (*trei pătrimi/Arg_2 de plăcintă/Arg_1* [Romanian]–['three quarters of pie']) and *one third/Arg_2 of the nation/Arg_1* (*o treime/Arg_2 din populaţia/Arg_1* [Romanian]–['one third from population-the'] and not *o treime de populaţia* – ['a third of population-the']). Here again, we notice the choice of the Romanian preposition *din* and not *de* when the second noun is definite. The preposition *from* indicates the idea of separation of the part from the whole, an idea which characterizes PART–WHOLE relations.

Portion–Mass relations cannot be encoded by N N structures in either English or Romanian and they are not symmetric in N P N constructions.

(4) Member–Collection

This subtype represents membership in a collection. Members are parts, but may not play any functional role with respect to their whole. That is, compared with Component–Integral instances such as *the knob of the door*, where the knob is a round handle one turns in order to open a door, in an example like *bunch of cats*, the *cats* don't play any functional role to the whole *bunch*.

This subtype can be further classified into a **basic** subtype (e.g., *the member of the team*), **count partitives** (e.g., *two of these people*), **fraction count partitives** (e.g., *two out of three workers*), and **vague measure partitives** (e.g., *a number/lot/bunch of cats*). Although the basic Member–Collection partitives are symmetric for N N (Romanian only) and N P N (English and Romanian), the other subtypes can be encoded only by N P N constructions and are not symmetric in English or in Romanian. For example, *the children/Arg_2 of the group/Arg_1* and *children/Arg_2 group/Arg_1* translate as *copiii/Arg_2 din grup/Arg_1* ('children-the from group') and as *grup/Arg_1 de copii/Arg_2* ('group of children').

The second and the third subtypes translate in Romanian as *doi/Arg_2 din aceşti oameni/Arg_1* ('two from these people') and *doi/Arg_2 din trei lucrători/Arg_1* ('two from three workers'), by always using the preposition *din* (*from*) instead of *de* (*of*). On the other hand, vague measure partitives translate as *un număr/Arg_1 de pisici/Arg_2* ('a number of cats') and not as *un număr din pisici* ('a number from cats'). Although all these subtypes need to have a plural modifier noun and are not symmetric, count partitives always have an existential interpretation, whereas fraction count and vague measure partitives have a generic meaning.

(5) Location–Area

This subtype captures the relation between areas and special places and locations within them. The parts are similar to their wholes, but they are not separable from them. Thus, this relation overlaps with the LOCATION relation. One such example is *the surface/Arg_2 of the water/Arg_1*. Both nouns can be either definite or indefinite and the relation is not symmetric when the part is a relational noun (e.g., *surface*, *end*). In Romanian,

both N *de* N and genitive-marked N N constructions are possible: *suprafaţa/*Arg_2 *apei/*Arg_1 ('surface-the water-GEN') and *suprafaţă/*Arg_2 *de apă/*Arg_1 ('surface of water'). The relation is symmetric only for N P N in both English and Romanian.

Table 11 summarizes the symmetry properties of all five PART–WHOLE subtypes accompanied by examples.

Thus, features such as the semantic classes of the two nouns (F1, F2), and the syntactic constructions in English and Romanian—more specifically, the preposition features for English (F5) and Romanian (F12) and the inflection feature for Romanian (F13)—can be used to train a classifier for the identification of the argument order in nominal phrases and compounds encoding different subtypes of PART–WHOLE relations. For example, the argument order for *Portion–Mass* instances can be easily identified if it is determined that they are encoded by N_2 *of/de* N_1 in English and Romanian and the head noun N_2 is identified as a *fraction* in the WordNet IS-A hierarchy, thus representing Arg_2 (the part). It is interesting to note here that all the other *Member–Collection* subtypes with the exception of the basic one are also encoded only by N *of/de* N, but here the order is reversed in both English and Romanian (N_1 *of/de* N_2), where the head noun N_1, if identified as a *collection* concept in WordNet, represents the whole concept (Arg_1).

This approach can also be applied to other symmetric relations by classifying them into more specific subtypes for argument order identification. Thus, local classifiers can be trained for each subtype on features such as those mentioned herein and tested on unseen instances. However, for training this procedure requires a sufficiently large number of examples for each subtype of the semantic relation considered.

Table 11
A summary of the symmetry properties of the five subtypes of PART–WHOLE semantic relation in CLUVI and Europarl. "–" means the semantic relation is not encoded by the syntactic construction, "✓" and "x" symbols indicate whether the relation is or is not symmetric.

		Symmetry				
		English		Romanian		
				genitive-marked		
No.	Semantic relations	N N	N P N	N N	N P N	Examples
1	**Component – Integral obj.**	✓	✓	x Arg_2 Arg_1	✓	En.: *chair#1/*Arg_1 *arm#5/*Arg_2 vs. *ham#1/*Arg_2 *sandwich#1/*Arg_1 Ro: *'braţul/*Arg_2 *scaunului/*Arg_1' (arm-the chair-GEN) vs. *'sandwich/*Arg_1 *cu şuncă/*Arg_2' (sandwich with ham)
2	**Stuff – Object**	x Arg_2 Arg_1	✓	x Arg_2 Arg_1	x Arg_1 de Arg_2	En.: *dress#1/*Arg_1 *of silk#1/*Arg_2 vs. *the silk#1/*Arg_2 *of the dress#1/*Arg_1 Ro.: *'rochie/*Arg_1 *de mătase/*Arg_2' (dress of silk) vs. *'mătasea/*Arg_2 *rochiei/*Arg_1' (silk-the dress-GEN)
3	**Portion – Mass**	–	x Arg_2 of Arg_1	–	x Arg_2 de Arg_1	En.: *half#1/*Arg_2 *of the cake#3/*Arg_1 vs. Ro: *'jumătate/*Arg_2 *de/din prajitură/*Arg_1' (half of/from cake)
4	**Member – Collection** (count, fraction count, and vague measure partitives)	–	x Arg_1 of Arg_2	–	x Arg_1 de Arg_2	En.: *a bunch#1/*Arg_1 *of cats#1/*Arg_2 Ro.: *'o grămadă/*Arg_1 *de pisici/*Arg_2', (a bunch of cats)
	Member – Collection (basic partitive)	x Arg_1 Arg_2	✓	✓	✓	En.: *president#4/*Arg_2 *of the committee#1/*Arg_1 vs. *committee#1/*Arg_1 *of idiots#1/*Arg_2 Ro.: *copiii/*Arg_2 *din grup/*Arg_1 (children-the from group) ('the children from the group') *grup/*Arg_1 *de copii/*Arg_2 ('group of children')
5	**Location – Area**	x Arg_1 Arg_2	✓	x Arg_2 Arg_1	✓	En.: *the swamps#1/*Arg_2 *of the land#7/*Arg_1 vs. *the land#7/*Arg_1 *with swamps#1/*Arg_2 Ro.: *oază în deşert* (oasis in desert) vs. *deşert cu oază în* (desert with oasis)

This analysis shows that the choice of lexico-syntactic structures in both English and Romanian correlates with the meaning of the instances encoded by such structures. In the next section we present a list of errors and situations that, currently, our system fails to recognize, and suggest possible improvements.

7. Error Analysis and Suggested Improvements

A major part of the difficulty of interpreting nominal phrases and compounds stems from multiple sources of ambiguity. These factors range from syntactic analysis, to semantic, pragmatic, and contextual information and translation issues. In this section we show various sources of error we found in our experiments and present some possible improvements.

A. Error analysis
Two basic factors are wrong part-of-speech and word sense disambiguation tags. Thus, if the syntactic tagger and WSD system fail to annotate the nouns with the correct senses, the system can generate wrong semantic classes which will lead to wrong conclusions. Moreover, there were also instances for which the nouns or the corresponding senses of these nouns were not found in WordNet. There were 42.21% WSD and 6.7% POS tagging errors in Europarl and 54.8% and 7.32% in CLUVI. Additionally, 6.9% (Europarl) and 4.6% (CLUVI) instances had missing senses.

There are also cases when local contextual information such as word sense disambiguation is not enough for relation detection and when access to a larger discourse context is needed. Various researchers (Spärck Jones 1983; Lascarides and Copestake 1998; Lapata 2002) have shown that the interpretation of noun–noun compounds, for example, may be influenced by discourse and pragmatic knowledge. This context may be identified at the level of local nominal phrases and compounds or sentences or at the document and even collection level. For example, a noun–noun compound modified by a relative clause might be disambiguated in the context of another argument of the same verb in the clause, which can limit the number of possible semantic relations. For instance, the interpretation of the instance *museum book* in the subject position in the following examples is influenced by another argument of the verbs *bought* in Example (18), and *informed* in Example (19):

(18) *the* [*museum book*]$_{TOPIC}$ *John bought in the bookshop at the museum*

(19) *the* [*museum book*]$_{LOCATION}$ *that informed John about the ancient art*

Prepositions such as spatial ones are also amenable to visual interpretations due to their usage in various visual contexts. For example, the instance *nails in the box* (cf. Herskovits 1987) indicates two possible arrangements of the nails: either held by the box, or hammered into it. We cannot capture these subtleties with the current procedure even if they are mentioned in the context of the sentence or discourse.

B. Suggested improvements
In this article we investigated the contribution of English and Romance prepositions to the task of interpreting nominal phrases and compounds, both as features employed in a learning model and as classification categories. An interesting extension of this approach would be to look into more detail at the functional–semantic aspect of these prepositions and to define various tests that would classify them as pure functional components with no semantic content or semantic devices with their own meaning.

Moreover, our experiments focused on the detection of semantic relations encoded by N N and N P N patterns. A more general approach would extend the investigation to adjective–noun constructions in English and Romance languages as well.

Another direction for future work is the study of the semantic (ir)regularities among English and Romance nominal phrases and compounds in both directions. Such an analysis might be also useful for machine translation, especially when translating into a language with multiple choices of syntactic constructions. One such example is *tarro de cerveza* ('glass of beer') in Spanish which can be translated as either *glass of beer* (MEASURE) or *beer glass* (PURPOSE) in English. The current machine translation language models do not differentiate between such options, choosing the most frequent instance in a large training corpus.

The drawback of the approach presented in this article, as for other very precise learning methods, is the need for a large number of training examples. If a certain class of negative or positive examples is not seen in the training data (and therefore it is not captured by the classification rules), the system cannot classify its instances. Thus, the larger and more diverse the training data, the better the classification rules. Moreover, each cross-linguistic study requires translated data, which is not easy to obtain in electronic form, especially for most of the world's languages. However, more and more parallel corpora in various languages are expected to be forthcoming.

8. Discussion and Conclusions

In this article we investigated the contribution of English and Romance prepositions to the interpretation of N N and N P N instances and presented a supervised, knowledge-intensive interpretation model.

Our approach to the interpretation of nominal phrases and compounds is novel in several ways. We investigated the problem based on cross-linguistic evidence from a set of six languages: English, Spanish, Italian, French, Portuguese, and Romanian. Thus, we presented empirical observations on the distribution of nominal phrases and compounds and the distribution of their meanings on two different corpora, based on two state-of-the-art classification tag sets: Lauer's set of eight prepositions (Lauer 1995) and our list of 22 semantic relations. A mapping between the two tag sets was also provided. A supervised learning model employing various linguistic features was successfully compared against two state-of-the-art models reported in the literature.

It is also important to mention here the linguistic implications of this work. We hope that the corpus investigations presented in this article provide new insight for the machine translation and multilingual question answering communities. The translation of nominal phrase and compound instances from one language to another is highly correlated with the structure of each language, or set of languages. In this article we measured the contribution of a set of five Romance languages to the task of semantic interpretation of English nominal phrases and compounds. More specifically, we showed that the Romanian linguistic features contribute more substantially to the overall performance than the features obtained for the other Romance languages. The choice of the Romanian linguistic constructions (either N N or N P N) is highly correlated with their meaning. This distinct behavior of Romanian constructions is also explained by the Slavic and Balkanic influences. An interesting future research direction would be to consider other Indo- and non Indo-European languages and measure their contribution to the task of interpreting nominal phrases and compounds in particular, and noun phrases in general.

Acknowledgments

We would like to thank all our annotators without whom this research would not have been possible: Silvia Kunitz (Italian) and Florence Mathieu-Conner (French). We also thank Richard Sproat, Tania Ionin, and Brian Drexler for their suggestions on various versions of the article. And last but not least, we also would like to thank the reviewers for their very useful comments.

References

Alexiadou, Artemis, Liliane Haegeman, and Melita Stavrou. 2007. *Noun Phrases in the Generative Perspective*. Mouton de Gruyter, Berlin.

Almela, Ramón, Pascual Cantos, Aquilino Sánchez, Ramón Sarmiento, and Moisés Almela. 2005. *Frequencias del Español. Dicctionario de estudios léxicos y morfológicos*. Ed. Universitas, Madrid.

Anderson, Mona. 1983. Prenominal genitive NPs. *The Linguistic Review*, 3:1–24.

Artstein, Ron. 2007. *Quality Control of Corpus Annotation Through Reliability Measures*. Association for Computational Linguistics Conference (ACL), Prague, Czech Republic.

Baker, Collin, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 86–90, Montreal.

Baldwin, Timothy. 2005. Distributional similarity and collocational prepositional phrases. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, pages 197–210, Kluwer, Dordrecht.

Baldwin, Timothy. 2006a. Automatic identification of English verb particle constructions using linguistic features. In *Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72, Trento, Italy.

Baldwin, Timothy. 2006b. Representing and Modeling the Lexical Semantics of English Verb Particle Constructions. In *The European Association for Computational Linguistics (EACL), the ACL-SIGSEM Workshop on Prepositions*, Trento.

Barker, Chris. 1998. Partitives, double genitives and anti-uniqueness. *Natural Language and Linguistic Theory*, 16:679–717.

Busa, Federica and Michael Johnston. 1996. Cross-linguistic semantics for complex nominals in the generative lexicon. In *AISB Workshop on Multilinguality in the Lexicon*, Sussex.

Cadiot, Piere. 1997. *Les prépositions abstraites en français*. Armand Colin, Paris.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *The International Conference on Language Resources and Evaluation LREC*, pages 1934–1940, Las Palmas.

Casadei, Federica. 1991. Le locuzioni preposizionali. Struttura lessicale e gradi di lessicalizzazione. *Lingua e Stile*, XXXVI: 43–80.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book*, 2nd edition. Heinle and Heinle, Boston, MA.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *The 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 132–139, Seattle, WA.

Cornilescu, Alexandra. 2001. Romanian nominalizations: Case and aspectual structure. *Journal of Linguistics*, 37:467–501.

Dorr, Bonnie. 1993. *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge, MA.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language*, 53:810–842.

Evans, Vyvyan and Paul Chilton, editors. 2009. *Language, Cognition and Space: The State of the Art and New Directions*. Advances in Cognitive Linguistics. Equinox Publishing Company, London.

Fang, Alex C. 2000. A lexicalist approach towards the automatic determination for the syntactic functions of prepositional phrases. *Natural Language Engineering*, 6:183–201.

Fellbaum, Christiane. 1998. *WordNet—An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Finin, Timothy W. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Giorgi, Alessandra and Giuseppe Longobardi. 1991. *The syntax of noun phrases*. Cambridge University Press, London.

Girju, Roxana, Alexandra Badulescu, and Dan Moldovan. 2006. Automatic discovery

of part-whole relations. *Computational Linguistics*, 32(1):83–135.
Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the Semantics of Noun Compounds. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):479–496.
Gleitman, Lila R. and Henry Gleitman. 1970. *Phrase and Paraphrase: Some Innovative Uses of Language*. Norton, New York.
Gocsik, Karen. 2004. *English as a Second Language*. Dartmouth College Press, Hanover, NH.
Grimshaw, Jane. 1990. *Argument Structure*. MIT Press, Cambridge, MA.
Herskovits, Annette. 1987. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge, MA.
Ionin, Tania, Ora Matushansky, and Eddy Ruys. 2006. Parts of speech: Toward a unified semantics for partitives. In *Conference of the North East Linguistic Society (NELS)*, pages 357–370, Amherst, MA.
Jensen, Per Anker and Jőrgen F. Nilsson. 2005. Ontology-based semantics for prepositions. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht.
Jespersen, Otto. 1954. *A Modern English Grammar on Historical Principles*. George Allen & Unwin Ltd., Heidelberg and London.
Johnston, Michael and Federica Busa. 1996. Qualia structure and the compositional interpretation of compounds. In Evelyne Viegas, editor, *Breadth and Depth of Semantics Lexicon*, pages 77–88, Kluwer Academic, Dordrecht.
Kim, Su Nam and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *The International Joint Conference on Natural Language Processing (IJCNLP)*, pages 945–956, Jeju.
Kim, Su Nam and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *The International Conference on Computational Linguistics / the Association for Computational Linguistics (COLING/ACL) - Main Conference Poster Sessions*, pages 491–498, Sydney.
Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *The National Conference on Artificial Intelligence (AAAI)*, pages 691–696, Austin, TX.
Kordoni, Valia. 2005. Prepositional arguments in a multilingual context. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht, pages 307–330.
Kordoni, Valia. 2006. PPs as verbal arguments: From a computational semantics perspective. In *The European Association for Computational Linguistics (EACL), the ACL-SIGSEM Workshop on Prepositions*, Trento, Italy.
Lapata, Mirella. 2002. The Disambiguation of nominalisations. *Computational Linguistics*, 28(3):357–388.
Lapata, Mirella and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2:1–31.
Lascarides, Alex and Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics*, 34:387–414.
Lauer, Mark. 1995. Corpus statistics meet the noun compound: Some empirical results. In *The Association for Computational Linguistics Conference (ACL)*, pages 47–54, Cambridge, MA.
Lees, Robert B. 1963. *The Grammar of English Nominalisations*. Mouton, The Hague.
Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13:249–263.
Lersundi, Mikel and Eneko Aggire. 2006. Multilingual inventory of interpretations for postpositions and prepositions. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht, pages 69–82.
Levi, Judith. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York.
Linstromberg, Seth. 1997. *English Prepositions Explained*. John Benjamins Publishing Co., Amsterdam/Philadelphia.
Litkowski, Kenneth C. and Orin Hargraves. 2005. The Preposition Project. In *The ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and*

Applications, pages 171–179, Colchester, UK.
Luraghi, Silvia. 2003. *Prepositions in Greek and Indo-European*. Benjamins, Amsterdam.
Lyons, Christopher. 1986. The syntax of English genitive constructions. *Journal of Linguistics*, 22:123–143.
Melis, Ludo. 2002. *Les prépositions du français*. L'essentiel français. Ophrys, Paris/Gap.
Meyers, A., R. Reeves, Catherine Maclead, Rachel Szekely, Veronika Zielinsk, Brian Young, and R. Grishman. 2004. The cross-breeding of dictionaries. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 1095–1098, Lisbon.
Mihalcea, Rada and Ehsanul Faruque. 2004. SenseLearner: Minimally supervised word sense disambiguation for all words in open text. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 155–158, Barcelona.
Moldovan, Dan and Adriana Badulescu. 2005. A semantic scattering model for the automatic interpretation of genitives. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 891–898, Vancouver.
Moldovan, Dan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *The Human Language Technology Conference / North American Association for Computational Linguistics Conference (HLT/NAACL), Workshop on Computational Lexical Semantics*, pages 60–67, Boston, MA.
Moldovan, Dan and Roxana Girju. 2003. *Proceedings of the Tutorial on Knowledge Discovery from Text*. Association for Computational Linguistics, Sapporo, Japan.
Nakov, Preslav and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *The 9th Conference on Computational Natural Language Learning*, pages 835–842, Vancouver.
O'Hara, Tom and Janyce Wiebe. 2003. Preposition semantic classification via Treebank and FrameNet. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 79–86, Edmonton.
Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *The International Computational Linguistics Conference / Association for Computational Linguistics (COLING/ACL)*, pages 113–120, Sydney.
Pantel, Patrick and Deepak Ravichandran. 2004. Automatically labeling semantic classes. In *The Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 321–328, Boston, MA.
Pennacchiotti, Marco and Patrick Pantel. 2006. Ontologizing semantic relations. In *The International Computational Linguistics Conference / Association for Computational Linguistics (COLING/ACL)*, pages 793–800, Sydney.
Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.
Pustejovsky, James, Catherine Havasi, Roser Sauri, Patrick Hanks, Jessica Littman, Anna Rumshisky, Jose Castano, and Marc Verhagen. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *The International Conference on Language Resources and Evaluation (LREC)*, pages 385–388, Genoa.
Romaine, Suzanne. 1995. *Bilingualism*. Blackwell, Oxford.
Rosario, Barbara and Marti Hearst. 2001. Classifying the semantic relations in noun compounds. In *Conference on Empirical Methods in Natural Language Processing*, pages 82–90, Pittsburgh, PA.
Rosario, Barbara, Marti Hearst, and Charles Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *The 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 247–254, Philadelphia, PA.
Saint-Dizier, Patrick. 2005a. PrepNet: A framework for describing prepositions: Preliminary investigation results. In *The 6th International Workshop on Computational Semantics*, pages 25–34, Tilburg.
Saint-Dizier, Patrick, editor. 2005b. *Syntax and Semantics of Prepositions*. Springer, Dordrecht.
Selkirk, Elisabeth. 1982a. Some remarks on noun phrase structure. In Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*. Academic Press, London.
Selkirk, Elisabeth. 1982b. *Syntax of Words.* MIT Press, Cambridge, MA.
Spärck Jones, Karen. 1983. Compound noun interpretation problems. In Frank Fallside and William A. Woods,

editors, *Computer Speech Processing*. Prentice-Hall, Englewood Cliffs, NJ, pages 363–381.
Tatu, Marta and Dan Moldovan. 2005. A semantic approach to recognizing textual entailmant. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Proceesing (HLT/EMNLP 2005)*, pages 371–378, Vancouver.
Turney, Peter. 2006. Expressing implicit semantic relations without supervision. In *The Computational Linguistics Conference / Association for Computational Linguistics Conference (COLING/ACL)*, pages 313–320, Sydney.
Tyler, Andrea and Vyvyan Evans. 2003. *The Semantics of English Prepositions: Spatial Sciences, Embodied Meaning, and Cognition*. Cambridge University Press, Cambridge, MA.
Vandeloise, Claude, editor. 1993. *La couleur des prépositions*, volume 110. Larousse, Paris.
Villavicencio, Aline. 2006. Verb-particle constructions in the World Wide Web. In Patrick Saint-Dizier, editor, *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht, pages 115–130.
Volk, Martin. 2006. How bad is the problem of PP-attachment? A comparison of English, German and Swedish. In *The European Association for Computational Linguistics (EACL), the ACL-SIGSEM Workshop on Prepositions*, pages 81–88, Trento.
Vossen, Peter. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Verlag.
Winston, Morton, Roger Chaffin, and Douglas Hermann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11:417–444.
Zelinski-Wibbelt, Cornelia, editor. 1993. *The Semantics of Prepositions*. Mouton de Gruyter, Berlin.

This article has been cited by:

1. Timothy Baldwin, Valia Kordoni, Aline Villavicencio. 2009. Prepositions in Applications: A Survey and Introduction to the Special IssuePrepositions in Applications: A Survey and Introduction to the Special Issue. *Computational Linguistics* **35**:2, 119-149. [Citation] [PDF] [PDF Plus]